(Un)Expected consequences of becoming a new format shopper: A causal approach

Xavi Vidal-Berastain Paul Ellickson Mitch Lovett*

Abstract

We quantify the causal impact of new format/platform entry on a broad range of consumer shopping behaviors, including overall spending, frequency of visit, breadth and depth of basket and brand loyalty. This is a large undertaking, made feasible only by recent innovations in data availability and statistical methodology. In particular, we exploit machine learning and data fusion techniques to combine a rich collection of data on consumer choice and store entry decisions into a unified longitudinal panel and employ and extend frontier synthetic control methods to address issues of selection and heterogeneous response inherent to this context.

Keywords: Grocery, shopping habits, new format adoption, cross category spillovers, synthetic methods

*Marketing department, Simon Business School, University of Rochester. Email: xavi.vidal.berastain@gmail.com, paul.ellickson@simon.rochester.edu, mitch.lovett@simon.rochester.edu, Our empirical results are derived based on data from The Nielsen Company (US), LLCand marketing databases provided by the Kilts Center for Marketing Data Center at The University of Chicago BoothSchool of Business. The conclusions drawn from the Nielsen data are those of the researchers and do not reflect theviews of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing theresults reported herein

1 Introduction

A large and historically stable industry, grocery retailing has witnessed almost continuous disruption over the past several decades. While the traditional supermarket format was an important post-war innovation, yielding the big box format that now defines modern retailing throughout the world, it has gradually been refined and replaced by a diverse collection of sophisticated retail platforms, including supercenters, club stores, limited assortment stores, organic specialists, and a small, but growing online channel. Consumers now face a rich array of shopping options that differ by size, location, assortment, and price, alongside other less quantifiable dimensions such as ambiance and convenience. The goal of this study is to quantify the causal impact of such new format/platform entry on a broad range of consumer shopping behaviors, including overall spending, frequency of visit, breadth and depth of basket, brand loyalty, and other dimensions. This is a large undertaking, made feasible only by recent innovations in data availability and statistical methodology. In particular, we exploit machine learning and data fusion techniques to create and combine a rich collection of data on consumer choice and store entry decisions into a unified longitudinal panel. We further employ frontier synthetic control methods (Abadie et al. [3]) to address the issues of selection and heterogeneous response inherent to this context, extending them to our dis-aggregated setting.

Grocery retailing is a large part of any modern economy. According to the Food Marketing Institute, U.S. shoppers spent 5.5% of their disposable income on food at home, yielding a total of $683 billion in sales in 2017. Owing to both their perishable nature and high frequency of purchase, these sales take place almost exclusively in physical outlets, with less than 1% occurring online. Moreover, grocery retailing is surprisingly complicated, as it involves what is essentially an exercise in joint production between consumer and retailer that takes place over a sophisticated two-sided platform linking consumer product manufacturers (brands) to their downstream shopping base. Amongst other things, consumers must decide whether, how often and how far to travel to a store, navigate trade-offs over assortment and price, and decide whether to purchase staple ingredients and packaged products or fully prepared meals. Perhaps not surprisingly, sophisticated retailers have responded by offering consumers a wide set of options, ranging from no-frills limited assortment stores that avoid national brands and target the urban poor to high end organic specialists that target the rich and time constrained to club stores offering bulk products to those who can store and stockpile products.

While there is now a small literature that examines how such formats compete with one another, much less is known about how these options impact consumer shopping behaviors. This dearth of research is likely due to the confluence of three significant challenges. First, the setting is inherently complex. Second, the data requirements are massive, demanding rich shopping histories spanning a diverse set of retail environments. Third, consumer store choice and store location (and format) decisions are strategic outcomes. The types of markets Whole Foods enters are quite

different from those that Wal-Mart targets. Similarly, the types of products and amenities that a typical Whole Foods shopper desires are undoubtedly different from those sought by a Wal-Mart customer. We address these challenges by combining novel data and recent econometric methods with a tailored conceptual model of the retail environment.

Turning to the first challenge, the features that distinguish one retail format from another are challenging to quantify, as they involve an inter-connected set of attributes such as store size, assortment, package size, private label emphasis and store ambiance that defy simple characterization. Moreover, the impact of new format entry depends on the characteristics and shopping behaviors of the particular consumers exposed to it. Consider the recent entry of the European limited assortment chain Lidl into the southeastern US. Lidl operates small footprint stores that emphasizes store branded products and low prices. Amenities are minimal, but their small size and convenient locations may expedite shopping time. A shopper exposed for the first time to a new Lidl might be expected to increase their frequency of visits, shift their purchases more toward unbranded products and decrease overall expenditure.

However, each of these predictions should further depend on how they shopped prior to Lidl's entry and what their choice set originally included. Would the prediction be the same if the shopper was already consuming private labels and consuming a limited bundle of products? Significant heterogeneity of such causal effects would mean that proper demand modeling would need to account for what formats the consumer was already exposed to, how much they were spending and how many trips per month they were making. Failing to control for these complex sources of variation could potentially invalidate any inference due to the existence of omitted, time-varying confounders.

Furthermore, one might also expect their behavior at existing stores to change. Perhaps the savings they enjoy via Lidl's low prices are diverted into greater engagement with organic products. Now contrast this with new entry by a club store. We might expect a club-prone consumer to purchase more items in bulk and consolidate their shopping trips into fewer visits. But who is the "club-prone" segment? And do they simply switch their basket to the new club, or does its limited offerings mean they now have to multi-home (i.e. shop at more than one platform)? To tackle these questions, we first develop a conceptual framework for organizing the set of customer demands and retail offerings, proposing concrete measures of each that can be constructed from available data sources.

Turning now to the second challenge, our empirical task is complicated by its inherently intensive data requirements. In particular, we require detailed purchase histories for shoppers spanning a wide variety of retail environments in which choice sets are changing. These histories must be available both before and after entry, and in a wide enough array of settings to span the feasible option sets. To address this challenge, we combine the rich Nielsen consumer panel (available through the Kilts Center) with novel information on store entries from a variety of public and proprietary

sources. To compile the dataset on store entries, involved an extensive text analytics application in order to extract the critical information. We develop a unique data set compiled from 4 different sources which combines 7,847 store locations with more 3,700 store openings from six retailers representing, two limited assortment retailer, one organic specialist, one wholesale club, one high-end supermarket and one discounter. These openings occurred between 2004 to 2015 in areas spread throughout the US, representing over 130,000 households tracked by Nielsen.

The third and final challenge lies in recovering causal treatment effects for highly selected and heterogeneous treatments. Note that classical methods, such as difference in difference, are inappropriate here due to the selective nature of firm entries and the complex and heterogeneous impact of consumers' past shopping histories. Both issues are likely to invalidate the parallel trends assumption by introducing time-varying confounders. To tackle this problem, we adopt and extend recent methods from the synthetic controls literature (Abadie et al. [3]). In particular, for each treated individual in the Nielsen sample, defined as a household that was exposed to novel entry by one of our six focal retailers, we construct a synthetic clone from a "donor pool" of untreated consumers who were not exposed. Using their detailed shopping histories, we first match the past behavior of the focal treated shopper to a synthetic pair built from the untreated "donors" and then define the treatment effect on that treated household as the difference in outcomes between these matched pairs. Inference is conducted using permutation methods.

In order to apply the synthetic control methods to our unbalanced pool of households, we extend these methods along two dimensions. First, we develop an approach that optimizes the tuning parameters for constructing the synthetic control based on the data available to each treated individual. This procedure optimally leverages the strengths of both households who are in the panel a long time, but have fewer donors because of the long tenure and households who are in the panel for a shorter time and have more donors. Second, we found that using the optimization techniques that have traditionally been used in the synthetic controls methodology makes computation of the thousands of treated individuals and donors nearly infeasible. We apply a new optimization approach to the synthetic controls methodology that has properties favorable to the problem scale of our panel data. In particular, we borrow from the machine learning literature and adapt the mirror descent algorithm to our problem (Nesterov et al. [39]). This algorithm provides a dimension-free convergence rate that provides large speed improvements in the repeated optimization problem we face.

The findings provide a rich pattern of results about how new retail platforms can shape household shopping behaviors. One strong theme in the results is that new entries with high ambience and unique assortments allow retailers to be added (rather than replace) to the set of retailers visited as well the total items and expenditure. This suggests that retailers can expand the market with such offerings, but we find that these effects are largely concentrated in households visiting many retailers. The third theme is that many of these retailers successfully shift consumption toward private labels, perhaps helping to explain the increase in private label consumption. Another clear

pattern is that bulk offerings change household behavior consistent with increased stockpiling. We find that households decrease their shopping trips and items purchased and that these effects are most concentrated in households that prior to the entry shopped at only one or two retailers. This pattern suggests that new formats with bulk offerings can shape how households use stockpiling to manage their shopping. Finally, we find that the discounter, which is the closest to a traditional supermarket model in our sample, doesn't appear to change the way households shop when it enters. This suggests that simply offering low prices does not fundamentally change shopping behaviors.

This research contributes to several distinct streams of literature in marketing and economics. First, we contribute to the store choice literature, which seeks to understand the key dimensions upon which consumers choose where to shop. This includes the foundational retail gravity model of Reilly [43], as well as more recent work on assortment (Broniarczyk et al. [8], Dreze et al. [17], Broniarczyk and Hoyer [7], Briesch et al. [5]), category breadth (Briesch et al. [6]), and shopping experience (Bell and Lattin [4]). The closest papers to ours are Fox et al. [25] and Cheng [12], which examine format choice, and Thomassen et al. [49], who study both store choice and expenditure, while emphasizes the distinct impact of single- versus multi-store shoppers.

We also contribute to the large literature on private label products. The seminal contributions of Hoch and Banerji [26] and Hoch and Lodish [27] focused attention on the key drivers of differential sales of private label products across different retailer types. Subsequent research has extended these ideas, examining how private label entry influences prices (Chintagunta et al. [13]), margins (Pauwels and Srinivasan [41]), and bargaining outcomes (Meza and Sudhir [38]; Ellickson et al. [23]). The most recent research on private label demand has focused on analyzing its connection with income (Dubé et al. [18], Cha et al. [11] ,Chintagunta et al. [13], Nevo and Wong [40], and Lamey et al. [32]). Our results point to the importance of exposure to entire formats whose focus is on emphasizing and highlighting private label offerings.

Finally, our paper is also related to the literature studying context dependent preferences. There is a large literature in psychology, marketing and economics studying how the addition of a new and very differentiated element (in our case, a new store of a different format) to the choice set of a consumer alters her perception of all the elements in her choice set. Research on consumer psychology has provided strong hints that consumers take grocery decisions relating observable attributes to pre-established stereotypes. These stereotypes are only reconsidered when a new and very differentiated element is introduced in the choice set. Important works in this vein include Dijksterhuis et al. [14], Dunning [20], Douglas [16], Rucker et al. [46], Kőszegi and Szeidl [31] and Zhu and Dukes [54].

The paper is organized as follows. Section 2 describes the data and institutional setting. Section 3 details our conceptual framework for shopping behavior and format types. Section 4 provides an overview of our empirical framework, identifying the key threats to causal inference and

the solution provided by synthetic controls. Section 5 contains our main results. Section 6 concludes.

2 Data and Setting

In this section, we introduce the two primary datasets used in our empirical analysis. First, we discuss the Kilts Center Nielsen Homescan Academic Data set. This data contains the household-level purchase histories for all retailers. When we introduce this data, we also discuss the key shopping behavioral variables under study. Second, we introduce the our data related to retail format entry locations and timing. We discuss the specific types of retail format entries as well as our data collection, which involved a combination of web scraping, text analytics, and purchased data.

2.1 The Nielsen Homescan Panel and Shopper Behavioral Measures

Data on households purchasing decisions comes from the Nielsen Homescan Panel of consumers, (henceforth, the HMS).[1] For the years spanning 2004-2015, AC Nielsen collects daily UPC-level purchase data for a stratified random sample of approximately 70,000 US households. More specifically, for each shopping trip that a panelist performs, we observe an identifier of the store visited, the whole basket of goods purchased at the upc level, the price paid, the quantity purchased and a battery of product descriptors such as its brand, whether it has the USDA-organic qualification or its Nielsen classification.[2] We also observe the retailer group owning each visited store. The exact location is not available but, as a proxy for it, Nielsen reports store locations at the 3 digit zip code level. Aside from the purchase data, Nielsen keeps yearly-updated and self-reported demographics which include information on income, race and the five digits zip code of the household's primary residence. In total, from year 2004 to 2015, the households in the HMS report more than 119 million shopping trips to any of the 65,726 classified stores, which amounts to more than 700 million transactions.

The HMS data are suited to answer our research question for three reasons. First, the household panel provides purchase information for all retailers for the same households, which is not available in the retail scan data or from loyalty card program data. Second, the geographic dispersion of the HMS data is sufficient to cover a broad range of store entries. Figure 1 is scatter plot of the where the households in the HMS live over the USA, all the major urban areas are perfectly covered irrespective of the state, county or region. Compared to the HMS, other panels of consumers

[1] We thank the University of Chicago and James M. Kilts Center for Marketing for letting us use this data. The views expressed in this paper are the authors' only.

