Retail Markups, Misallocation, and Store Variety Across U.S. Cities* 

Colin J. Hottman

Board of Governors of the Federal Reserve System†

May 25, 2017

Abstract

I develop a structural model of consumer demand and oligopolistic retail competition to study three mechanisms through which the retail sector affects the attractiveness of cities and consumer welfare: the average retail markup in a location; misallocation from the dispersion in retail markups in a location; and the variety of retail stores available in a location. To quantify the importance of these mechanisms, I estimate my model using detailed retail scanner data with prices and sales at the barcode level from thousands of stores across the United States. I estimate that retail markups are, on average, smaller and exhibit less dispersion in larger cities and that retail markets such as New York City and Los Angeles are close to the undistorted monopolistically competitive limit. However, I find that the distortion and consumption misallocation from retail markups can be economically significant. Furthermore, I show that retail store variety significantly affects the cost of living such that the consumer price index is lower in larger cities and that this difference in prices could be an important consumption-based agglomeration force.

JEL CLASSIFICATION: D12, D43, D61, L81, R32
KEYWORDS: markups, variety, price index, misallocation, retail, cities

*I am grateful to Jeffrey Campbell, Don Davis, Keshav Dogra, Gilles Duranton, Etienne Gagnon, Juan Carlos Hallak, Jessie Handbury, Daniel Hosken, Amit Khandelwal, John Morrow, Devesh Raval, Michael Riordan, John Rogers, Pasquale Schiraldi, Patrick Sun, John Van Reenen, Jonathan Vogel, David Weinstein, and seminar participants at Columbia University, University of Pennsylvania, London School of Economics, Wharton, University of Colorado-Boulder, University of Connecticut, University of Massachusetts-Amherst, Johns Hopkins SAIS, the Federal Trade Commission, USDA ERS, Census CES, and the Federal Reserve Board for helpful comments. All remaining errors are mine. All results are calculated based on data from the Nielsen Company (U.S.), LLC, and provided by the Marketing Data Center at the University of Chicago Booth School of Business. These data were accessed while the author was a doctoral student at Columbia University. The views expressed in this paper are solely that of the author and do not represent the views of the Board of Governors of the Federal Reserve System or any other person associated with the Federal Reserve System.

†20th Street and Constitution Avenue N.W., Washington, D.C. 20551. E-mail: colin.j.hottman@frb.gov.
1 Introduction

Retailers have an important economic function: They transport, store, and display thousands of products for consumers to browse and buy. Despite the clear role of retailers as intermediaries, it is common in economic theory to model producers as if they sold costlessly and directly to consumers. These models ignore how retailers influence consumption. There are three reasons why retailers matter for economic outcomes. First, because 30% of U.S. household consumption comes from packaged goods bought from retailers, distortions in retail may be important for allocative efficiency and consumer welfare. Second, variation in retail competition across locations may be important for understanding regional variation in markups, which plays an important role in recent economic geography models. Third, retail stores are differentiated, so the variety of retail stores available to consumers matters for consumer welfare across locations. I analyze these mechanisms through which retailers affect the U.S. economy and show them all to be important channels affecting national welfare and the attractiveness of different cities.

I study these three aspects of the retail sector in a unified framework based on nested constant elasticity of substitution (CES) demand. Retail markups are variable despite CES demand, because I allow retailers to internalize their effect on the market price index. Retail competition is thus oligopolistic, and retailers with larger market shares set higher markups. Distortions in retail stem from variation in markups across stores within a location, resulting in an endogenous misallocation of consumption across retail stores. Across locations, differences in retail market concentration generate differences in retail markups. Under CES demand, the representative consumer has a love for variety. All else being equal, a greater number of retail stores operating in a location will raise consumer welfare. The importance of all three mechanisms (average markups, markup dispersion, and retail store variety) for consumer welfare depends on one key parameter: the substitutability across stores in a location. As stores become closer substitutes, retailers set lower markups, the losses from consumption misallocation are smaller, and the consumer gains from additional store variety become smaller.

Quantifying the importance of these three retail channels requires estimating the substitutability across stores. The ideal data to estimate this parameter would be measures of store level price indices and store market shares. Previous studies have not been able to make this estimate, however, because store-level prices are typically unobservable. In this paper, I use retail store scanner data, which include prices and sales at the barcode level from about 16,000 stores from 72 retail chains across 55 metropolitan statistical areas (MSAs) in the United States. Using these barcode data, one could construct store price
indices as a simple average of barcode prices or as a store-level unit value. Instead, I use the structural model of nested CES demand to build up to store-level price indices from barcode prices and sales, correcting for differences in product variety across stores. I then estimate the substitutability across stores using a generalized method of moments (GMM) approach.

An important contribution of this paper is to develop a tractable framework that allows me to estimate markup variation across locations. This markup heterogeneity plays an important role in many models in international trade and economic geography that predict that larger markets feature lower markups as a result of tougher competition. In economic geography models, the variation in competition across cities acts as an agglomeration force because consumers benefit from the lower markups in larger cities and only the most productive firms can produce there. This is the first paper to show that markups are lower in larger cities. Moreover, my results on markups and market size also shed light on the question of how large a market size is necessary for oligopolistic competition to converge to the monopolistically competitive limit. As Dhingra and Morrow (2013) discuss, “While the [monopolistically competitive] CES limit is optimal despite imperfect competition, it is an open empirical question whether markets are sufficiently large for this to be a reasonable approximation to use in lieu of richer variable elasticity demand” (page 22). I provide the first answer to this question of what market size is sufficiently large. My results show that larger cities have significantly lower markups than smaller cities in the United States. New York City, with a population of about 19 million people, is estimated to have a lower share-weighted average markup by 10 to 30 percentage points relative to Des Moines, which has a population of about 570,000 people. Furthermore, New York City and Los Angeles are found to be approximately at the undistorted monopolistically competitive limit in terms of markups and the deadweight loss from misallocation. These findings are robust to different market definitions (e.g., county versus metropolitan statistical area) and assumptions about which decision-making unit sets markups (e.g., the retail chain or the individual stores).