[2] Nielsen classifies each upc into one of 9 categories, namely, alcoholic beverages, dairy, deli, dry grocery, fresh produce, frozen foods, general merchandise, health & beauty care, magnet data, non-food grocery, packaged meat. Each of these categories is then divided into modules, which represent lower level groupings.

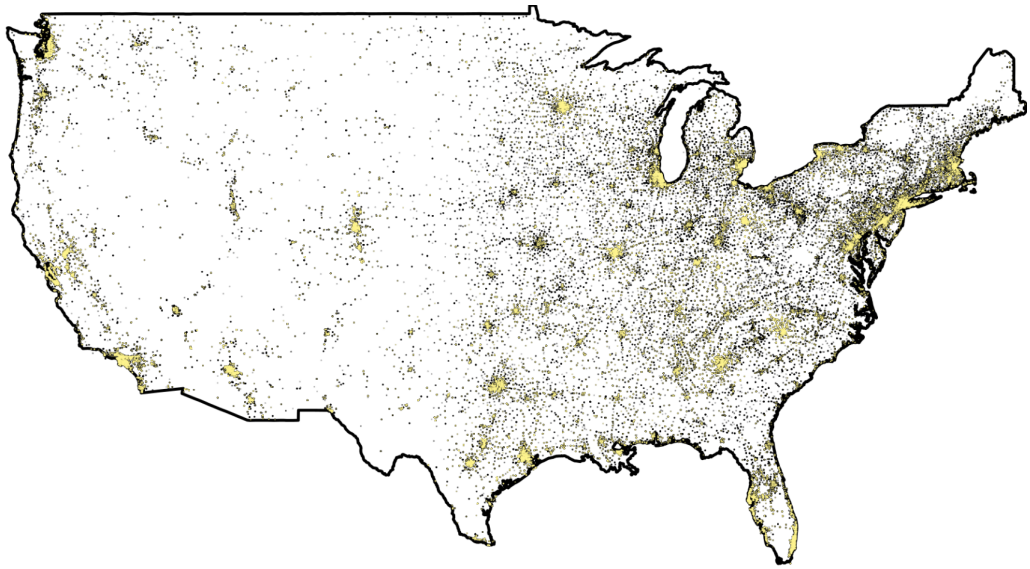


Figure 1: HMS locations. Black dots represent a 5 digits zip code with a household living there. Yellow dots indicate a household living in that zip codes and inhabiting the HMS since 2004 to 2015, namely the whole time span

available to researchers focus on very specific regions. The IRI academic panel of consumers for example is restricted to households living nearby Eau Claire, Wisconsin and Pittsfield, Massachusetts where only 4 openings occurred in the time span they cover. Furthermore, such concentration on where the households live makes it impossible to construct control groups to identify the effect on the treated because there are so few markets. Finally, our methodology relies on time series variation to ensure the match between control and treated cases is valid. In the HMS 80% of the panelists stays at least 3 years and more than 7 thousand panelists stay more than 10 years in the panel.

We follow Dubé et al. [18] in constructing the final data set used for the estimation, performing a number of filters on households, trips and transactions to trim inconsistent observations. Specifically, to be included in the final data set, households must make at least one trip per month to stores classified as grocery, wholesale club, discounter, convenience store or dollar store. Households entering and leaving the panel are trimmed out. In addition, because our study focuses on entry, we need to also control for household migration between areas to avoid migration in and out of areas with the focal format. To address this concern, we trim households living in more than 3 different zip codes. On average we conserve around 80% of the households with year 2006 being the lowest (76%) and year 2014 being the highest(83%).

From the Homescan data, we draw several measures of household shopping behaviors that form the focus of our analysis. These measures capture the breadth of shopping, the frequency of shopping, the breadth of purchases (items and expenditure), the extent that products are purchased on promotion, and the extent of private label use. All of the measures are constrained to shopping

for groceries, so that shopping trips where only non-grocery items are purchased are dropped and only grocery items and costs are included. The measures are aggregated to the monthly level and available for every household in the HMS sample. The specific measures for the shopping behaviors of interest are:

Retailers visited: the average number of unique retailers (retailer codes) visited per month. This measure captures the breadth of the retailers the household typically visits.

Shopping trips: the average number of shopping trips per month. This measure captures the frequency of shopping.

Items: the total number of items purchased per month, which captures the breadth of purchases in terms products.

Expenditure: the total expenditure of the household on groceries, which also captures the breadth of purchases, but in terms of dollars.

Items disc: the number of grocery items purchased on discount per month, which captures the extent that the household leverages promotions in their shopping behaviors.

Private label share: the household's share of total expenditures that private labels make up. This captures how heavily the household uses private labels.

Private label items: the number of items the household purchases that are private label, which captures the breadth of purchases in terms of products, but for private labels only.

Private label expenditure: the household's total expenditure on private label products, which captures the breadth of purchases in terms of dollar spend, but for private labels only.

2.2 The Entry Formats and Store Data

We now describe the retail formats that we focus on in this research and the data collection related to identifying store openings and locations. Our analysis focuses on entries by six different retailers spanning five different types of retailers: Limited Assortment Retailers (LARs), Organic/Natural

Foods Specialists (Org), Warehouse Clubs (Club), Discounters (Disc), and the broader set of supermarkets (Sup). We choose these formats because, in the years spanned in the HMS, limited assortment retailers, organic specialists, wholesale clubs, have been extremely prolific in terms of opening stores, they have attracted a lot attention of retail managers and been able to capture size-able shares of the grocery market. We also include one more traditional supermarket that holds an upmarket position, and a discounter that holds a more established position. We note that due to data limitations, we are not permitted to reveal the identity of the retailers. We describe each set and general information about the retailers below.

Limited Assortment Retailers (LARs) Limited assortment retailers are characterized by stores smaller than average sized supermarkets that offer a limited amount of products. Often they will have minimal or no perishables offerings like meat or produce. Usually they have lower prices than most food stores and are dominated by private label products (up to 95% of their listings). They often provide only one size per variety and typically offer lower prices for like products, even when comparing private label products against other private labels of traditional supermarkets. In our data, we have two LARs that offer different format positions and primary geographic focuses. We have a total of 1,188 store openings for these two LARs.

Organic/Natural Foods Specialists (Org) Sometimes regarded as ecosupermarkets or organic-supermarkets, these retailers target high income customers. Stores are upscale and tend to have the same breadth of product categories as standard supermarkets, but the assortment is generally focused on products that are unprocessed and unrefined, or processed and refined as little as possible, before being consumed. The pricing strategy is closer to promotions than to EDLP and they tend to include a large array of fair-trade products. Typically, their price points are relatively high because of the types of products they offer within any given product category. We have one such retailer with 285 store openings in our data.

Warehouse Clubs (Club) Warehouse Club stores are no frills operations where products are often displayed as pallets or fixtures that were shipped directly from the manufacturer. Most warehouse clubs require a membership to be able to shop in their stores in their club stores. They tend to focus on selling products at low prices, in large quantities. These goods usually are sold in larger package sizes, multi-packs, or bulk-packaged, so that the assortments match well with large families and businesses. We have one such retailer with 172 openings in our data.

Discounters (Disc) or discount stores offer a wide assortment of goods with a focus on price rather than service, display, or store location. They differ from LARs stores in offering more breadth and depth of products, i.e. they sell many name-brand products from many different categories. This

large heterogeneity in the assortment creates a wide price range of the items offered. We observe 1536 store openings for this firm.

Supermarkets (Sup) A store between 20,000 and 50,000 square feet that offers a full line of grocery, meat and produce with an average of 10,000 to 30,000 sku's and a sales volume over $2 million annually. We have collected data for one such retailer in our data. This retailer is well-established in a number of regional markets (like most retailers of this type) and enters into new markets during our data period. The new market store entries tend to be upscale. We observe 32 new store openings.

As mentioned for each of these retailers and retail formats, we observe that these retailers open new stores and we wish to study how consumers near those stores change their shopping behaviors in response to these new retail format entries. Importantly, we need to know exactly where and when the store openings occur in order to both clearly delineate the period of interest and to avoid duplicate stores covering a household. The challenge is that the Homescan data does not provide information at a sufficiently detailed level for this analysis. As a result, we conduct an extensive effort to collect the necessary data.

This process involves two main steps. First, we seek to understand the number and general identity of the store openings for each of the retailers. This is the easier of the two tasks. The second task requires us to identify the exact location and the timing (within a month) of each store opening. This task is much more difficult and involves applying various data mining strategies including web scraping and machine learning on text data in addition to purchasing data.

We obtain store-level information about all retailers from TDLinx for the year 2004-2006. TDLinx collects store level data from every supermarket operating in the U.S. The data are sold to marketing firms and food manufacturers for marketing purposes. This data set has been previously used in Ellickson and Misra [21] and Ellickson and Misra [22]. This data provides a baseline, but does not cover our full data period.

The second data source that we use is the website of each of our focal retailers. We use a webcrawler to obtain all of the current store locations from their websites. Combined with the TDLinx data, we obtain a clear identification of the new stores that opened between 2004 and 2017.

Our next step is to find the opening date and more detailed location for each of the store entries we know opened in the years spanned by the HMS. We note that for two of our retailers, their public website contains the opening dates and exact locations for the new stores. For the rest, we use a combination of methods to triangulate on these openings. We describe these methods briefly below:

1. **News and Text classification:** The first approach is to leverage text data from local newspapers and magazines. Due to the importance of the grocery shopping in everyday life of people, it is quite common in local newspapers and magazines to announce store openings. News-banks like Lexis-Nexis or Factiva archive these news stories. We recover the news mentioning each of our focal retailers. However, such a query is imprecise and recovers many articles that are not relevant to store openings. For example, a search for one of our retailers in Lexis-Nexus for January of 2004 alone produces over 2,000 articles. With so many articles, manually reading them is not feasible. We adapt a feature selection technique namely the term frequency inverse document frequency `tf-idf` to trim out articles that are likely orthogonal to our topic and to sort the newspaper articles based on the probability of being relevant to store openings. Formally, we define $\mathbf{C}$ to be the pool of the news that we collect from Lexis-Nexis or Factiva. The pool $\mathbf{C}$ is a long vector where each component contains a newspaper article. In the text mining jargon, $\mathbf{C}$ is referred to as the corpus, and each instance of the corpus is regarded as a document and noted by d. Finally each document d in corpus $\mathbf{C}$ is a set of words w. Formally:

$$\mathbf{C} \triangleq \bigcup_{i=1,\cdots,N_{\mathbf{c}}} d_i \tag{1}$$

$$d_i \triangleq \bigcup_{i=1,\cdots,N_{d_i}} w_i \tag{2}$$

We next define the two statistics that determine how likely a newspaper article is related to the opening of a store given retailer code in a given location. The term frequency is a measure of how prevalent the term j is in document d, whereas the inverse document frequency is a measure of how specific the term j is relative to the the whole corpus $\mathbf{C}$. Formally they are given by:

$$\texttt{tf}(j,d) \triangleq 0.5 + 0.5 \cdot \frac{\sum_{i\in 1\cdots N_{d_i}} \mathbb{1}_{w_i=j}}{\max_{s\in d_i}\left[\sum_{i\in 1\cdots N_{d_i}} \mathbb{1}_{w_i=s}\right]} \tag{3}$$

in equation (3) the denominator is just a correction for the bias created by the fact that some documents are longer than others. This measure takes higher values the more times a word appears. The **idf** measures how specific the term j is and is defined as

$$\texttt{idf}(j,\mathbf{C}) \triangleq \ln(N_{d_i}) - \ln\left(\sum_{d=1,\cdots,N} \mathbb{1}_{j\in d}\right). \tag{4}$$

The `tf-idf`(j,d,t) is defined as

$$\texttt{tf-idf}(j,d,t) \triangleq \texttt{tf}(j,d) \times \texttt{idf}(j,\mathbf{C}) \tag{5}$$

The filtering rule that we use to determines the relevance of a document is the sum `tf-idf`

for a set of terms that we consider informative. Our set of terms $\mathbf{S}$, is a dictionary of words that contains enties such as "open", "grand-opening", "ribbon-cutting", and both the name of the towns where we know retail entry occurs and the name of the retailers. Formally:

$$\texttt{inx}(d) \triangleq \sum_{j \in \mathbf{S}} \texttt{idf}(j, d, \mathbf{C}) \tag{6}$$

We sort our news based on `inx` and manually analyze the newspaper article that receive highest `inx`.[3]

2. **yelp.com:** The second approach leverages consumer review data to signal a new retailer entry. We use Yelp.com, an online site where user generated content and crowd-sourced reviews are published about local businesses, such as restaurants, bars and supermarkets. The company generates revenues from serving advertisements, training small businesses in how to respond to reviews, hosting social events for reviewers, and providing data about businesses, including health inspection scores. When a new business opens, the first user that posts a comment receives a "first-reviewer" mark on his profile. Heavy posters, compete to collect "first-reviewer" marks as a kind of badge of honor. We code a python crawler to find the first review of each of the stores that opened at some point in time between 2006 to 2015.