A second contribution of this paper is providing the first estimate of deadweight loss from retail misallocation. Although the mean retail markup matters because of the standard monopoly deadweight loss, what matters for retail misallocation is the dispersion of retail markups in a location. Variable markups across retail stores distort the relative prices faced by consumers and thus the equilibrium share of retail sales across stores. Because more productive (or higher-quality) retail stores will have higher markups, the equilibrium share of retail sales of the relatively productive retail stores is too low relative to the undistorted equilibrium. This misallocation of sales across retail stores makes
consumers and society worse off. Retail misallocation can be the source of a large dead-
weight loss because a significant fraction of household consumption comes from retail
goods. Retail market concentration is also an active area of interest for policymakers. For
example, the Federal Trade Commission challenged supermarket mergers in 134 of the
153 markets it investigated between 1998 and 2007 (Hanner et al. 2011). My framework
allows me to quantify the deadweight loss from retail misallocation by using the struc-
tural model to compute a counterfactual equilibrium in which I remove the dispersion in
retail markups while keeping the mean markup unchanged. My results show that losses
from retail misallocation are economically significant. Misallocation losses for consumers
are between 1% and 4.6% of aggregate packaged goods consumption, depending on the
nature of competition. The value of this lost consumption to consumers is $918 million to
$4.4 billion per year. The social deadweight loss from retail misallocation is $302 million
to $2.2 billion per year. These deadweight losses represent between 0.3% and 2.3% of total
yearly sales.

The third contribution of this paper is providing the first estimates of the consumer
gain from access to a greater variety of retail stores. To understand why consumers would
gain from a greater variety of stores, consider stores differentiated by location. When
more stores are available, consumers save on travel costs. Additionally, stores are dif-
ferentiated by other characteristics, such as store amenities or product variety. The gains
from variety in my framework depend on the substitutability across stores in a location. If
stores are viewed by consumers as close substitutes, then the consumer gains from addi-
tional retail stores will be small. By estimating the substitutability across stores, I am able
to construct retail store variety-adjusted consumer price indices across locations. My esti-
mates imply that retail store variety has a significant effect on the cost of living and could
be an important consumption-based agglomeration force. Retail store variety-adjusted
county price indices are 50% lower in the largest counties (e.g., Los Angeles County) rel-
ative to counties with populations of 150,000 people (e.g., Johnson County, Texas). One
concern with my price index is that some counties are very large, like Los Angeles County,
and consumers may not actually shop far from where they live and work. To address this
potential concern, I alternatively construct price indices using truncated (first three dig-
its) zip code areas instead of counties. This approach breaks up Los Angeles County (and
other large counties) into smaller areas. My results on the gains from retail store variety
are qualitatively unchanged by using these zip code areas instead of counties.

My paper is related to several parts of the literature. My estimate of the importance of
retail misallocation complements the larger literature studying misallocation across pro-
ducers (e.g., Banerjee and Duflo 2005, Restuccia and Rogerson 2008, Hsieh and Klenow
In this literature, recent papers have focused on variable markups as a potential source of endogenous misallocation (Epifani and Gancia 2011; Edmond, Midrigan, and Xu 2015; Peters 2011; Holmes, Hsu, and Lee 2014; Dhingra and Morrow 2013). However, these papers ignore retailers in their models and only consider misallocation across producers. In contrast, I focus on variable markups across retailers as a source of potential misallocation. I find that the consumption losses from retail misallocation are about the same magnitude as the losses from producer misallocation in the United States as a result of either financial frictions (Gilchrist, Sim, and Zakrajsek 2013), job creation and destruction frictions (Hopenhayn and Rogerson 1993), or consumer packaged goods producers’ markups (Hottman, Redding, and Weinstein 2016).

This paper also contributes to the literature on markups and market size. Standard models of international trade (Melitz 2003) and economic geography (Krugman 1991) feature constant markups across markets of different sizes. Recent models predict that larger markets have lower markups because of increased competition (e.g., Melitz and Ottaviano 2008 and Feenstra 2014 in international trade and Baldwin and Okubo 2006, Behrens and Murata 2009, Combes et al. 2012, Behrens and Robert-Nicoud 2014, and Behrens et al. 2017 in economic geography). Prior theoretical work also predicts that the efficiency loss from markups disappears as markets grow large in the limit (e.g., Postlewaite and Schmeidler 1978, Hart 1979, Guesnerie and Hart 1985, Dhingra and Morrow 2013). My empirical results confirm the theoretical predictions from the literature.

In terms of the prior empirical literature on markups and market size, there is very little direct evidence on markups. Some papers indirectly examine how models with variable markups fit the data in terms of other facts, such as how the number of establishments and establishment sizes vary with city size (e.g., Holmes and Stevens 2002, Campbell and Hopenhayn 2005, Campbell 2005, Dunne et al. 2009, Manning 2010, Combes and Lafourcade 2011). Syverson (2007) studies ready-mix concrete and shows that average prices and price dispersion are both lower in denser markets, although he does not estimate markups. Badinger (2007) uses aggregate manufacturing data and a crude accounting measure of markups at the country-industry level to study how markups vary with market size. Bellone et al. (2016) use French manufacturing data to examine how production-function-based estimates of firm-level markups vary with proxy measures of domestic industry market size. Two other recent papers similarly use production data to examine how estimated manufacturer markups vary with regional industry concentration in China (Zhao 2011 and Lu, Tao, and Yu 2014). These papers—based only on manufacturing data—observe plant-level unit values at best, typically use industry-level price deflators, have relatively aggregated definitions of products and face difficulties due to
multi-product plants. Unlike these papers, I observe very disaggregated prices and quantities within retail stores. I know that consumers are local, and I use retail market shares defined at the U.S. county level.

My paper also contributes to the literature estimating consumer gains from variety. New economic geography models predict that larger cities have lower price indices and that this price difference is an important consumption-based agglomeration force (e.g., Krugman 1991, Helpman 1998, Glaeser, Kolko, and Saiz 2001, Ottaviano, Tabuchi and Thisse 2002). The existing evidence from product prices and product variety is consistent with this prediction (Handbury and Weinstein 2015, Li 2012, Handbury 2013), but differences in variety-adjusted price indices across cities are relatively small. The gains for consumers from greater restaurant variety in larger cities appears larger (Berry and Waldfogel 2010; Schiff forthcoming; and Couture 2013). However, measuring restaurant prices and controlling for differences in restaurant quality is difficult. This paper is the first to estimate the consumer gains from the greater variety of retail stores in bigger cities, a setting in which I can control for store prices and quality. Furthermore, my results suggest a re-interpretation of prior results. In particular, Handbury and Weinstein (2015) find that product variety-adjusted consumer price indices are lower in larger cities, but their city-level price indices implicitly assume that consumers can costlessly shop from any store in the city. When I build product variety-adjusted price indices at the store level, I find that the average store price index is rising with city size. However, I do find that larger cities have lower price indices when the variety of retail stores is taken into account. These two findings tell us that larger cities do not have lower price indices because the typical store contains more product variety, but instead because there is greater store variety. Lastly, a related paper is Atkin, Faber, and Gonzalez-Navarro (forthcoming). They use a retail scanner dataset for Mexico and a nested constant elasticity of substitution (CES) demand structure that are similar to mine. However, their focus is different. They investigate the welfare impacts of foreign retail entry in Mexico. I focus on differences in retail markups, retail misallocation, and the gains from store variety across U.S. cities.

The rest of the paper is structured as follows. Section 2 describes the data used. Section 3 derives the structural model. Section 4 outlines the estimation strategy. Section 5 presents the estimation results. Section 6 concludes.

2 Data

My main data come from the Kilts retail database from Nielsen and contain barcode-level point-of-sales data from 16,680 stores from 72 retail chains operating in 55 MSAs in the
United States.\footnote{My results are calculated based on data from the Nielsen Company (U.S.), LLC, and provided by the Marketing Data Center at the University of Chicago Booth School of Business. These data were accessed while the author was a doctoral student at Columbia University. Information on availability and access to the data is available at \url{http://research.chicagobooth.edu/nielsen}.} A list of the 55 MSAs is given in the appendix. Nielsen collects the retailer data directly from store point-of-sales systems. Some of the retailers that Nielsen contracts with declined to make their data available to researchers. However, if a retailer is in the Kilts retailer data, then generally the data contain all of that retailer’s store locations. For each store, I observe the price and quantity sold for every barcoded product sold in a given week from 2006 through 2010. There are approximately 3 million unique barcodes observed in the database. Nielsen assigns the barcode-level products into product categories called \textit{product groups} based on where they are generally located within a retail store. The data are organized into 106 product groups. For example, the data include health and beauty product groups such as cosmetics and over-the-counter pharmaceuticals; non-food grocery product groups such as detergent, batteries, and pet care; household supply product groups such as cookware, computer/electronic, and film/camera; and grocery food product groups such as carbonated beverages and bread. For the typical city, the observed store-level data contain about one-third of all retail grocery, pharmacy, and mass-merchandise sales occurring during this period. This fraction ranges from about two-thirds to about fifteen-hundredths across the cities. The data are aggregated to a quarterly frequency to avoid issues such as consumer stockpiling, store inventory management, temporary promotional sales, and stickiness in price setting, which would require the theoretical model to feature dynamics.

I use two additional sources of data along with the Kilts scanner data. The first additional data source is the 2007 Census of Retail Trade data on county-level sales by NAICS code for grocery, pharmacy, and mass-merchandise retail stores.\footnote{The NAICS codes are the following: 445110 for Supermarkets and Grocery Stores (excluding convenience stores), 446110 for Drugstores and Pharmacies, 452112 for Discount Department Stores, and 452910 for Warehouse Clubs and Supercenters.} Because the Kilts data do not contain the universe of sales, I need the Census of Retail Trade data to define the total sales in a market. These Census data make it possible to construct county-level market shares for the stores in the Kilts data. The second additional data source is the 2009 Nielsen Market Scope data on market shares by retail chain for each MSA. These data provide MSA market shares for the universe of retail chains and thus includes the retail chains not observed in the Kilts scanner data. I use these data in a robustness check.

Table 1 shows summary statistics on the Kilts retail data. The table shows that there is substantially more variation in the number of stores across markets than in the number of retail chains. The 90th percentile county has more than ten times as many stores as the
10th percentile county, but only two times as many retail chains. This fact suggests that while sales per store is likely falling as market size rises, the relationship between market size and sales per chain is not as clear. Furthermore, the competitive model is unlikely to apply to this retail sector, as even the largest city has only 16 retail chains.

Table 1 also demonstrates the importance of modeling grocery, pharmacy, and mass-merchandise retailers as multi-category retailers. The average store in the data sells products in 98 product groups, while the 10th percentile number of product groups offered by a store is 86. These retail stores sell thousands of different barcodes, on average more than 19,000, with the 10th percentile number of barcodes sold being 4,683.

<table>
<thead>
<tr>
<th></th>
<th>Avg</th>
<th>Median</th>
<th>Std. Dev.</th>
<th>10th Percentile</th>
<th>90th Percentile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td># Retail chains per county</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td># Stores per county</td>
<td>62</td>
<td>36</td>
<td>82</td>
<td>9</td>
<td>138</td>
<td>679</td>
</tr>
<tr>
<td># Retail chains per city</td>
<td>8</td>
<td>9</td>
<td>2</td>
<td>6</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td># Stores per city</td>
<td>303</td>
<td>211</td>
<td>288</td>
<td>75</td>
<td>609</td>
<td>1555</td>
</tr>
<tr>
<td># Product groups per store</td>
<td>98</td>
<td>100</td>
<td>10</td>
<td>86</td>
<td>105</td>
<td>106</td>
</tr>
<tr>
<td># UPCs per store</td>
<td>19,338</td>
<td>19,422</td>
<td>10,029</td>
<td>4,683</td>
<td>33,065</td>
<td>37,873</td>
</tr>
</tbody>
</table>

Note: Calculated based on data from the Nielsen Company (U.S.), LLC, and provided by the Marketing Data Center at the University of Chicago Booth School of Business.

To summarize, the data have key features that my model needs to incorporate. The model needs to allow retailers to sell products in many product categories and to provide a way to summarize the prices of thousands of barcoded products. The model also needs to allow retail chains to internalize the effect of their price changes across their many stores in the same market. The next section outlines the theoretical model.

## 3 Theoretical framework

The roadmap for this section is as follows. First, I explain my choice of market definition. Second, I describe consumer preferences. I conclude this section by solving the retailer problem.

### 3.1 Market definition

The market definition I use for my benchmark case is the county, so stores will compete for consumers within a county. This is the smallest market area in the publicly available...
Census of Retail Trade data. In the Kilts data, I can observe store locations at the sub-county level, but only at the truncated (first three digits) zip code level. As Hanner et al. (2011) note, “Many studies which focus on localized competition between retailers use relatively small geographic market definitions such as a county. This definition is reasonable when using a demand-side definition of a market: consumers do not travel far to purchase food and are likely most familiar with the retailers in operation near where they live and work” (page 9). Using the county market definition is more disaggregated than using the MSA, the market definition used in a recent Federal Trade Commission analysis of the effects of grocery retail mergers (Hosken et al. 2012). My results will be robust to using the MSA as the relevant market definition instead of the county.