3. **Market research companies:** The third source of openings is to purchase data. The same company, TDlinx, has a data base of store opening dates and imputed revenues that they sell commercially and to academics. Papers in the literature using this data includes Zheng [53], Holmes [28] and Ellickson and Misra [21]. They claim to be the best available store-level sales and location data in the discount retail. We purchased access to store opening data for two of our focal retailers, where the above methods left meaningful gaps.

The combination of these approaches yields 3,213 store openings over the 2004-2015 period of the consumer panel data with the necessary location information for our analysis. Across the retailer store entries, we have over 32,000 treated households.

3 Conceptual Framework

In this section, we introduce a framework for understanding how retail format differ from one another in service of organizing our predictions about the impact of retail entry. We first discuss characteristics of retail formats and how these characteristics are predicted to influence shopping behaviors. We argue that these characteristics are not evenly viewed by consumers and instead that the prominent attributes of the retailer shape it's perceived format. We then discuss the measures we

[3] In a pre-analysis process, duplicates have been removed

use to operationalize these characteristics and a machine learning method that follows our argument of prominent attributes in order to transform the measures of the retail format characteristics into a classification system that can be applied to any retailer in the Nielsen HMS data.

3.1 Retailer Format Characteristics and Their Implications

To the best of our knowledge, there is no consensus on what characteristic or set of characteristics determines whether a retailer belongs to a given format.[4] Hence, to organize our predictions about the causal effects of retail format entry on shopping behaviors, we introduce a set of characteristics of retailers that relate to the perceived retail format.

Our framework starts from the assumption that a small set of prominent attributes of a retailer define it in the mind of consumers. We present a set of candidate characteristics that we argue can serve as prominent attributes in shaping consumer perceptions of the retail formats. Below we discuss each of these characteristics and how they are expected to influence new format shoppers after entering a new market.

Ambience: Many retailers spend resources to create engaging store environments (see the examples in Levy et al. [33] and Zentes et al. [52]). These engaging store environments can imbue the retailer with a sense of ambience, which consumer behavioral research define as the (positive) sensorial experiences perceived during shopping episodes(Mehrabian and Russell [35], Mehrabian and Russell [36]). Previous consumer behavior literature has provided evidence that such pleasant shopping environments are positively correlated with store spending (Sweeney and Wyber [48],Robert and John [44]), social interaction (Robert and John [44], Dubé et al. [19], Sherman et al. [47], Sweeney and Wyber [48] Yalch and Spangenberg [50], Yoo et al. [51]), visit duration(Robert and John [44]), and visit satisfaction (Sherman et al. [47], Sweeney and Wyber [48], Yalch and Spangenberg [50], Yoo et al. [51]). These consequences of visiting a store with ambience, directly suggest that retail formats with higher levels of ambience are more likely to increase household expenditures and expand the items purchased upon adoption of such a format. Further, the higher levels of satisfaction and derived utility from shopping at such a store suggest that households may find such a store attractive enough to visit for the utility of shopping there as much as filling necessities. Hence, in contrast to store environments that are less engaging and more likely used simply to purchase required goods, retail formats with more ambience are more likely to increase the number of shopping trips and, potentially, increasing the size of the set of retailers visited. One other growing element of the retail shopping environment that engenders ambience is fresh and prepared food [42]. We predict that

[4] We find that there is not even agreement on how many retail formats are there in total. We have been able to list up to 34 different (so called) retail formats. The list includes a wide range of examples ranging from the well known cases of, limited assortment retailers or wholesale clubs to lesser known examples such as full line store or combination store, or off-price superstores

this element could lead households to increase the items they purchase, as they replace restaurant purchases with grocery purchases.

Organic prevalence: Many recent format entries and existing formats offer a larger assortment of organic products. These products generally cater to a higher income level due to their higher prices, and increase the set of options available within a category. Although organic prevalence is often associated with better ambience, for new shoppers of a retailer with this prominent attribute, we expect the organic prevalence to further increase the number of items purchased and the total expenditure.

Private label prevalence: Most of the retailers we examine, and most retailers today, have private labels in many categories. Traditionally, private labels serve as me-too products at a lower price. However, some retailers specialize more in providing higher quality or unique private label offerings. Thus, we expect two possible effects. First that high private label prevalence for a traditional private label offering would lead to increased private label share, expenditure, and items, but potentially lower total expenditure due to the cost savings. Second, for a higher quality, unique private label offering, we would expect the private label offering to be similar to the organic prevalence, individuals seek out to purchase more items and expand items purchased.

Full-line coverage: Acknowledging the importance of the assortment, store managers carefully choose a store's breadth (number of product categories to offer) and depth (number of varieties in each category) to attract and retain consumers. Research on how assortment affects store choice started with the law of retail gravitation (Reilly [43]) which states that the probability of choosing a retailer is positively correlated with the product of its breadth and depth.[5] However, empirical evidence on the validity of the law of retail gravitation is mixed, see the works of Broniarczyk and Hoyer [7], Broniarczyk et al. [8], Briesch et al. [5], Briesch et al. [6], Dreze et al. [17], Fox et al. [25]. Studies in consumer psychology suggest that an important moderator governing the influence of assortment on store choice is variety-seeking (Kahn [30], Menon and Kahn [37], McAlister and Pessemier [34]). Thus, variety-seeking consumers that shop at stores with limited assortment are likely to find the assortment insufficient to meet their needs. This will lead them to add the store to their shopping basket rather than replace an existing store. Hence, they will increase not only the number of retailers visited but also their number of shopping trips. In contrast, stores with full-line coverage can meet the needs of the variety seekers and serve to replace existing retailers upon starting to shop there. This implies no change in the number of retailers visited and shopping trips. For the consumers that are not high variety-seekers, the more limited assortment may be less problematic for meeting their needs, leading to smaller changes in shopping visits and retailers

[5] See Brown [10] for a review on the law of retail gravitation

visited.

Convenience: The convenience offered by a given store is reflective of shopping costs such as the distance traveled, in-store search time, and the time spent waiting to check out. While travel time is important, we do not consider it here, since it varies by individual. The remaining aspects related to in-store convenience are shaped meaningfully by the size of the store, with the depth of the assortment generally reflecting this size. Smaller stores can allow shopping to be faster and less time consuming. This implies convenience will lead to more frequent shopping trips and more likely that the store will be added to the retailers visited. Further, the recent marketing literature has explored how attention on the number of varieties available poses a costly penalization in the consideration stage of the decision funnel (Bronnenberg [9],Huang and Bronnenberg [29]). By lowering consideration costs, small stores can further increase the convenience of shopping at smaller stores with less variety.

Bulk offering prevalence: While the availability and extent of bulk product offerings reflects the opportunity to buy more product, usually at a discount, high bulk offering prevalence can require it. In principle, buying bulk sizes offerings implies that the product will last longer before repurchase is necessary. Such a stockpiling effect would imply that shopping frequency and the number of items purchased is likely to decrease. However, to the extent that the bulk offerings are indeed cheaper per unit, this can lead to a decrease in price and a resulting income effect that could offset the decrease in items purchased.

Price: While there is large literature in quantitative marketing, economics, and consumer behavior on the importance of perceived and actual price in shopping behaviors, this factor plays a smaller role in our analysis of changes in shopper behavior upon new retail format entry. While a important feature of retailers, the price levels in themselves are unlikely to offer a radically new option in terms of prices alone. This is the case both because of the persistent vertical structure of the grocery industry so that most populous markets generally cover the range of price levels and because the new formats price offers are connected to other attributes such as bulk packaging (for low price) or organice prevalence (for high price). Further, the broader pricing strategy of retailers as everyday low price (EDLP) vs. high-low, where high-low offers frequent promotions and EDLP offers consistent (low) prices both existed prior to our analysis period. When EDLP first became common it probably had the capacity to change how consumers shop for promotions, and, as a result, the number of items purchased on discount. By the time of our data, most households appear to already face some EDLP formats, again, making this aspect of the price feature not as prominent.

In summary we have five retail format characteristics that can serve as prominent attributes of retail formats when that format enters a new market. These characteristics are ambience, organic

prevalence, private label prevalence, full-line coverage, convenience, bulk offering prevalence, and price. Each of these characteristics except price is likely to shape the behavior of shoppers after entry as we describe of above.

3.1.1 Single and Multi-homer Response to Retail Formats

The literature identifies that households differ in the number of retailers they regularly visit and these differences create fundamentally different values from shopping Thomassen et al. [49]. For our purposes, we define multi-homers as households that visit more than two unique retailers on average in a month in the year prior to the treatment, and otherwise they are single-homers. Because households are choosing to be single or multi-homing, this reveals their preferences and constraints. Compared to multi-homing households, single-homing households reveal themselves to either have less taste for variety or tighter time budgets.

Because of this revealed difference in the household, our predictions about the role of prominent attributes differ based on whether the household is multi-homing or single-homing. In general, we expect multi-homing households to respond more to the ambience, unique assortments, and (in-store) convenience of the new format options. Thus, for most of the predicted effects, they are more likely to exhibit the effects or will exhibit them with a greater magnitude.

In contrast, because of the limited time budgets, single-homing households are likely to exhibit specific differences in the predictions. In general, we expect single-homing households to have lower effect sizes because of the constraints related to time budgets and their lower (revealed) interest in unique assortments and varieties. However, two specific predicted effects might be more pronounced for single-homing households. First, the influence of bulk offerings might be stronger. Bulk offerings provide time constrained single-homing households with the opportunity to reduce their time commitments to shopping through stockpiling and shopping less frequently. Thus, for single-homing households, we expect the influence of bulk offerings to strongest, reducing both the number of shopping trips and number of items purchased per month. Second, retailers that offer the fresh and prepared foods may provide single-homing households with another time saving device. We capture this aspect through the ambience characteristic, and so we expect retail formats with higher ambience to potentially change single-homing households more. The specific changes expected here would be an increase in shopping trips and items purchased.

3.2 Retail Format Characteristic Measures and Classification

In the HMS data, Nielsen assigns to each retailer a so-called channel type. Nielsen's channel types, in many respects, match an intuitive classification of store formats with categories such as grocery

stores, wholesale clubs, discounters, and dollar stores. However, the channel type variable provided by Nielsen ignores distinctions that are of interest in our research. For instance, the grocery store category neither discriminates an organic specialist from a traditional grocery store nor a limited assortment retailer from an organic specialist. Further, the classification contains some categories that are hard to distinguish (e.g., hypermarkets vs. supermarkets), in terms of predictions we would make about shopping behaviors after such a format entered. As a result, we propose an approach for classifying retailers that focuses around our research question.

This classification system is a natural outcome of this research and can support extensions to the current analysis. Some example extensions including considering how the set of existing retail formats a consumer faces in the market shapes the response to a new format. Can some formats insulate better against the transformative effects of new format entries? Similarly, how can the formats a consumer currently visits influence the impact of a new format entry?

Of course, we have already determined the format types for our focal formats. Here, we provide the measures of the retail format characteristics for these retailers to demonstrate the measurement underpinnings of our classification system. We note that our description of these retail formats was informed by these measures. We then turn to describe the machine learning methods used to classify retailers into different formats.

3.2.1 Measures of the Retail Format Characteristics

Based on the above descriptions, we propose specific operationalizations of the retail format characteristics. We start with the measure of ambience, which is constructed from information outside the Nielsen data. The rest of the measures are constructed from the Nielsen HMS data.

The attribute `ambience` is a measure of how engaging the retail environment is of a given retail format. There is no information in the HMS on these retail environment dimensions. We use as proxy, rankings of the store ambience available on line[6]. To avoid disclosing the identity we cannot be explicit on which position and ranking each retailer occupies each of our stores. We note that the list does not provide the full set of retailers. For the retailers that are missing from the identified ambience list, we set them to 1, which is exactly the median value, since normalize each measure by the median across retailers. This missing data assumption follows from the notion that the missing cases do not have retail environments that are extremely positive, since in that case we would expect them to be on the list of best store environments. This assumption should not generate bias since we focus on *prominent* attributes and these normalized characteristics are by definition not prominent.

[6] In the future we plan to construct a classifier based on online reviews

The measure for organic prevalence, `Org. Prev.`, is defined as the share of the revenue that a retailer receives from selling organic USDA classified upcs. The measure of `Org. Prev` for retailer k is a sum of upcs, u, from the set, U_k, of all purchases in that the retailer. Specifically, we sum the revenue for each upc, $rev_{u,k}$, if that purchase is organic, $\mathbb{1}_{\texttt{Org}_u}$. The equation can be written,

$$\texttt{Org. Prev}_k \triangleq \frac{\frac{\sum_{u \in U_k} \mathbb{1}_{\texttt{Org}_u} \times rev_{u,k}}{\sum_{u \in U_k} rev_{u,k}}}{\texttt{med}\left(\bigcup_{n \in \texttt{HMS}} \left(\frac{\sum_{u \in U_n} \mathbb{1}_{\texttt{Org}_u} \times rev_{u,n}}{\sum_{u \in U_n} rev_{u,n}}\right)\right)} \tag{7}$$

The measure for private label prevalence, `PL Prev.` is defined similar to that of organic prevalence. Private label purchases are indicated by $\mathbb{1}_{\texttt{PL}_u}$, so that the formal definition is

$$\texttt{PL Prev}_k \triangleq \frac{\frac{\sum_{u \in U_k} \mathbb{1}_{\texttt{PL}_u} \times rev_{u,k}}{\sum_{u \in U_k} rev_{u,k}}}{\texttt{med}\left(\bigcup_{n \in \texttt{HMS}} \left(\frac{\sum_{u \in U_n} \mathbb{1}_{\texttt{PL}_u} \times rev_{u,n}}{\sum_{u \in U_n} rev_{u,n}}\right)\right)} \tag{8}$$

The measure `Full-line` measures both the characteristics of full-line coverage and convenience, where convenience is equivalent to low levels of full-line coverage. The measure is defined as the average number of upcs per department, $N_{upc,s}$, divided by the average number of modules per store within store, $N_{dep,s}$. For this measure, we have an additional challenge related to sparsity. We wish to identify the completeness of upc coverage from household purchases. Because some retailers have too few household visits, including them could bias the results downward for those retailers, since the limited visits might artificially result in a lower number of observed upcs. As a result, we have to put a lower bound on the number of visits and households to a given store before we are willing to use this measure. Formally, we sum over stores, s in the observed set of stores for retailer k, S_k, but include only those with a sufficient number of visits, indicated by $1_{\{visited\}}$. The formal definition is

$$\texttt{Full-line}_k \triangleq \frac{\frac{\sum_{s \in S_k} 1_{\{visited\}} \left[\frac{N_{upc,s}}{N_{dep,s}}\right]}{\sum_{s \in S_k} 1_{\{visited\}}}}{\texttt{med}\left(\bigcup_{n \in HMS} \left(\frac{\sum_{s \in S_n} 1_{\{visited\}} \left[\frac{N_{upc,s}}{N_{dep,s}}\right]}{\sum_{s \in S_n} 1_{\{visited\}}}\right)\right)} \tag{9}$$

The final attribute we measure is bulk offering prevalence. The measure, `Bulk`, captures how large the product offerings are that the retailer sells and is formally defined by

$$\texttt{Bulk}_k \triangleq \frac{\sum_{u \in U_k} \frac{N_{u,k}}{N_k} \times \frac{size_u^{nielsen}}{\texttt{med}\ size_{m(u)}}}{\texttt{med}\left(\bigcup_{n \in \texttt{HMS}} \left(\sum_{u \in U_n} \frac{N_{u,n}}{N_n} \times \frac{size_u^{nielsen}}{\texttt{med}\ size_{m(u)}}\right)\right)} \tag{10}$$

N_k is the total number of units sold in retailer k, which is equal to the sum, $\sum_{u \in U_k} N_{u,k}$. The

variable $size_u^{nielsen}$ is the size of the upc reported by Nielsen and `med` $size_{m(u)}$ is the median size of the module where the upc u belongs to.

Figure 2 presents the measures for each of our focal retailers in a polar chart format. Our approach to classification is that households will most likely perceive retailer formats based on their most prominent attributes. For both LARs, the most prominent attribute is the private label prevalence. For LAR I, bulk offerings is the second most prominent, whereas for LAR II, ambience and organic prevalence are also relatively prominent. Thus, even among LARs the next most prominent attributes can differ. For the wholesale club, bulk offering prevalence stands out and for the organic specialist, unsurprisingly organic prevalence and ambiance are prominent. For the discounter, only full-line coverage stands out, whereas for the supermarket, full-line coverage as well as organic prevalence and ambience all appear to be prominent.

3.2.2 Classification based on prominent attributes

In our current analyses, we evaluate the effect of the size of the set of retailers the household visits, but would like to extend this to consider the types of retailers in those sets. In order to do this, we develop a natural extension of our framework–a classifiction algorithm of retailers into retailer types based on their prominent attributes. Although we haven't folded the resulting data into our results, we present here our algorithm to classify the remaining retailers in the Nielsen HMS data.

To accomplish the classification task, we use a supervised machine learning approach. As the training, "labeled" data, we use the format definitions for our focal retailers along with a set of additional retailers, which we de-identified using the RMS data. We now describe in detail the algorithm.

We use a nearest neighbor algorithm, finding the `K1` nearest neighbors based on a distance measure $d(\cdot)$ that over-weights prominent attributes. Specifically, given a vector of known retailers, RC, containing retailers $RC_1, \ldots, RC_N$ with known formats, the format of an arbitrary retailer $k > N$ is given by:

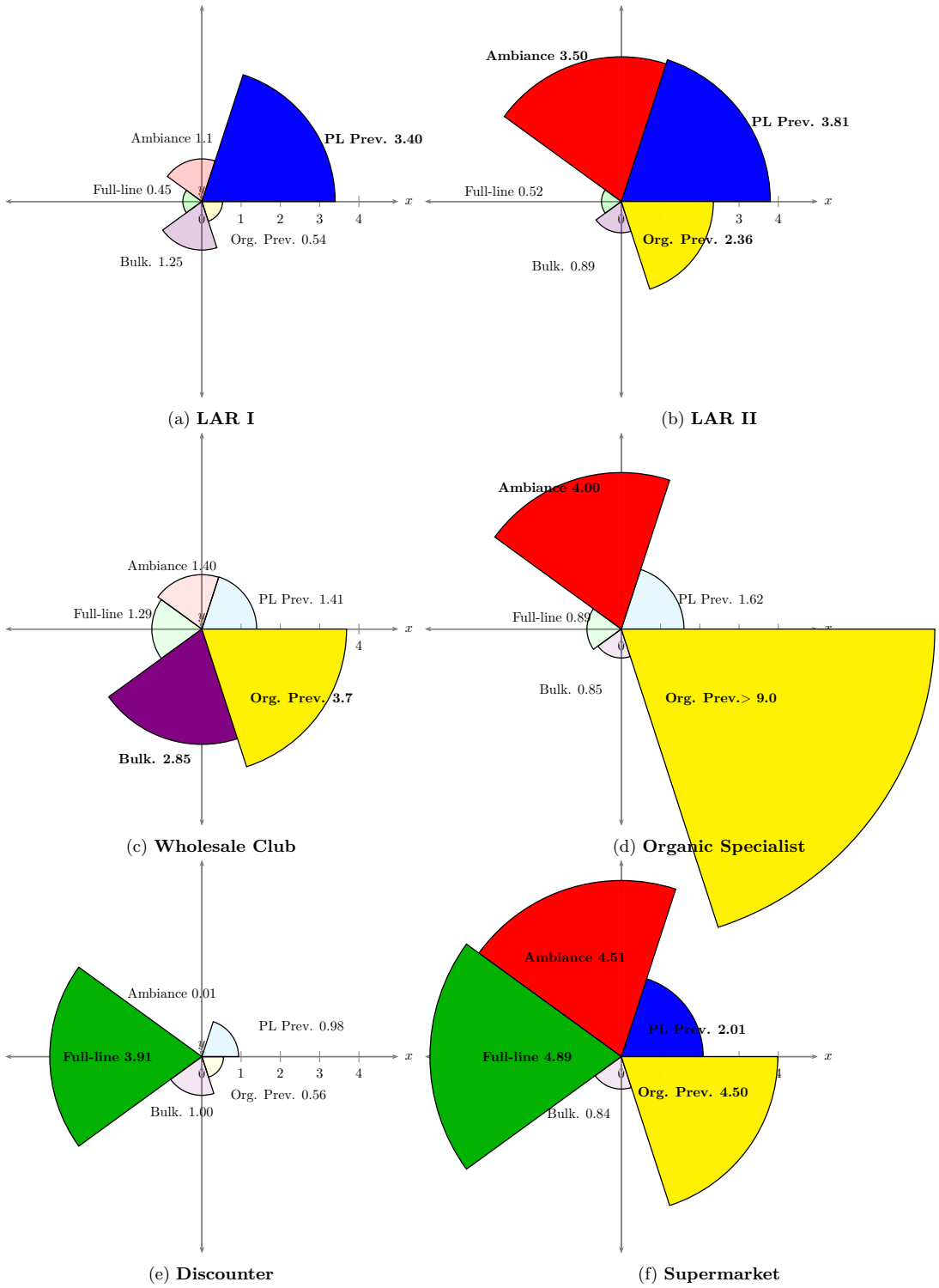
$$K1 - NN_{\|x\|_p}(RC) \triangleq \arg\min \{d(RC_1, RC_k), \cdots, d(RC_N, RC_k)\} \tag{11}$$

with:

$$d(RC_i, RC_j) \triangleq \texttt{abs}(\|\texttt{Attr}_{RC_i}\|_p - \left\|\texttt{Attr}_{RC_j}\right\|_p) \tag{12}$$

and $\|x\|_p$ being the L_p norm which is defined as:

$$\|x\|_p \triangleq (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{\frac{1}{p}}$$



(a) **LAR I**

(b) **LAR II**

(c) **Wholesale Club**

(d) **Organic Specialist**

(e) **Discounter**

(f) **Supermarket**

Figure 2: Polar chart representation of retail formats

where each x_h is a measure of one of the retail format characteristics. The hyper-parameter $p \geq 1$ is chosen by the researcher to impose how much attention is focused on all of the attributes versus a single prominent attribute. Setting $p = 1$ implies equal weights to all the attributes while $p \to \infty$ implies only the most prominent attribute is relevant for retail format determination. We select the retail format of RC_j as that of his closest neighbor among the known retailers, conditional on being closer than κ. If no retailer is within κ, we classify the retailer as a "generic" supermarket.

Intuitively, our classification has two advantages. First, given that it is based on the prominent attributes, we don't need to know their real identity of each retailer and hence can be applied to all of the retailers in the HMS without deidentifying them. Second it is computationally lighter and robust to the inclusion of more attributes, including potentially non-informative attributes. In fact, we can include any number of attributes, but may need to adjust the tuning parameters, p and κ. As a proof of concept, we the method to a subset of the Nielsen HMS data. For instance, we found that we detect 3 limited assortment retailers, 6 organic specialists, and 7 wholesale clubs. These retailers appear to be accurate based on our limited ability to de-identify the retailers. Our intent is to use this classification system in order to leverage it for future research.

4 Inferring causality with synthetic controls using unbalanced and disaggregated data

In this section, we introduce and explain the methodology used to evaluate the causal effects in our study. We first introduce the language and notation of the canonical Rubin casual model (Rubin [45]) and then apply it to study the causal effects induced by the entry of the different types of retailers on the shopping habits of interest. We then present the synthetic controls methodology (Abadie and Gardeazabal [1],Abadie et al. [3], Abadie and l'Hour [2]), introduce our proposed method for handling unbalanced panels and show how to perform inference both for the individual specific and average treatment effects. We conclude the section presenting our adaptation of the mirror descent machine learning algorithm. The mirror descent algorithm is a deterministic proximal method for convex optimization specially designed to work on extremely large scale problems. For our purposes it has the advantages of offering a dimension-free convergence rate (requiring neither to compute nor to invert the hessian of the objective function). Our final implementation of the algorithm is massively "embarassingly" parallelizable.

4.1 Rubin's (1978) classical model of causality

Formally stated, we aim to quantify the effect of a non-reversible and binary treatment on a measurable outcome relying solely on observational data. To that end, we have access to a panel data

set up described by the following tuple.

$$\mathcal{P} \triangleq \left\{T_i, T_i^t, \mathbf{Z}_i, \mathbf{y}_i\right\}_{i\in\mathcal{I}} \tag{13}$$

In equation (13), $\mathcal{I}$ is the set of all the individuals[7], T_i is the set of all the time periods that the individual i inhabits the panel, $T_i^t \subset T_i$ is the set of time periods that the individual i is treated, $\mathbf{Z}_i$ is a vector of K observables, $\mathbf{y}_i$ is a vector of length equal to $|T_i|$ containing the observed values of the variable of interest.[8] In the current example of studying how the entry of a retail format affects private label expenditure, T_i is the set of periods that household i stays in the HMS data set and T_i^t is the subset of these time periods with a newly opened store in her neighborhood. The vector $\mathbf{y}_i$ contains measures of her expenditures on private labeled goods in each time period, and $\mathbf{Z}_i$ contains descriptors of the individual such as demographics of the HMS panelists. The assumption of treatment irreversibility imposes that joining the treatment group is an absorbing event and is formally defined as:

$$\forall s > t \ : \ t \ \in T_i^t \Rightarrow s \in T_i^t \tag{14}$$

The assumption of non-reversible treatment allows us to use convenient $pre-post$ treatment notation, specifically, we let $T_i^{pre} \equiv t \in \left[\underline{T}_i, \underline{T}_i^t\right)$ and $T_i^{post} \equiv \left[\underline{T}_i^t, \overline{T}_i\right]$. For individuals in the panel that are never treated we set $T_i^{pre} = T_i$ and $T_i^{post} = T_{i,t} = \emptyset$.