3.2 Consumer preferences

Consumer behavior features multi-stage budgeting that occurs in three stages. Figure 1 shows the stages of the budgeting process. In the first stage, consumers in a county decide which store to buy from based on the store price indices. In the second stage (conditional on shopping at a given store), consumers decide in which product group (e.g., carbonated beverages, bread) to buy a product based on the product group price indices. In the third and final stage (conditional on shopping in a given store and product group), consumers decide which barcode (e.g., 12 oz. Coke) to purchase based on the barcode prices. The demand of the representative consumer will be CES demand at every stage. This preference structure is isomorphic to a nested logit model with a population of heterogeneous consumers who each choose a single option at each stage (Anderson, de Palma, and Thisse 1992).
Two reasons motivate my choice of the nested CES functional form for consumer utility. First, this form allows my model to nest prior work in the literature as a special case. For example, my framework will nest the constant markup CES model (used in Krugman 1991, Melitz 2003, and in the misallocation literature by Hsieh and Klenow 2009), which is an important benchmark, as well as the monopolistically competitive limit case in Dhingra and Morrow (2013). The CES model is also used in the literature on consumer gains from variety (e.g., Handbury and Weinstein 2015, Li 2012, Couture 2013). The second reason I use nested CES is for analytical tractability. This functional form makes it possible to provide an analytical solution to the multi-store, multi-product retail chain pricing problem. The functional form also makes it possible to conduct an exact additive decomposition of consumer welfare.

### 3.2.1 Utility function

Utility of the representative consumer in county \( c \) at time \( t \) is assumed to be given by

\[
U_{ct} = \left[ \sum_{s \in R_{ct}} \left( \varphi_{st} C_{st} \right)^{\sigma_S^{-1}} \right]^{\sigma_S^{-1}}, \quad \sigma_S > 1, \, \varphi_{st} > 0, \quad (1)
\]

where \( C_{st} \) is the consumption index of store \( s \) at time \( t \); \( \varphi_{st} \) is the quality of store \( s \) at time \( t \); \( R_{ct} \) is the set of stores in county \( c \) at time \( t \); and \( \sigma_S \) is the constant elasticity of substitution across stores within the county.
The consumption index of each store, \( C_{st} \), is itself a CES aggregator and is given by

\[
C_{st} = \left[ \sum_{g \in G_{st}} \left( \varphi_{gst} C_{gst} \right)^{\frac{1}{\sigma_{G}}} \right]^{\frac{1}{\sigma_{G} - 1}}, \quad \sigma_{G} > 1, \ \varphi_{gst} > 0, \tag{2}
\]

where \( C_{gst} \) is the consumption index of product group \( g \) from store \( s \) at time \( t \); \( \varphi_{gst} \) is the quality of product group \( g \) at store \( s \) at time \( t \); \( G_{st} \) is the set of product groups in store \( s \) at time \( t \); and \( \sigma_{G} \) is the constant elasticity of substitution across product groups within the store.

As with stores, the consumption index of each product group, \( C_{gst} \), is itself also a CES aggregator and is given by

\[
C_{gst} = \left[ \sum_{u \in U_{gst}} \left( \varphi_{ust} C_{ust} \right)^{\frac{1}{\sigma_{U_{g}}}} \right]^{\frac{1}{\sigma_{U_{g}} - 1}}, \quad \sigma_{U_{g}} > 1, \ \varphi_{ust} > 0, \tag{3}
\]

where \( C_{ust} \) is the consumption of UPC \( u \) from store \( s \) at time \( t \); \( \varphi_{ust} \) is the quality of UPC \( u \) at store \( s \) at time \( t \); \( U_{gst} \) is the set of UPCs within product group \( g \) in store \( s \) at time \( t \); and \( \sigma_{U_{g}} \) is the constant elasticity of substitution across UPCs within product group \( g \) within the store.

Because the utility function is homogeneous of degree one in quality, I will need to choose a normalization of the quality parameters.\(^3\) The following normalizations will prove convenient:

\[
\left( \prod_{u \in U_{gst}} \varphi_{ust} \right)^{\frac{1}{N_{gst}}} = \left( \prod_{g \in G_{st}} \varphi_{gst} \right)^{\frac{1}{N_{st}}} = 1, \tag{4}
\]

where \( N_{gst} \) is the number of barcodes in product group \( g \) in store \( s \) at time \( t \) and \( N_{st} \) is the number of product groups in store \( s \) at time \( t \). Thus, I will normalize the geometric mean barcode quality to be equal to one for each product group and time period. I also normalize the geometric mean product group quality to be equal to one for each store and time period.

Although I could choose the same normalization for store quality, I will instead choose a different normalization. I pick the largest drugstore (by sales) that is present in every city in my data and for each county and time period normalize the store quality of the highest selling store from this drugstore chain to be equal to one. This normalization

\(^3\)This normalization will not matter for any of my main results.
means that my store quality parameters for each county are all expressed relative to the store quality of the same drugstore chain.

Having defined the utility function, I next solve for the consumer budgeting decisions via backward induction, starting from the problem of allocating expenditure across UPCs in a given product group and store.

3.2.2 Lowest-tier: allocating expenditure across barcodes within product groups

In the lowest tier of demand, the representative consumer allocates expenditure across barcodes within a given product group in a given store. Barcode $u$’s share of consumer spending in product group $g$ at store $s$ in county $c$ at time $t$ is given by

$$S_{ust} = \frac{(P_{ust} / \phi_{ust})^{1-\sigma_u}}{\sum_{k \in U_{gst}} (P_{kst} / \phi_{kst})^{1-\sigma_u}}, \quad \sigma_u > 1, \ \phi_{kst} > 0,$$

(5)

where $P_{ust}$ is the retail price of UPC $u$ at store $s$ at time $t$; $\phi_{ust}$ is the quality of UPC $u$ at store $s$ at time $t$; $U_{gst}$ is the set of UPCs within product group $g$ at store $s$ at time $t$; and $\sigma_u$ is the constant elasticity of substitution across UPCs in product group $g$.