The fundamental problem of causality states that, a posteriori, for a researcher it is never possible to observe more than one potential outcome. It is common in the literature to make explicit such impossibility by means of an observation equation. Specifically, any $y_{i,t}$ in the vector of observed outcomes $\mathbf{y}_i$ contains:

$$y_{i,t} \equiv \begin{cases} y_{i,t}\left(0\right) & \text{if } \ i \text{ not treated at } t \\ y_{i,t}\left(1\right) & \text{if } \ i \text{ treated at } t \end{cases} \tag{15}$$

In equation (15), the term $y_{i,t}(0)$ (respectively $y_{i,t}(1)$) is the outcome variable had individual been assigned to the control group (respectively, to the treatment group). Under the stable unit treatment value assumption (see Imbens and Rubin), equation (15) can be rewritten as:

$$y_{i,t} = y_{i,t}(0) + \underbrace{\tau_{i,t}}_{y_{i,t}(1)-y_{i,t}(0)} \tag{16}$$

The parameter $\tau_{i,t}$ is the effect of being treated on individual i. We call it an individual treatment effect (**ITE**), since it does not average over a sample. This ITE is the direct analogue to the "case" value for the typical synthetic controls that examines aggregate outcomes. However, our ITE by

[7] Throughout, we refer generically to the unit of analysis as an individual, but note that in our data setting this is actually a household.

[8] It will be convenient to adapt the standard notation to reference elements in ordered sets. Throughout $\underline{T}_i$ and $\overline{T}_i$ represents the min and the max of the set T_i and $|T_i|$ represents its dimensionality

definition will contain measurement error and whenever we evaluate outcomes, we evaluate these as averages over these ITEs. Equation (16) implies that, by construction, any estimate of the **ITE** should be close to zero for any $(i,t)-$pair if untreated.[9] For a specific time period t, we regard as $\mathbf{CATE}_t(\mathbf{Z})$ to be the average treatment effect across the distribution of all the individuals, $F_{i|t,z}$ conditional on demographics $\mathbf{Z}$. Formally, it is defined by

$$\mathbf{CATE}_t(\mathbf{Z}) = \int \tau_{i,t} dF_{i|t,z} \tag{17}$$

Finally, the last layer of aggregation is the ATE is the

$$\mathbf{ATE}_t \equiv \int \tau_{i,t} dF_{i|t} \tag{18}$$

$$\equiv \sum_{\mathbf{Z}} \left[\pi(\mathbf{Z}) \times \mathbf{CATE}_t(\mathbf{Z})\right]. \tag{19}$$

The choice of whether to use $\mathbf{ATE}_t$, $\mathbf{CATE}_t(\mathbf{Z})$ or $\mathbf{ITE}_t$ is determined by the research question and the type of data available. In our case, since we have panel data we are able to identify the ITE.

We conclude the summary of the Rubin's classic model of causality noticing that the problem of estimating the $\tau_{i,t}$ is completely equivalent to the problem of estimating the $y_{i,t}(0)$. The literature on causal estimation using observational data has provided two ways to estimate the $y_{i,t}(0)$. The first is to construct a decision theoretic model of consumer choice, estimate all its structural parameters and then use the calibrated model to predict the behavior of the agent in counterfactual scenarios. The other approach is to use machine-learning, data-driven algorithms to create a prediction about the counteractual case for the treated unit based on the untreated units, and then assume that this function would still prevail as the treated unit receives the effect of the treatment. Synthetic controls takes the latter approach.

4.2 Synthetic controls

Synthetic controls are a blend of the variables of non-treated individuals used to represent the counterfactual, non-treated outcomes for the treated units. As a result, the treatment effects estimated from synthetic controls necessarily are for the treated units, not the general population. We indicate this by regularly referring to the effects as those on new format shoppers, or just shoppers.

Formally, we denote $\boldsymbol{\mathcal{D}}_i$ as the donor pool of individual i, the set of never-treated-individuals observed for the same time periods as the treated unit. We require full coverage in the current

[9] This trivial statement turns out to be relevant to perform causal inference, since it underlies the central concept of the permutation test used.

approach, which is formally described by,

$$\begin{aligned}\boldsymbol{\mathcal{D}}_i \equiv s \in \mathcal{I} \text{ such that:} \\ &(i)\, \underline{T}_s \leq \underline{T}_i \\ &(ii)\ \min\left\{\underline{T}_s^{post}, \overline{T}_s\right\} \geq \overline{T}_i \end{aligned} \tag{20}$$

In equation (20), the first restriction states that for an individual to be assigned to $\boldsymbol{\mathcal{D}}_i$, she has to be in the panel on or before the treated individual. The second restriction states that a treated individual, in practice, can indeed belong to a donor pool of another treated individual as long as the latter receives the treatment after the former has already left the data. Given a donor pool $\boldsymbol{\mathcal{D}}_i$ a synthetic control is completely described by a set of weights, denoted by ω_i, attached to each element in the donor pool.

There are many different potential synthetic controls, each defined by a the ω vector. The set of these potential synthetic controls, $\Delta_{|\boldsymbol{\mathcal{D}}_i|}$ is given by

$$\Delta_{|\boldsymbol{\mathcal{D}}_i|} \equiv w \in \mathbb{R}_+^{|\boldsymbol{\mathcal{D}}_i|} \text{ such that } \sum_{s \in \boldsymbol{\mathcal{D}}_i} w_s = 1. \tag{21}$$

The actual synthetic control described by w is given by $\sum_{s \in \boldsymbol{\mathcal{D}}_i} w_s \cdot x_s$. Valid examples of synthetic controls are $\boldsymbol{\omega}_i = (w_1, \cdots, w_{|\boldsymbol{\mathcal{D}}_i|}) = (0, 0, 1, 0, \cdots)$ and $w = (\frac{1}{N}, \frac{1}{N}, \frac{1}{N}, \frac{1}{N}, \cdots)$. In the former case the actual synthetic control contains all the observables of the third element in the donor pool whereas the latter is just the mean. A synthetic control estimator is just a synthetic control optimally chosen to be as informative as possible about the term $y_{i,t}(0)$.

4.3 Estimation of individual treatment effects using penalized synthetic methods in an unbalanced panel of consumers

The original synthetic controls estimator was designed for balanced panels. For our purposes, we must adapt the methods in Abadie and Gardeazabal [1], Abadie and l'Hour [2] and Doudchenko and Imbens [15] to settings in which households that have stayed in the panel for a different number of time periods are treated by the same event. In the context of our research question, a store opening in period t equally affects a household that has participated in the HMS for ten years as a household that joined the HMS just two years ago. Efficient estimation has to account for the fact that, for the first household, we will have a long time series and a relatively thinner donor pool, whereas, for the second household, we will have a very large donor pool with relatively few time-series observations. We develop an approach to efficiently and automatically chooses the estimation approach based on the data, and specifically, depending on the time-series dimensionality and donor pool characteristics of each treated unit.

In this subsection, we first explain the estimator for a single treated individual, conditional on a penalizing parameter, λ_i. Second, we provide two separate approaches to identify the optimal λ_i and how to choose between them for each treatment individual. Finally, we present how our method can be used to test hypotheses for an individual's causal effect, the ITE, and for the CATE and ATE.

4.3.1 Estimation of Treatment Effects Conditional on λ_i

For each given treated household i, her donor pool, $\boldsymbol{\mathcal{D}}_i$, a vector of pre-treated time periods T_i^{pre}, and the $\mathbf{Z}_i$ we define $X^{\overline{T_i^{pre}}}$ to be the vector containing outcome variables for such pre-periods along with the other relevant individual-level observables:

$$X_i^{T_i^{pre}} \equiv \left[\begin{array}{cccccc} y_{i,\left(\overline{T_i^{pre}}-1\right)} & \cdots & y_{i,\left(\underline{T_i^{pre}}\right)} & Z_{i,1} & \cdots & Z_{i,K} \end{array} \right]'.$$

Similarly, for a given vector of weights, we define the synthetic counterpart of $X_i^{T_i^{pre}}$ as

$$\hat{X}^{T_i^{pre}}\left(\boldsymbol{\omega}_i, \boldsymbol{\mathcal{D}}_i\right) \equiv \sum_{s \in \boldsymbol{\mathcal{D}}_i} w_s X_s^{T_i^{pre}},$$

where X_s is the vector of corresponding X values for the donor pool individual s. For a given λ_i, the λ_i-restricted estimator minimizes the mapping $\boldsymbol{\Omega}_i\left(\boldsymbol{\omega}_i \mid \cdot\right)$ given by

$$\boldsymbol{\Omega}_i\left(\boldsymbol{\omega}_i \mid \lambda_i, \boldsymbol{\mathcal{D}}_i, T_i^{pre}\right) \equiv \left\| X_i^{T_i^{pre}} - \sum_{s \in \boldsymbol{\mathcal{D}}_i} w_s X_s^{T_i^{pre}} \right\|^2 + \lambda_i \sum_{s \in \boldsymbol{\mathcal{D}}_i} w_s \left\| X_i^{T^{pre}{}_i} - X_s^{T_i^{pre}} \right\|^2 \tag{22}$$

Equation (22) is the sum of two penalization terms usually regarded as distance penalizer and regularizing penalizer. The standard synthetic estimator approach in Abadie and Gardeazabal [1] and Abadie et al. [3] only minimizes the distance penalization which equates to the case $\lambda_i = 0$.

However, if the optimal synthetic control is a combination of donors with very different characteristics then interpolation bias can invalidate the estimator. Moreover, this problem can occur even if the objective function for selecting the weights achieves the zero lower bound.[10] Such interpolation bias motivates the inclusion of a regularizing penalization, as pointed out in Doudchenko and Imbens [15], Abadie et al. [3] and Ferman and Pinto [24]. By adding a regularizing penalization in the objective function, the estimation task is no longer focused solely on fitting the pre-treatment cases. Instead, the estimator also needs to ensure that the donors with the most weights are relatively similar to one another. The relative size of the regularizing penalization is increasing in λ_i. In contrast to $\lambda = 0$, if $\lambda \to \infty$, then the weights are degenerate, i.e., all but one donor becomes

[10] This occurs when the focal $X_i^{T_i^{pre}}$ belongs to the convex hull of $\boldsymbol{\mathcal{D}}_i$.

irrelevant (values of 0). Such a degenerate solution is termed the nearest neighbor solution. Thus, the λ_i-restricted estimate is defined as

$$\boldsymbol{\omega}_i^{\star}\left(\lambda_i, \boldsymbol{\mathcal{D}}_i, T_i^{pre}\right) \equiv \arg \min_{\boldsymbol{\omega}_i \in \Delta_{|\boldsymbol{\mathcal{D}}_i|}} \left\{\boldsymbol{\Omega}_i\left(\boldsymbol{\omega}_i \mid \lambda_i, \boldsymbol{\mathcal{D}}_i, T_i^{pre}\right)\right\} \tag{23}$$

and its implied treatment effect is given by:

$$\hat{\tau}_{i,t}\left(\lambda_i, T_i^{pre}, \boldsymbol{\mathcal{D}}_i\right) \equiv y_{i,t} - \sum_{s \in \boldsymbol{\mathcal{D}}_i} w_s^{\star}\left(\lambda_i, T_i^{pre}, \boldsymbol{\mathcal{D}}_i\right) \times y_{s,t} \tag{24}$$

4.3.2 λ_i Calibration

Calibrating the λ_i in an unbalanced panel provides a challenge that has not been discussed yet in the literature. For balanced panels, Doudchenko and Imbens [15] and Abadie and l'Hour [2] suggest calibrating the scale of the regularizing penalization to minimize the estimated treatment effect for the cases where we know there is no active treatment whatsoever. This technique is sometimes referred to as a placebo or pseudo-treatment test and the identifying assumption is that the data on which it is implemented is not treated, which is by design true for the pretreatment periods. For our purposes, there are two potential ways to approach the placebo tests implementation–one oriented around splitting the time-series of the pre-treatment into a training and validation sample and one oriented around using the donors as a pseudo-treated unit. The value of these two approaches depends on the number of pretreatment periods, which implies in an unbalanced sample that it varies by treated unit. We provide an approach to selecting the optimal approach for each treated unit here.

We also propose to extend these ideas in the completed dissertation. In this extension, we will consider the impact of the decision of how many pretreatment periods to include as observations for an individual. To see the direction of the future development, note that the more pretreatment periods an individual is observed, the harder it is to find donors to match that individual, so that the size of the donor pool shrinks and the quality of the matches might shrink as well. Thus, the choice of the pretreatment observation window has implications for the quality of the causal inference. We plan to address this issue in future developments.

λ_i - Calibration Based on Pre-treatment Periods: When the focal household has a long time-series of pretreatment periods, then relatively few untreated households are likely to have been in the panel and overlapped for the full period. As a result, the size of the donor pool is likely to be relatively small. To leverage the longer time-series and provide a better basis to select the λ_i penalization for dissimilarity of the donors, we consider an approach that partitions the pre-treatment period into a training set and validation set. In this way, the validation set can compare

the small number of donors against the actual pre-treatment cases for the treated unit.

Our approach here is to calculate the squared treatment effect in each validation period and average these squared effects. Since these effects should be zero, we construct an objective function that is the sum of these squared errors. One can evaluate this objective function for any subset of the pre-treatment periods being used as the training data (and the rest as the validation data). We average over multiple such subsets and illustrate this objective function with a convex set of the upper limit of a contiguous training period that begins with the first pre-treatment period. We denote $\overline{T}_i^{pre,train}$ and $\underline{T_i}^{pre,train}$ as the lowest and highest upper limit of the training period. Given the assumption that no treatment occurs during the pre-treatment periods, we select the minimizing $\equiv \lambda_{i,ts}^{\star}$ as the value that produces the minimum value,

$$R_{i,ts}^{\star}(\cdot) \equiv \min_{\lambda_i \geq 0} \left\{ \frac{\sum_{s=\underline{T_i}^{pre,train}}^{\overline{T}_i^{pre,train}} \left(\frac{\sum_{s+1}^{\overline{T}_i^{pre}} \left[\hat{\tau}_{i,t}\left(\lambda_i, \left\{ \underline{T_i}^{pre}, \cdots, T_i^s \right\}, \boldsymbol{\mathcal{D}}_i \right) \right]^2}{\left| s+1 : \overline{T}_i^{pre} \right|} \right)}{\left| T_i^{pre,train} \right|} \right\} \tag{25}$$

The idea behind expression (25) is to replicate the estimating procedure on the same treated household against her same donor pool but on a set of periods where we know the treatment did not occur. This method has the advantage of being implemented on the same unit as the final estimation, which should in principle increase the validity the synthetic control. However, to use this approach requires a relatively long pre-treatment time series for the treated individual.

λ_i **- Calibration Based on the Donor Pool:** When the focal household has a relatively few pre-treatment periods, then using the approach above can lead to over-fitting, so that it is usually more conservative to search for the optimal λ_i using a cross-validating task on a pseudo-treated unit from the donor pool. In this case we drop the treated unit and suppose instead that one of the households in the donor pool received the effect of the treatment on the same time period as in the main case. In this way, we preserve the full length of the pre-treatment period for calibration. Since, by assumption, we are certain that households in the donor pool were not treated, we still have a valid placebo treatment effect that should be zero. We select the λ_i to be the one that minimizes the pseudo-treatment effect across individuals in the donor pool. We do by finding the $\lambda_{i,cs}^{\star}$ that produces the minimum value

$$R_{i,cs}^{\star}(\cdot) \equiv \min_{\lambda_i \geq 0} \left\{ \frac{\sum_{j \in \boldsymbol{\mathcal{D}}_i} \left[\hat{\tau}_{j,t}\left(\lambda_i, T_i^{pre}, \boldsymbol{\mathcal{D}}_{i,-j}\right) \right]^2}{\left| \boldsymbol{\mathcal{D}}_i \right|} \right\} \tag{26}$$

where $\boldsymbol{\mathcal{D}}_{i,-j}$ is the donor pool less the treated unit and the pseudo-treated unit. We can combine

both methods efficiently and define our individual specific estimator as:

$$\widehat{\tau}^{\star}_{i,t} = \begin{cases} y_{i,t} - \sum_{j \in \boldsymbol{\mathcal{D}}_i} w^{\star}_i(\lambda^{\star}_{i,ts}, \boldsymbol{\mathcal{D}}_i) \times y_{j,t} & \text{if } R^{\star}_{i,ts} \leq R^{\star}_{i,cs} \\ y_{i,t} - \sum_{j \in \boldsymbol{\mathcal{D}}_i} w^{\star}_i(\lambda^{\star}_{i,cs}, \boldsymbol{\mathcal{D}}_i) \times y_{j,t} & \text{if } R^{\star}_{i,ts} > R^{\star}_{i,cs} \end{cases}$$

For each treated individual, we propose to search for the optimal λ_i using each approach, and to select the best of the two approaches. This will naturally allow us to take advantage of the cases with longer pre-treatment periods, while not over-fitting the cases with fewer pre-treatment periods.

4.4 $\tau_{i,t}$ Inference

To make inferences on the value of the ITE based on $\widehat{\tau}^{\star}_{i,t}$ we have to account for the fact that our estimator is just an approximation to the real value of the ITE. Specifically one needs to be explicit on what are the sources of error of the estimator. In our set up, when estimating ITE for a treated household and conditional on her donor pool, the only source of variation of the estimator stems from its incapacity to perfectly predict the path of the variable of interest had our focal household remained untreated. Under the assumption that such error is independent of the specific treated unit, we can get a measure of how accurate our predictor is by running a permutation test. The intuition behind a permutation test is to estimate the treatment on the cases where we actually know the real value of the treatment and measure how close the prediction is to that value. The test is implemented as follows, we randomly choose S households of the donor pool and estimate the effect of the treatment.

$$\mathbb{V}\text{ar}\left(\hat{\tau}^{\star}_{i,t}\right) \leq \frac{1}{S} \sum_{s=1}^{S} \left(\widehat{\tau}^{\star}_{s,t}\right)^2 \tag{27}$$

We can create error bounds for $\hat{\tau}^{\star}_{i,t}$ that determines the set of values that can assumed to caused by inacuracy of the prediction using $\left(\hat{\tau}^{\star}_{i,t}\right) \pm 2 \times \sqrt{\mathbb{V}\text{ar}\left(\hat{\tau}^{\star}_{i,t}\right)}$ and hence if the zero belongs to the tails of the distribution we will conclude that the ITE is very unlikely to be zero.

4.5 Aggregation

Given that our $\hat{\tau}^{\star}_{i,t}$ is an unbiased estimate for the **ITE**, one valid estimator for the CATE(X) is

$$\widehat{CATE}(X) \equiv \frac{\sum_i \left(\mathbf{1}_{\{X_i = x\}} \widehat{SD\left(\hat{\tau}^{\star}_{i,t}\right)}\right)^{-1} \hat{\tau}^{\star}_{i,t}}{\sum_i \left[\mathbf{1}_{\{X_i = x\}} \left(\left(\widehat{SD\left(\hat{\tau}^{\star}_{i,t}\right)}\right)^{-1}\right)\right]}$$

This estimator is efficient since it assigns more weight to the treatment effects that are more precisely estimated. The problem is that the estimate of its standard deviation does not have a closed form solution. Using a second layer of permutation tests would be computationally infeasible. We instead use the $\widehat{CATE}(X)$ to calculate

$$\sqrt{\widehat{\mathbb{V}\text{ar}}\left(\widehat{CATE}(X)\right)} \leq \frac{1}{N} \max\left[\widehat{SD\left(\hat{\tau}^{\star}_{1,t}\right)}, \cdots, \widehat{SD\left(\hat{\tau}^{\star}_{N,t}\right)}\right]. \tag{28}$$

4.6 Estimation algorithm

The task of estimating individual treatment effects and their standard deviations is computationally burdensome for two reasons. First, we have thousands of households, from whom we aim to recover long vectors of estimated treatment effects (one for each time period) and for multiple variables. This problem can be mitigated using embarrassingly parallel algorithms on super computers. Given that our estimates are household specific, our program meets this condition, and so we leverage massive parallelization to accomplish the computational challenge.

However, even this is insufficient. The second problem is that, even for a single household, we have to repeatedly solve a difficulty convex optimization with thousands of constraints. Specifically, we need to find a way to solve the following program as fast as possible:

$$\begin{aligned}
\min_{\mathbf{w}} \left\| X_i^{T_i^{pre}} - \sum_{s \in D_i} w_s X_s^{T_i^{pre}} \right\|^2 + \lambda_i \sum_{s \in D} w_s \left\| X_i^{T^{pre_i}} - X_s^{T_i^{pre}} \right\|^2 \\
\text{subject to: } \sum_{w \in D_i} w_i = 1 \\
w_1 > 0 \\
\vdots \\
w_{|D_i|} > 0
\end{aligned}$$

. Using unconstrained optimizing techniques and imposing the restrictions by means of functional forms does not work.[11] A more practical but still sub-optimal approach is to recast the problem as quadratic programming problem and use a optimized software, for instance, that available in R. This is what Abadie et al. [3] use in their package `synth`, which relies on another `R` package `kernlab`. The original method was intended for relatively small donor pools and a single treated case, and this approach does not appear to work at he scale we need (thousands of donors and treated units).

[11] We note that we considered other approaches and each has limitations. For example, Nelder-Mead converges very slowly with so many variables and the kinds of functional forms required, BFGS requires approximating the hessian and inverting the matrix, which in our case would mean inverting a matrix of 4,000 by 4,000 elements, gradient descent methods are not valid because the problem does not operate in Hilbert vector space but in a Banach space. Finally, our experience is that stochastic optimization does not perform well in this context.

As a result, we identified an algorithm developed and used for machine learning problems. Specifically, we borrow from the literature on machine learning methods and adapt the mirror descent algorithm (Nesterov et al. [39]). The mirror descent algorithm has the advantage of offering a dimension-free convergence to a local optima. The idea behind the algorithm is to break the optimization problem into a sequence of local easy-to-solve optimization problems yielding (potentially) in-feasible solutions. Such (potentially) in-feasible solutions are projected into the space of feasible values. The algorithm is described by the following:

$$\min_{x \in X} f(x) \tag{29}$$

start with $x_1 = x \in X$ and $\gamma_1 = 1$ at iteration t:

$$\kappa_{t+1} = \texttt{Prox}_{x_t}(\gamma_t \cdot \nabla f(x_t)) \tag{30}$$

with:

$$\nabla f(x_t) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \tag{31}$$

$$\texttt{Prox}_{x_t}(\xi) = \frac{1}{\left[\sum_i x_i \exp -\xi_i\right]^{-1}} \left[x_1 \exp\left(-\xi_1\right), \cdots, x_n \exp\left(-\xi_n\right)\right] \tag{32}$$

set:[12]

$$x_{t+1} = \begin{cases} \kappa_{t+1} & \text{if } f\left(x_{t+1}\right) - f\left(x_t\right) > 0 \\ x_t & \text{if } f\left(x_{t+1}\right) - f\left(x_t\right) < 0 \end{cases} \tag{33}$$

$$\gamma_{t+1} = \begin{cases} \gamma_t & \text{if } f\left(x_{t+1}\right) - f\left(x_t\right) > 0 \\ 0.98 \times \gamma_t & \text{if } f\left(x_{t+1}\right) - f\left(x_t\right) < 0 \end{cases} \tag{34}$$

Importantly, if the function has a unique local min then we can show that in iteration T the error, i.e. the difference between the solution to (29) and the function f evaluated at x^T is bounded by:

$$f(x^T) - \min_{x \in X} f(x) \leq \sqrt{\frac{2\ln(n)}{T}} \times L_{\|\cdot\|_1}(f), \tag{35}$$

meaning that for each T, we can ensure to be as far as $\varepsilon \triangleq \sqrt{\frac{2\ln(n)}{T}} \times L_{\|\cdot\|_1}(f)$ from the local optima.

[12] In our case the gradient admits a closed form solution given by $\nabla f(x_t) = \left(-2 \cdot \mathbf{X}'\mathbf{X}^{Donor} + 2\omega\mathbf{X}'\mathbf{X}^{Donor}\omega + \lambda\left(\mathbf{X} \otimes \texttt{rbind(1)} - \mathbf{X}^{Donor}\right)\right)$

5 Results

In this section we present the estimates of the treatment effects and discuss their implications. We organize our discussion into two parts. In section 5.1 we discuss the average treatment effects. For these main results, we present the results for each shopping behavioral variable, linking the results to the retail format characteristics. In section 5.2, we present the conditional average treatment effects, where we break the treated cases into the single-homing and multi-homing groups. We focus that discussion around the overall pattern of results for each of these groups. The main average treatment effects are presented in table 1 and the conditional average treatment effects are presented in table 2.

5.1 Average Treatment Effects on the Treated

Retailers visited: Our results indicate that starting to shop in **LAR - I**, **LAR - II**, **Club** or **Org**. increases the size of the set of retailers where consumers shop. The effect sizes are somewhat larger for **Club** and **Org**., but these differences are not statistically distinguishable. Intuitively, this result suggests that households are effectively adding **LAR - I**, **LAR - II**, **Club**, and **Org**. to the list of stores that they frequently visit. Each of these stores offers a large degree of unique assortments (private label and/or organic prevalence), ambience, and/or convenience, each of which are predicted to potentially lead the retailer to be added to the set of retailers visited. In contrast, starting to shop at **Disc**. does not alter the size of the set of visited retailers, which indicates that consumers substitute **Disc**. for a retailer that they previously visited for groceries. This is consistent with consumers seeking out lower prices and replacing an existing traditional grocery supermarket with this discounter. Finally, our results indicate that **Sup**. is the only retailer that has the capacity to shrink the set of retailers that the consumers visit. In the year after **Sup**. entered their local market, out of the 156 panelists that started shopping there, 131 of them have negative estimated treatment effects on the size of their shopping set (97 of the treated households received such effect was significant at the 95% confidence level). These means that 86% of the households that started to shop at **Sup**., also decided to replace more than one of the retailers in the consumers shopping set. This phenomena suggests that incorporating the full-line coverage with high ambience and organic prevalence might not only attract customers but allow them to meet more of their shopping needs in a single location. In contrast, the convenience and ambience provided by smaller stores such as **LAR - I**, **LAR - II** or **Org**. suffices to attract new consumers but their assortment is not wide enough to cover all the necessities of the average consumer. The same is true for **Club**. Our results suggest that the assortment of **Club**, possibly because it's distinctive prominent attributes (bulkiness) is not capable of fulfilling all the needs of the consumers. This result also puts a new perspective on the law of retail gravitation (see section 3), where full-line isn't necessary to attract consumers, but rather to serve as a replacement for another retailer or displacement of multiple retailers.