The corresponding price index for product group $g$ at store $s$ at time $t$ is then given by

$$P_{gst} = \left[ \sum_{k \in U_{gst}} \left( \frac{P_{kst}}{\phi_{kst}} \right)^{-1-\sigma_u} \right]^{-1-\sigma_u}.$$

(6)

3.2.3 Middle-tier: allocating expenditure across product groups within stores

With the price indices for each product group known, I can now solve for the allocation of expenditure across product groups in a given store. Product group $g$’s share of spending in store $s$ at time $t$ is given by

$$S_{gst} = \frac{(P_{gst} / \phi_{gst})^{1-\sigma_g}}{\sum_{k \in G_{gst}} (P_{kst} / \phi_{kst})^{1-\sigma_g}}, \quad \sigma_g > 1, \ \phi_{kst} > 0,$$

(7)

where $P_{gst}$ is the product group price index given by equation 6; $\phi_{gst}$ is the quality of product group $g$ at store $s$ at time $t$; $G_{gst}$ is the set product groups at store $s$ at time $t$; and $\sigma_g$ is the constant elasticity of substitution across product groups within the store.

The price index for store $s$ at time $t$ is then given by

$$P_{st} = \left[ \sum_{k \in G_{gst}} \left( \frac{P_{kst}}{\phi_{kst}} \right)^{-1-\sigma_g} \right]^{-1-\sigma_g}.$$

(8)
3.2.4 Highest-tier: allocating expenditure across stores within a county

With the price indices for each store known, I can now solve for the allocation of expenditure across stores in a given county. The share of consumer spending on store $s$ within county $c$ at time $t$ is given by

$$S_{sct} = \frac{(P_{st}/\varphi_{st})^{1-\sigma_s}}{\sum_{k \in R_{ct}} (P_{kt}/\varphi_{kt})^{1-\sigma_s}}, \quad \sigma_s > 1, \quad \varphi_{kst} > 0,$$

(9)

where $P_{st}$ is the store price index given by equation 8; $\varphi_{st}$ is the quality of store $s$ at time $t$; $R_{ct}$ is the set of stores in county $c$ at time $t$; and $\sigma_s$ is the constant elasticity of substitution across stores within the county.

The price index for county $c$ at time $t$ is then given by

$$P_{ct} = \left[ \sum_{k \in R_{ct}} \left( \frac{P_{kt}}{\varphi_{kt}} \right)^{1-\sigma_s} \right]^{\frac{1}{1-\sigma_s}}.$$

(10)

3.2.5 Barcode quantity demand

Having solved for the expenditure shares at each stage of consumer budgeting, I can now solve for the quantity demanded of each barcode in each store. The sales of barcode $u$ in product group $g$ at store $s$ in county $c$ at time $t$ is given by

$$E_{ust} = S_{ust}S_{gst}S_{sct}E_{ct},$$

(11)

where $E_{ust}$ is barcode $u$’s sales and $E_{ct}$ is the expenditure on retail in county $c$ at time $t$.

Demand for barcode $u$ in terms of quantities can be written as

$$Q_{ust} = \frac{E_{ust}}{P_{ust}},$$

(12)

where substituting for the share terms in equation 11 and re-writing gives the following

$$Q_{ust} = \varphi_{st}^{\sigma_s-1} \varphi_{gt}^{\sigma_g-1} \varphi_{ut}^{\sigma_u-1} E_{ct}^\sigma p_{st}^{\sigma_s-1} p_{gt}^{\sigma_g-1} p_{ut}^{\sigma_u-1}.$$

(13)

3.3 Retailer problem

I will define the retail chain as the parent company that owns the retail stores. This definition is the broadest possible and will be the most economically substantitive when the same parent company owns multiple retail banners (i.e., store brands). In my case, the retail chain market share in a county will be the sum of the county market shares across
all the stores owned by the same company. In this approach, the parent company will be the decision-making unit setting optimal prices, taking into account substitutability across all the stores its owns. I will consider the alternative polar case of each individual store independently setting prices as a robustness check.

Importantly, I will allow retail chains to be large relative to the county retail market. Retail chains will thus internalize their effects on the county price index, the magnitude of which will depend on retail chain market shares. Despite CES demand, the retail chains will thus face perceived elasticities of demand that vary with chain market share. However, I will assume that retail chains are small relative to the overall county economy and thus take county expenditure and factor prices as given.\footnote{See D’Aspremont et al. (1996) for a discussion of the case when firms are allowed to internalize their impact on aggregate expenditure.}

3.3.1 Retailer technology

Retail store \( s \) in county \( c \) at time \( t \) has a total variable cost for supplying barcode \( u \) in product group \( g \) of

\[
V_{ust} (Q_{ust}) = z_{ust} Q_{ust}^{1+\delta_g},
\]

(14)

where \( Q_{ust} \) is the total quantity supplied of barcode \( u \) by store \( s \); \( \delta_g \) determines the convexity of marginal cost with respect to output for barcodes in product group \( g \); and \( z_{ust} \) is a store-barcode-specific shifter of the cost function. Costs are incurred in terms of a composite factor input that is chosen as the numeraire. One reason to expect \( \delta_g > 0 \) is the presence of fixed factors in the retailer production function. This type of convex cost function is also generated by inventory-capacity problems (Gallego et al. 2006). The same kind of cost function at the barcode level is used in Burstein and Hellwig (2007) and Broda and Weinstein (2010).

Retail store \( s \)’s marginal cost of supplying barcode \( u \) depends on the quantity supplied and is given by

\[
m_{ust} = (1 + \delta_g) z_{ust} Q_{ust}^{\delta_g}.
\]

(15)

Each retail store operating in county \( c \) at time \( t \) must also pay a fixed market access cost of \( H_{ct} > 0 \).

3.3.2 Profit maximization

The total profit of retail chain \( r \) in county \( c \) at time \( t \) is as follows:

\[
\pi_{rct} = \sum_{u \in U_{rct}} [P_{ust} Q_{ust} - V_{ust}(Q_{ust})] - H_{ct},
\]

(16)
where $U_{rct}$ is the set of barcodes sold in county $c$ at time $t$ at stores owned by retail chain $r$.

In the case of Bertrand competition, each retail chain chooses their prices $\{P_{ust}\}$ to maximize profits. The first-order conditions take the following form:

$$Q_{ust} + \sum_{k \in U_{rct}} [P_{kst} \frac{\partial Q_{kst}}{\partial P_{ust}} - \frac{\partial V_{kst}(Q_{kst})}{\partial Q_{kst}} \frac{\partial Q_{kst}}{\partial P_{ust}}] = 0. \quad (17)$$

Solving the first-order conditions allowing retail chains to internalize their effect on the county price index (derivation in the appendix), the optimal price is then given by

$$P_{ust} = \mu_{rct} m_{ust}, \quad (18)$$

where $\mu_{rct}$ is a markup over marginal cost that is the same across all products within retail chain $r$ in county $c$ at time $t$.