Table 1: Average treatment effects in the year after entering

	LAR - I		**LAR - II**		**Org.**	
	ATE	S.E.	ATE	S.E.	ATE	S.E.
Shopping Behaviors						
Retailers visited	**0.104**	**0.037**	**0.139**	**0.041**	**0.16**	**0.053**
Shopping trips	**0.482**	**0.174**	0.400	0.22	**1.87**	**0.897**
Items	0.048	1.183	**0.679**	**0.225**	**1.43**	**0.614**
Expenditure	0.34	2.011	**1.71**	**0.325**	**1.65**	**0.64**
Private label						
Expenditure	**1.439**	**0.238**	**1.007**	**0.259**	0.581	0.339
Share	**0.117**	**0.049**	**0.201**	**0.078**	0.022	0.073
Items	**0.926**	**0.408**	**0.213**	**0.038**	0.399	0.484

	Sup.		**Club**		**Disc.**	
	ATE	S.E.	ATE	S.E.	ATE	S.E.
Shopping Behaviors						
Retailers visited	**-0.084**	**0.039**	**0.178**	**0.063**	0.049	0.061
Shopping trips	0.871	2.011	**-0.292**	**0.110**	1.365	1.105
Items	-0.282	4.126	**-1.372**	**0.142**	1.565	2.014
Expenditure	1.77	2.317	**1.87**	**0.870**	0.87	1.17
Private label						
Expenditure	**4.869**	**1.893**	**1.761**	**0.432**	0.487	0.366
Share	**5.185**	**2.061**	**0.22**	**0.070**	-0.030	0.075
Items	**2.932**	**1.324**	0.092	0.685	0.447	0.684

Bolded items represented permutation test indicates significant effects.

Shopping frequency: The second measure we consider is the number of shopping trips per month. We find that the point estimates for all retail format entry types are positive except **Club**. Of the positive effects, **Org**. and **LAR - I** significantly increase the consumers' shopping frequency, and **LAR - II** increases it at the 90% confidence level. When the median household starts shopping at **Org**. she increases by 25% her grocery shopping frequency. We argue this is likely due to the ambience of the store, which includes the fresh and prepared foods. Shoppers of **LAR - I** also increase the shopping frequency, but the median shopper increases only by 9%. The increase for **LAR - I** shoppers seems more likely due to the convenience of shopping at this new store that leads to adding shopping trips as the retailer is added. In contrast, **Disc**., **LAR - II**, and **Sup**. do not increase the shopping trips at traditional significance levels, though the magnitudes are fairly large with large standard errors rather than a tightly estimated zero. Taken as a tightly estimated zero, the result would imply that when adding the retailer to the set, the household replaces trips to other retailers with ones to the new retailer. However, our later results suggest that for **LAR - II**and **Sup**., the null effect arises from meaningful heterogeneity. In contrast to these positive and null effects, **Club** presents a very different picture. Consumers that start to shop at **Club** significantly reduce their shopping frequency. On average, households starting to shop at **Club** make 3.75% fewer trips per month. After becoming a **Club** shopper, the household makes 4 fewer grocery shopping trips per year. This result is consistent with the bulk offerings that are a prominent feature of **Club** leading households to start stockpiling through bulk purchases.

Number of items: The effects of becoming new format shoppers on the number of items purchased is similar to that of the shopping trips–**LAR - II** and **Org**. have positive significant effects, **LAR - I** and **Disc**. have insignificant effects with large standard errors, and **Club** has a negative effect. The significant positive effect for **Org**. is more than twice the size of the effect for **LAR - II** whereas the significant negative effect for **Club** is similarly sized to that of **Org**., but negative. Considering first the significant positive results, **LAR - II** and **Org**. have unique assortments (private label or organic prevalence) and strong ambience, which for **Org**. in particular includes fresh and prepared foods. Again, we find that the insignificant effects actually have large standard errors reflective of heterogeneity we address in the next section. The significant negative effect for starting to shop at **Club** is consistent with the same stockpiling behavior as we saw with the decreased shopping frequency. Thus, the bulk offerings of **Club** seems to provide a clear pattern of stockpiling across multiple measures. Interestingly, although households could previously stockpile by purchasing multiple items per trip, the bulk offerings actually reduce the shopping frequency, suggesting changes in stockpiling behaviors.

Expenditures: The estimates of the effects of shopping at **LAR - II**, **Org**. and **Club** on the total monthly grocery expenditure are positive, statistically significant and similar in magnitude. In addition, the effect size for **Sup**. is also similar in magnitude, but the standard error is larger, leading

to the insignificant effect. For **LAR - II** and **Org**. the pattern of results for total expenditures is similar to that of the number of items, and again consistent with the ambience and the unique assortments related to organic prevalence, suggesting that these variables may in particular be able to expand demand for new format shoppers. Interestingly, shoppers at **Club** increases expenditure while reducing the number of items, suggesting that they consume more through bulk purchases. Even though stockpiling brought on by bulk offerings reduces shopping frequency and the number of items, it leads to an increase in consumption. This increase in consumption is fairly broad with **Club Org**. and **LAR - II**, respectively, having 92%, 87%, and 90% of the new format shoppers increasing their expenditures on groceries. In contrast, **Disc**. does not statistically increase total expenditure. When combined with the insignificant effects for retailers visited, shopping frequency, and number of items, this suggests that this format largely substitutes for existing purchases and shopping rather than expanding demand or altering the way that households shop.

Private labels: Starting to shop at **LAR - I**, **LAR - II**, **Sup**. and **Club** significantly increases private label use, whereas **Org**. and **Disc**. do not. The positive and significant results indicate that the private label prevalence of these stores is particularly attractive, either because these offerings are less expensive or because they offer unique assortments. **Club** presents a slightly different pattern where the private label expenditure and share increase, but the items do not. This pattern is consistent with the strong private label offering combined with the bulk offerings that lead to stockpiling and lower total number of items purchased. As in previous results, the **Disc**. effects are not significant, again suggesting this store operates purely through substitution rather than changing fundamentally the way shoppers behave. Perhaps as interesting from these results is the lack of significant results for the **Org**. case. This retailer does change behavior in terms of total items, shopping frequency, and increasing the size of the set of retailers visited, but does not significantly increase the private label use. This might be consistent with private label offerings not covering as much of the needs in the organic domain, the most prominent attribute of this retail format. With the exception of **Org**., these results suggest that new formats are successfully using private labels to shift consumption towards products only available in their stores. The results for **Disc**. further supports this finding, suggesting that traditional formats may have not generated the same kind of differentiation on through their private label products.

5.2 Conditional Treatment Effects on the Treated for Single and Multi-homing Shoppers

In section 3.1.1, we argued that single-homing households would behave differently in response to new format entry than multi-homing households due to the time budget constraints that single-homing represents. We now examine to what extent the causal effects we obtain are consistent with the predictions of those arguments.

Before we examine the results, recall from section 3.1.1 that we define single-homing and multi-homing as relating to the central tendency for a household's number of unique retailers visited per month. We formally define this as

$$\texttt{n\_RC}^{pre} \triangleq \texttt{mode}\left(n\_RC_{t-1}, \cdots, n\_RC_{t-12}\right) \tag{36}$$

where `mode` is the mode operator and n_RC_h is the number of retailers visited in month h. In words, $\texttt{n\_RC}^{pre}$ is the number of retailers visited per month that most often occurs during the previous year before being treated. We consider a household to be single-homing if $\texttt{n\_RC}^{pre} \leq 2$ and multi-homing otherwise. This definition follows the notion that single-homing households will be more (time) budget constrained and that multi-homing households will be accustomed to shopping at multiple retailers in order to obtain the variety and prices they seek. We define the theoretical treatment effect conditional on being a single ($Single$) or multi-homing ($Multi$) household as

$$\mathbb{E}\left[\tau_t \mid \texttt{n\_RC}^{pre} \in \{Single = (0,1,2), Multi = (3,4,5,6)\}\right] \tag{37}$$

Equation (37) is the definition of the average treatment effect at a specific time period for households exhibiting low or high levels of cross-shopping.

The results are presented in table 2. We first discuss the multi-homing shoppers and then turn to the single-homing shoppers.

Multi-homing shoppers Multi-homing shoppers starting to shop at **LAR - I**, **LAR - II**, **Org.**, and **Sup.** follow the same general pattern as the average treatment effects, but have treatment effects that are generally slightly stronger than the average treatment effects. This suggests that the influence of ambience, unique assortments (private label and organic prevalence), convenience, and full-line offerings match with the previous findings for multi-homers. The effects for **LAR - I** shows one exception, which is that now the effect of the number of items is positive and significant, suggesting multi-homers' variety-seeking may lead them to seek out the unique assortments provided by **LAR - I** or that the prices are lower so that more items are purchased as an income effect.

The multi-homing household results for **Club** and **Disc.** differ more markedly. For **Disc.** the effects are actually generally consistent with the results for **LAR - I**, **LAR - II**, and **Org.**, where the number of retailers visited in a month, shopping visits, and items purchased all increase. This suggests that part of the reason that we find no significant effects for **Disc.** in the average treatment effects is due to heterogeneity. For **Club** the effects are largely different from the average treatment effects. Mutli-homers have a larger increase in the retailers visited suggesting multi-homers are more likely add **Club** to their shopping set rather than replacing an existing retailer, consistent with a desire for unique assortments. Consistent with this interpretation, multi-homers do not decrease the shopping frequency or number of items purchased. Both of those effects are explained by purchasing

bulk products as a replacement for multiple smaller items in order to stockpile. Further, multi-homers increase the total expenditure, and do so while increasing the private label expenditures and items purchased. This overall pattern of results is highly consistent with stockpiling that is offset by the increased expenditure and items consumed as a result of seeking the unique assortment (private label and bulk offerings) this retail format provides. Thus, multi-homing households appear to increase consumption, while also increasing bulk purchasing, and do so by adding **Club** to the set of retailers, but substituting trips to the other retailers for trips to **Club**Ȯverall, these results are consistent with the arguments made in section 3.1.1 and suggest that for multi-homers they take advantage of the prominent features of the new retail formats.

Single-shoppers For households that were single-homing prior to the entry of the new retail format, we find a starkly different pattern from the average treatment effects. First, for **LAR - I** **LAR - II** and **Org**. the pattern of response is muted in terms of the increase in retailers visited, shopping frequency, and items purchased. The muted effects lead to much smaller or insignificant effects. For example, **LAR - II** entry does not increase the number of unique retailers for shoppers, indicating that for these single-homing households, they replace an existing retailer with **LAR - I**İn fact, these smaller effects become insignificant effects in almost all cases with the main exception being **Org**. where the effects are still positive and significant, but smaller than the average treatment effects. This pattern of results is consistent with the more time constrained, less variety-seeking nature of the single-homing households. These group traits lead this group to not take advantage of the unique assortments, ambience, and in-store convenience that change the way multi-homers shop. Further, the single-homing households in general do not increase their total expenditures significantly in response to new entry formats. This finding is consistent with these households being more budget constrained.

Org. though following the muted pattern, still has strong enough effects for shopping frequency and items that even single-homers appear to seek the unique organic prevalence and ambience. Similarly, unlike the multi-homers, single-homers increase their shopping frequency and items purchased for the new **Sup**. shoppers. Interpreting the ambience as including fresh and prepared foods, this finding suggests a willingness on the part of single-homers to visit more often to take advantage of these prepared foods, substituting from the outside option of restaurants. Since both **Org**. and **Sup**. demonstrate this response and both have high ambience, this provides some consistency. However, for this group, the value is substituting for other products since total expenditures does not increase. Recall also that this pattern is very different from the multi-homers shopping for the first time at **Sup**.İn contrast, that group reduced their retailers visited, did not significantly change their shopping frequency or number of items purchased.

Overall, these findings suggest that single-homing and multi-homing households do not respond similarly to the same retail format entries. The response differences are relatively sharp

and range from muted effects for the single-homers to responses that even more directly reflect the single-homing constraints such as an increase in use of stockpiling. These results elaborate the main results and because the pattern provides a consistent picture of the role of the prominent attributes, these results provide further support for the proposed framework and the predictions that come out of that framework.

6 Conclusion

In this research, we study the causal effects of new retail platform entries on household shopping behavior. We present a conceptual framework that solves the challenge of organizing this complex setting. Our framework considers the prominent attributes of retailers and makes predictions for how these attributes will influence shopping behaviors. We then collect a unique dataset the combines household-level panels data on purchases and store openings. Finally, we conduct causal inference by developing and applying a modification of the synthetic controls methods. Our results suggest a rich pattern of causal effects of these retail platforms on household shopping behaviors.

Table 2: Conditional Average Treatment Effects Conditioning on Single-Homers vs. Multi-Homers

	retailers visit		**trips**		**items**		**Total exp**		**PL exp**		**PL share**		**PL items**	
	ATE	**s.e.**	**ATE**	**s.e.**	**ATE**	**s.e.**	**ATE**	**s.e.**	**ATE**	**s.e.**	**ATE**	**s.e.**	**ATE**	**s.e.**
LAR - I														
single	0.035	0.009	0.023	0.155	-0.235	0.295	0.258	0.546	0.770	0.218	0.089	0.878	0.931	0.107
multiple	0.138	0.410	0.531	0.156	1.471	0.304	0.436	0.461	1.387	0.200	1.337	0.652	0.771	0.102
LAR - II														
single	0.008	0.008	-0.147	0.161	-0.194	0.298	-0.035	0.553	-0.602	0.195	-0.42	0.774	-0.149	0.088
multiple	0.194	0.012	0.588	0.181	1.370	0.332	1.198	0.536	1.962	0.236	2.920	0.752	0.583	0.102
Org.														
single	0.024	0.011	0.446	0.203	1.510	0.377	0.873	0.670	1.263	0.260	2.078	0.998	0.286	0.308
multiple	0.176	0.015	3.46	1.351	2.092	0.429	1.775	0.662	0.946	0.310	0.508	0.987	0.412	0.425
Sup.														
single	0.057	0.033	0.94	0.330	1.064	0.737	1.418	1.457	5.714	0.559	10.549	2.269	1.759	1.237
multiple	-0.095	0.041	0.194	0.610	-1.152	1.246	1.961	1.981	4.364	0.969	4.686	2.945	4.038	1.393
Club														
single	0.024	0.011	-0.796	0.285	-2.808	0.453	0.300	1.514	0.013	0.340	0.052	1.345	-0.553	0.154
multiple	0.200	0.018	-0.013	0.295	0.386	0.607	2.141	0.794	2.067	0.379	1.903	1.193	0.831	0.198
Disc.														
single	-0.021	0.010	0.562	0.220	0.339	0.408	0.715	0.836	-0.244	0.257	-1.135	0.963	0.082	0.119
multiple	0.130	0.020	1.765	0.328	3.106	0.603	1.220	0.916	1.565	0.378	1.506	1.167	0.683	0.443

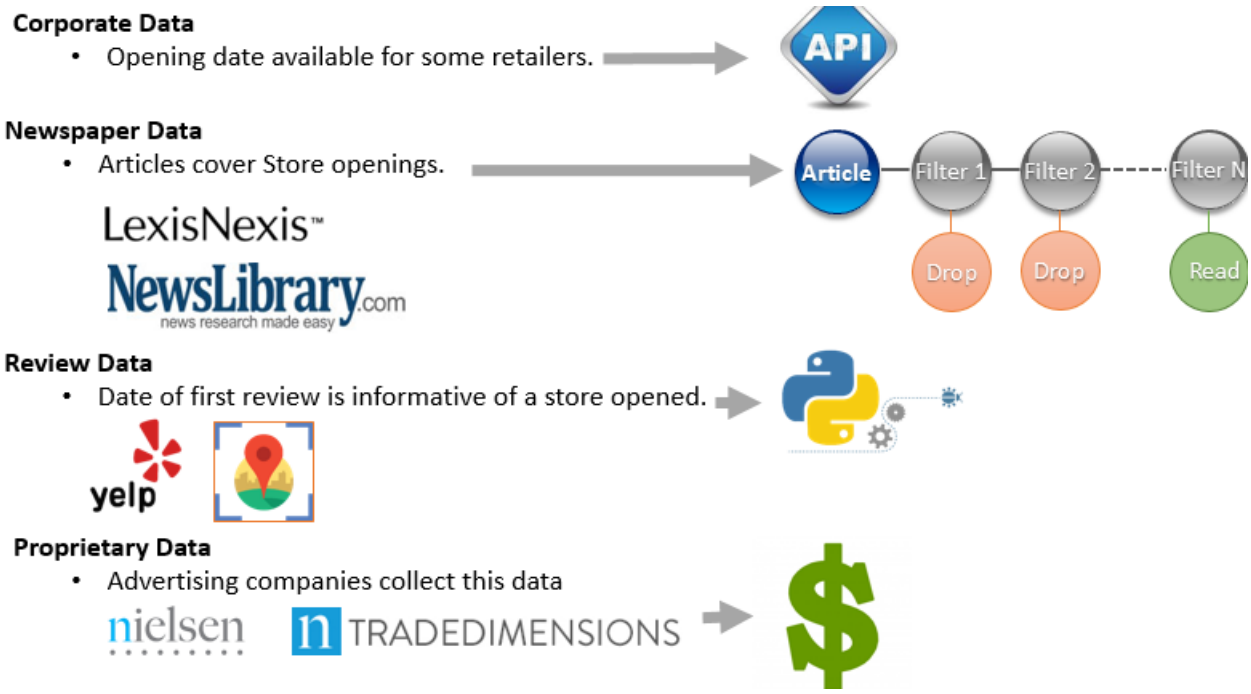


Figure 3: Source of Data

References

[1] Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *The American Economic Review*, 93(1):113–132, 2003.

[2] Alberto Abadie and Jeremy l'Hour. A penalized synthetic control estimator for disaggregated data. *Working Paper*, 93(1):113–132, 2017.

[3] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

[4] David R Bell and James M Lattin. Shopping behavior and consumer preference for store price format: Why 'large basket' shoppers prefer edlp. *Marketing Science*, 17(1):66–88, 1998.

[5] Richard A Briesch, Pradeep K Chintagunta, and Edward J Fox. How does assortment affect grocery store choice? *Journal of Marketing Research*, 46(2):176–189, 2009.

[6] Richard A Briesch, William R Dillon, and Edward J Fox. Category positioning and store choice: The role of destination categories. *Marketing Science*, 32(3):488–509, 2013.

[7] Susan M Broniarczyk and Wayne D Hoyer. Retail assortment: More$\neq$ better. In *Retailing in the 21st Century*, pages 225–238. Springer, 2006.

[8] Susan M Broniarczyk, Wayne D Hoyer, and Leigh McAlister. Consumers' perceptions of the assortment offered in a grocery category: The impact of item reduction. *Journal of marketing research*, pages 166–176, 1998.

[9] Bart J Bronnenberg. The provision of convenience and variety by the market. *The RAND Journal of Economics*, 46(3):480–498, 2015.

[10] Stephen Brown. *Retail location theory: The legacy of Harold Hotelling*. 1989.

[11] William Minseuk Cha, Pradeep K Chintagunta, and Sanjay K Dhar. Food purchases during the great recession. 2015.

[12] Chaoqun Cheng. Competition among retail formats. 2016.

[13] Pradeep K Chintagunta, Andre Bonfrer, and Inseong Song. Investigating the effects of store-brand introduction on retailer demand and pricing behavior. *Management Science*, 48(10): 1242–1267, 2002.

[14] Ap Dijksterhuis, Pamela K Smith, Rick B Van Baaren, and Daniel HJ Wigboldus. The unconscious consumer: Effects of environment on consumer behavior. *Journal of Consumer Psychology*, 15(3):193–202, 2005.

[15] Nikolay Doudchenko and Guido W. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Working Paper 22791, National Bureau of Economic Research, October 2016. URL http://www.nber.org/papers/w22791.

[16] Susan P Douglas. Cross-national comparisons and consumer stereotypes: A case study of working and non-working wives in the us and france. *The Journal of Consumer Research*, 3(1): 12–20, 1976.

[17] Xavier Dreze, Stephen J Hoch, and Mary E Purk. Shelf management and space elasticity. *Journal of Retailing*, 70(4):301–326, 1994.

[18] Jean-Pierre Dubé, Günter J Hitsch, and Peter E Rossi. Income and wealth e ects on private-label demand: Evidence from the great recession. 2016.

[19] Laurette Dubé, Jean-Charles Chebat, and Sylvie Morin. The effects of background music on consumers' desire to affiliate in buyer-seller interactions. *Psychology & Marketing*, 12(4):305–319, 1995.

[20] David Dunning. Self-image motives and consumer behavior: how sacrosanct self-beliefs sway preferences in the marketplace. *Journal of Consumer Psychology*, 17(4):237–249, 2007.

[21] Paul B Ellickson and Sanjog Misra. Supermarket pricing strategies. *Marketing science*, 27(5): 811–828, 2008.

[22] Paul B Ellickson and Sanjog Misra. Enriching interactions: Incorporating outcome data into static discrete games. *Quantitative Marketing and Economics*, 10(1):1–26, 2012.

[23] PB Ellickson, P Kong, and MJ Lovett. Private labels and retailer profitability: Bilateral bargaining in the grocery channel. Technical report, Working Paper, 2017.

[24] Bruno Ferman and Cristine Pinto. Placebo tests for synthetic controls. 2017.

[25] Edward J Fox, Alan L Montgomery, and Leonard M Lodish. Consumer shopping and spending across retail formats. *The Journal of Business*, 77(S2):S25–S60, 2004.

[26] Stephen J Hoch and Shumeet Banerji. When do private labels succeed? *Sloan management review*, 34(4):57, 1993.

[27] Stephen J Hoch and Leonard M Lodish. Store brands and category management. *The Wharton School, University of Pennsylvania*, pages 1–38, 1998.

[28] Thomas J Holmes. The diffusion of wal-mart and economies of density. *Econometrica*, 79(1): 253–302, 2011.

[29] Yufeng Huang and Bart J Bronnenberg. Pennies for your thoughts: Costly product consideration and purchase quantity thresholds. 2015.

[30] Barbara E Kahn. Dynamic relationships with customers: High-variety strategies. *Journal of the Academy of Marketing Science*, 26(1):45–53, 1998.

[31] Botond Kőszegi and Adam Szeidl. A model of focusing in economic choice. *The Quarterly journal of economics*, 128(1):53–104, 2012.

[32] Lien Lamey, Barbara Deleersnyder, Marnik G Dekimpe, and Jan-Benedict EM Steenkamp. How business cycles contribute to private-label success: Evidence from the united states and europe. *Journal of Marketing*, 71(1):1–15, 2007.

[33] Michael Levy, Barton A Weitz, and Dhruv Grewal. *Retailing management*, volume 6. McGraw-Hill/Irwin New York, 2012.

[34] Leigh McAlister and Edgar Pessemier. Variety seeking behavior: An interdisciplinary review. *Journal of Consumer research*, 9(3):311–322, 1982.

[35] Albert Mehrabian and James A Russell. A measure of arousal seeking tendency. *Environment and Behavior*, 5(3):315, 1973.

[36] Albert Mehrabian and James A Russell. *An approach to environmental psychology.* the MIT Press, 1974.

[37] Satya Menon and Barbara E Kahn. The impact of context on variety seeking in product choices. *Journal of Consumer Research*, 22(3):285–295, 1995.

[38] Sergio Meza and K Sudhir. Do private labels increase retailer bargaining power? *Quantitative Marketing and Economics*, 8(3):333–363, 2010.

[39] Yurii Nesterov et al. Gradient methods for minimizing composite objective function, 2007.

[40] Aviv Nevo and Arlene Wong. The elasticity of substitution between time and market goods: Evidence from the great recession. Technical report, National Bureau of Economic Research, 2015.

[41] Koen Pauwels and Shuba Srinivasan. Who benefits from store brand entry? *Marketing Science*, 23(3):364–390, 2004.

[42] Lynn Petrak. Outlook for grocery prepared foods remains strong. *Progressive Grocer*, 2017. Accessed: 2018-06-27.

[43] William J. Reilly. The law of retail gravitation. new york: Knickerbocker press. *New York: Knickerbocker Press*, 1931.

[44] Donovan Robert and Rossiter John. Store atmosphere: an environmental psychology approach. *Journal of retailing*, 58(1):34–57, 1982.

[45] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

[46] Derek D Rucker, Adam D Galinsky, and David Dubois. Power and consumer behavior: How power shapes who and what consumers value. *Journal of Consumer Psychology*, 22(3):352–368, 2012.

[47] Elaine Sherman, Anil Mathur, and Ruth Belk Smith. Store environment and consumer purchase behavior: mediating role of consumer emotions. *Psychology and Marketing*, 14(4):361–378, 1997.

[48] Jillian C Sweeney and Fiona Wyber. The role of cognitions and emotions in the music-approach-avoidance behavior relationship. *Journal of services marketing*, 16(1):51–69, 2002.

[49] Øyvind Thomassen, Howard Smith, Stephan Seiler, and Pasquale Schiraldi. Multi-category competition and market power: a model of supermarket pricing. *American Economic Review*, 107(8):2308–51, 2017.

[50] Richard F Yalch and Eric R Spangenberg. The effects of music in a retail setting on real and perceived shopping times. *Journal of business Research*, 49(2):139–147, 2000.

[51] Changjo Yoo, Jonghee Park, and Deborah J MacInnis. Effects of store characteristics and in-store emotional experiences on store attitude. *Journal of Business Research*, 42(3):253–263, 1998.

[52] Joachim Zentes, Dirk Morschett, and Hanna Schramm-Klein. *Strategic retail management*. Springer, 2007.

[53] Fanyin Zheng. Spatial competition and preemptive entry in the discount retail industry. 2016.

[54] Yi Zhu and Anthony Dukes. Prominent attributes under limited attention. *Marketing Science*, 36(5):683–698, 2017.