This markup is given by

$$\mu_{rct} = \frac{\varepsilon_{rct}}{\varepsilon_{rct} - 1}, \quad (19)$$

where $\varepsilon_{rct}$ is retail chain $r$’s perceived elasticity of demand in county $c$ at time $t$ and is given by

$$\varepsilon_{rct} = \sigma_s - (\sigma_s - 1) S_{rct}, \quad (20)$$

where $\sigma_s$ is the constant elasticity of substitution across stores in the county and $S_{rct}$ is the market share of retail chain $r$ in county $c$ at time $t$.

In the case of Cournot competition, the markup is given as in equation 19 where the retail chain $r$’s perceived elasticity of demand in county $c$ at time $t$ is now given by

$$\varepsilon_{rct} = \frac{1}{1 - \sigma_s} \left( \frac{1}{\sigma_s} - 1 \right) S_{rct}. \quad (21)$$

A key property of this setup is that while demand is CES, markups vary across retail chains in a county. As can be seen in equation 20 for the Bertrand case or equation 21 for the Cournot case, retail chains with higher market shares in a county face a lower perceived elasticity of demand and thus set higher markups, as in prior work in the literature (Atkeson and Burstein 2008; Edmond et al. 2015; Hottman et al. 2016). A similar relationship between markups and market shares arises under other commonly used demand systems such as linear demand, Translog, or logit demand. This markup variation across retail chains within a county will be the source of distortions in relative prices of retail stores and thus endogenous misallocation.
This model nests the standard CES monopolistic competition case of a constant markup as a special case. As retail chain market shares approach zero, the markup approaches the standard CES markup of $\frac{\sigma_s}{\sigma_s - 1}$. The quantitative question of how close retail markups are to the monopolistically competitive limit thus depends critically on the magnitude of retail chain market shares in the data. The difference in absolute terms between oligopolistic retail markups and the monopolistically competitive limit also depends on the magnitude of $\sigma_s$. Note that both oligopolistic retail markups and monopolistically competitive retail markups converge to zero as $\sigma_s \to \infty$, when stores thus become perfect substitutes, and the retail market becomes perfectly competitive.

In this setup, markups are constant across all products within a retail chain in a given county at a given time because of the weak separability implied by multistage budgeting. There is thus no within-store variable retail markup distortion. This analytic solution to the multi-store, multi-product retail chain’s pricing problem will prove very convenient to work with in later counterfactual exercises. Relaxing multistage budgeting and thus the constant markup within the chain property would require solving for markups numerically and is computationally intractable with the large number of products and stores in the data.

3.4 Decomposing the different channels for retail sector effects on consumer welfare

In this section I use the structure of the model to provide an exact decomposition of consumer welfare. First, note that consumer welfare in county $c$ at time $t$ (denoted by $W_{ct}$) is given by the ratio of county expenditure to the county price index:

$$W_{ct} = \frac{E_{ct}}{P_{ct}}. \quad (22)$$

Using equation 9 to express the share of store $s$ as a fraction of the geometric mean share of stores in county $c$, solving for the quality of store $s$, and substituting this quality into equation 10, I can re-write the county price index as

$$\ln P_{ct} = \ln \bar{P}_{st} - \frac{1}{\sigma_s - 1} \ln N_{ct} - \frac{1}{\sigma_s - 1} \ln \left[ \frac{1}{N_{ct}} \sum_{k \in R_{ct}} S_{kt} \tilde{S}_{st} \right] - \ln \bar{\phi}_{st}. \quad (23)$$

Equation 23 decomposes the county price index into four terms. The first term on the right hand side is the log of the geometric mean of store price indices in the county. Because store price indices reflect markups, this term captures the average retail markup
in a county. Store price indices also reflect product variety, so the first term also captures differences in available product variety across counties.

The second term is the log of the number of stores in the county. This term captures consumer gains from differences in available retail store variety across counties. These gains depend on $\sigma_s$, the elasticity of substitution across stores. As $\sigma_s \rightarrow \infty$, stores become perfect substitutes, the second term disappears and there are no gains from retail store variety.

The third term is the log of the average ratio of store market share to the geometric mean store market share in the county. This term is a measure of share dispersion and will capture the consumer losses from retail misallocation. Because retail chains with larger market shares set higher markups, the retail stores from chains with higher markups have smaller market shares in equilibrium than they would if all retail stores across all chains set the same markup. This substitution away from higher productivity (or quality) retail stores toward lower productivity (or quality) retail stores costs consumers in terms of welfare. The welfare effects of retail misallocation depend on the elasticity of substitution across stores.

The fourth term is the log of the geometric mean store quality in the county. This term captures consumer gains from having higher quality stores on average in their county. This term will not play an important role in later analysis.

4 Structural estimation

This section explains how I estimate the structural model. First, I outline how to recover the unobserved qualities at a given tier of demand given the elasticity of substitution at that tier. Second, I describe how to recover the unobserved markups and retailer marginal costs given the elasticity of substitution across stores. The rest of this section discusses the strategy for estimating the elasticities of substitution at each tier of demand.

4.1 Recovering unobservable qualities, retailer markups, and retailer marginal costs

4.1.1 Quality

Consider the lowest tier of the demand system. Given $\sigma_{uu}$, equation 5 defines a relationship between barcode prices and shares in which only the qualities are unobserved. This equation can thus be used to solve for the unobserved qualities, up to the normalization discussed earlier. After solving for the barcode qualities, the product group price index
can then be constructed from equation 6. This process for solving for unobservable qualities can then continue in the same way at the next tier of the demand, given the elasticity of substitution for that tier.

4.1.2 Retailer markups and retailer marginal costs

Given $\sigma_s$, equation 20 then defines the perceived elasticity of demand facing the retail chain. The perceived elasticity can then be used to compute the retail chain’s markup $\mu_{rct}$ for either Bertrand or Cournot competition. Retailer marginal costs can then be computed from the observed retail prices from $m_{uset} = \frac{p_{uset}}{\mu_{rct}}$.

4.2 Estimating the elasticities of substitution

4.2.1 Lowest tier of demand

Estimation of $\sigma_u$ in the lowest tier of demand follows the approach in Broda and Weinstein (2010), based on Feenstra (1994). A similar idea for identification has also been proposed in more recent papers (Rigobon 2003, Lewbel 2012). The identification is as follows. The slope of the demand and supply curves for a given product group, $\sigma_u$ and $\delta_g$, are assumed to be constant across barcodes and over time, but their intercepts are allowed to vary across barcodes and time. As Leontief (1929) points out, if the supply and demand intercepts for a given barcode are orthogonal, there is a rectangular hyperbola in $(\sigma_u, \delta_g)$ space that best fits the observed price and shares data of that barcode. This hyperbola can be seen in Figure 2. The orthogonality assumption alone does not provide identification: A higher value of $\sigma_u$ but a lower value of $\delta_g$ will keep the expectation at zero. If the variances of the supply and demand intercepts are heteroskedastic across barcodes in the product group, then the hyperbolas that fit the data are different for each barcode.\(^5\) Because the slopes of the demand and supply curves are the same, the intersection of the hyperbolas of the different barcodes in the product group separately identifies the demand and supply elasticities (Feenstra 1994). The rest of this subsection defines the orthogonality conditions for each barcode in terms of its double-differenced supply and demand intercepts and outlines the GMM procedure for estimating the slopes of demand and supply for each product group.

\(^5\)I can reject the null of homoskedasticity in a White test for generalized heteroskedasticity for the product groups in the data.
Start from the demand equation 5 and take the time difference and difference relative to another barcode in the same brand, product group, and store. This double differencing gives

$$\Delta k, t \ln S_{ust} = (1 - \sigma_U) \Delta k, t \ln P_{ust} + \omega_{ust}, \quad (24)$$

where the unobserved error term is

$$\omega_{ust} = (1 - \sigma_U) \left[ \Delta t \ln \phi_{kst} - \Delta t \ln \phi_{ust} \right].$$

Next, start from the pricing equation 18. Using equation 15 for marginal cost and the fact that $Q_{usct} = S_{ust} / P_{ust}$, the pricing equation can be written in double-differenced form as

$$\Delta k, t \ln P_{ust} = \frac{\delta_g}{1 + \delta_g} \Delta k, t \ln S_{ut} + \kappa_{ust}, \quad (25)$$

where the unobserved error term is

$$\kappa_{ust} = \frac{1}{1 + \delta_g} \left[ \Delta t \ln z_{usct} - \Delta t \ln z_{kst} \right].$$

The orthogonality condition for each barcode is then defined as

$$G(\beta_g) = \mathbb{E}_{T} \left[ \chi_{ust}(\beta_g) \right] = 0, \quad (26)$$
where \( \beta_g = \left( \frac{\sigma_U}{\delta_g} \right) \) and \( x_{ust} = \omega_{ust}\kappa_{ust} \).

This condition assumes the orthogonality of the idiosyncratic demand and supply shocks at the barcode level, after barcode and brand-quarter fixed effects have been differenced out. This orthogonality is plausible because product characteristics are fixed for each barcode and advertising typically occurs at the level of the brand. Supply shocks such as labor strikes or changes in manufacturing costs are unlikely to be correlated with quarterly barcode demand shocks at the store level.

For each product group, stack the orthogonality conditions to form the GMM objective function

\[
\hat{\beta}_g = \arg\min_{\beta_g} \left\{ G^*(\beta_g)'WG^*(\beta_g) \right\},
\]  

(27)

where \( G^*(\beta_g) \) is the sample counterpart of \( G(\beta_g) \) stacked over all barcodes in product group \( g \) and \( W \) is a positive definite weighting matrix. Following Broda and Weinstein (2010), I give more weight to barcodes that are present in the data for longer time periods.

### 4.2.2 Middle tier of demand

Given estimates of \( \sigma_u \), I can then construct product group price indices as outlined in section 4.1.1. Time difference the product group demand equation 7 and difference the result relative to another product group within the same store \( s \) to get

\[
\Delta^{g,t} \ln S_{gst} = (1 - \sigma_g) \Delta^{g,t} \ln P_{gst} + \omega_{gst},
\]

(28)

where the unobserved error term is \( \omega_{gst} = - (\sigma_g - 1) \Delta^{g,t} \ln q_{gst} \).

Ordinary least squares (OLS) estimation of equation 28 is expected to be biased as a result of endogeneity because the unobserved error term is likely correlated with the double-differenced product group price index. This correlation occurs because a relative increase in product group quality raises the quantity demanded of the barcodes within the product group and thus raises the product group price index because barcode supply curves are upward sloping. Estimation of \( \sigma_g \) will therefore use an instrumental variables approach as in Hottman et al. (2016).

Note that the double-differenced CES product group price index can be written as

\[
\Delta^{g,t} \ln P_{gst} = \Delta^{g,t} \ln \hat{P}_{ust} + \frac{1}{1 - \sigma_U} \Delta^{g,t} \ln \left[ \sum_{u \in U_{gst}} \frac{S_{ust}}{S_{ust}} \right],
\]

(29)

\[\text{This expression reflects the normalization that } \hat{\varphi}_{ust} = 1.\]
where a tilde indicates the geometric mean across the barcodes within product group $g$.

The first term on the right-hand side is the natural log of the geometric mean of barcode prices within the product group. This term is the reason why the product group price index is correlated with the error term in equation 28. The increase in product group price from movements along upward-sloping barcode supply curves due to increases in product group demand is fully captured in this term.

The second term on the right-hand side reflects the dispersion of shares across barcodes within the product group. This term captures how much lower the product group price index is due to gains from barcode variety. This term is plausibly uncorrelated with the error term in equation 28.

There are two reasons why the second term would not be uncorrelated with the error term in the demand equation. Orthogonality would be violated if changes in the relative shares of existing barcodes were correlated with contemporaneous changes in the demand for one product group relative to another within the store. This correlation is unlikely because quality is fixed at the barcode level and product group demand shocks are likely uncorrelated with idiosyncratic barcode supply shocks. Note that product group fixed effects and any common quarterly product group demand and supply shocks within the store (e.g., store advertising) will be differenced out in the demand equation. The other reason orthogonality would be violated would be if the introduction of new barcodes was correlated with contemporaneous changes in the demand for one product group relative to another within the store. This correlation is possibly an issue, although if there is such a correlation, the most plausible case is that product groups that gain relative share within the store add more barcodes relative to the other product groups. This pattern would induce a negative covariance between the instrumented value of the product group price index and the error term in the second stage regression. In that case, this remaining endogeneity would bias the estimated substitutability across product groups upward (away from zero). Simulations strongly suggest that the effect of an upward bias in $\sigma_g$ is to increase the estimated value of $\sigma_s$. As I will discuss in the next section, an upward bias in the substitutability across stores would mean that my results about the distortions from misallocation, the differences in markups across cities, and the gains from store variety are all biased toward zero and thus are lower bound estimates.

Keeping the discussion in the previous paragraph in mind, I estimate $\sigma_g$ using the second term in equation 29 as an instrument for the product group price index in equation 28. The moment condition for instrumental variables is
\[ E \left[ \omega_{gst} \Delta^{g,t} \ln \left( \sum_{u \in U_{gst}} \frac{S_{ust}}{\tilde{S}_{ust}} \right) \right] = 0. \] 

(30)

### 4.2.3 Upper tier of demand

Given an estimate of \( \sigma_g \), I can then construct store price indices. Time difference the store demand equation 9 and difference the result relative to another store within the same chain and county \( c \) to get

\[ \Delta^{s,t} \ln S_{sct} = (1 - \sigma_s) \Delta^{s,t} \ln P_{st} + \omega_{st}, \]

(31)

where the unobserved error term is \( \omega_{st} = -(\sigma_s - 1) \Delta^{s,t} \ln \varphi_{st} \).

As in the middle tier of demand, estimation of \( \sigma_s \) will use an instrumental variables approach. Note that the store price index can be written as\(^7\)

\[ \Delta^{s,t} \ln P_{st} = \frac{1}{1 - \sigma_s} \Delta^{s,t} \ln \left( \sum_{g \in G_{gst}} \frac{S_{gst}}{\tilde{S}_{gst}} \right) + \Delta^{s,t} \{ \frac{1}{N_{Gst}} \sum_{g \in G_{gst}} \frac{1}{1 - \sigma_U} \ln \left( \sum_{u \in U_{gst}} \frac{S_{ust}}{\tilde{S}_{ust}} \right) \} + \Delta^{s,t} \{ \frac{1}{N_{Gst}} \sum_{g \in G_{gst}} \ln \tilde{P}_{ust} \}. \]

(32)

I estimate \( \sigma_s \) using the sum of the first two terms in equation 32 as an instrument for the store price index in equation 31. The moment condition for instrumental variables is

\[ E \left[ \omega_{st} \Delta^{s,t} \{ \ln \left( \sum_{g \in G_{gst}} \frac{S_{gst}}{\tilde{S}_{gst}} \right) + \frac{1}{N_{Gst}} \sum_{g \in G_{gst}} \ln \left( \sum_{u \in U_{gst}} \frac{S_{ust}}{\tilde{S}_{ust}} \right) \} \right] = 0. \]

(33)

This moment condition assumes that changes in the relative demands for product groups within the store, changes in barcode assortment for the average product group, or changes in the relative quality of existing barcodes for the average product group are uncorrelated with the contemporaneous change in the demand for one store relative to another store within the same chain in the same county. Note that store fixed effects and any quarterly demand and supply shocks that are common across the retail chain within the county (e.g., chain advertising, chain-level product rollout) will be differenced out in the demand equation. Remember that I have excluded variation in the price index due to movements along upward-sloping barcode supply curves, a likely source of endogeneity.

A possible concern with this identification strategy is that stores might stock more barcodes when they experience positive demand shocks. This pattern would induce a

---

\(^7\)This expression reflects the normalization that \( \tilde{\varphi}_{gst} = 1 \).
negative covariance between the instrumented value of the store price index and the error term in the second stage regression. In that case, this remaining endogeneity would bias the estimated substitutability across stores upward (away from zero). However, an upward bias in the substitutability across stores would mean that my results on the distortions from misallocation, the differences in markups across cities, and the gains from store variety are all biased toward zero and thus are lower bound estimates.

5 Estimation results

5.1 Model parameters

Table 2 shows the results of estimating $\sigma_u$ and $\delta_g$ for 106 product groups. As expected, OLS estimates of $\sigma_u$ are much lower than the GMM estimates based on Feenstra (1994). The GMM estimates are reasonably precise. The confidence intervals for $\sigma_u$ do not cross for the estimates between the 10th and 90th percentile. I can also reject $\delta_g \geq 1$ for all product groups. The estimates of $\delta_g$ are all less than 1 and imply that marginal cost is inelastic with respect to quantity. This result is consistent with the findings of Gagnon and López-Salido (2014), who estimate, using different data and methods, that supermarket supply curves are relatively flat in the short run. The median $\sigma_u$ of 7 means that a one percent increase in the price of a given barcode will reduce the quantity demanded of that barcode by seven percent.\footnote{This calculation is assuming that the barcode has a near-zero market share within its product group.} The results show that demand for a given barcode in a given category in a given store is very elastic. Consumers are very willing to purchase different barcodes in response to price changes.

<table>
<thead>
<tr>
<th>Percentile</th>
<th>$\sigma_u$ OLS (95% CI)</th>
<th>$\sigma_u$ GMM (95% CI)</th>
<th>$\delta$ GMM (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.3 (-0.7, 0.5)</td>
<td>3.6 (3.4, 3.8)</td>
<td>0.01 (-0.004, 0.01)</td>
</tr>
<tr>
<td>5</td>
<td>0.6 (0.3, 0.7)</td>
<td>3.8 (3.6, 3.9)</td>
<td>0.02 (0.01, 0.02)</td>
</tr>
<tr>
<td>10</td>
<td>0.8 (0.6, 0.9)</td>
<td>4.3 (4.0, 4.4)</td>
<td>0.02 (0.02, 0.03)</td>
</tr>
<tr>
<td>25</td>
<td>1.0 (0.9, 1.2)</td>
<td>5.4 (4.8, 5.6)</td>
<td>0.03 (0.03, 0.04)</td>
</tr>
<tr>
<td>50</td>
<td>1.5 (1.4, 1.6)</td>
<td>7.0 (6.0, 7.6)</td>
<td>0.09 (0.07, 0.10)</td>
</tr>
<tr>
<td>75</td>
<td>2.0 (2.0, 2.1)</td>
<td>10.6 (9.2, 11.8)</td>
<td>0.13 (0.11, 0.16)</td>
</tr>
<tr>
<td>90</td>
<td>2.3 (2.2, 2.4)</td>
<td>16.0 (13.5, 18.4)</td>
<td>0.18 (0.15, 0.23)</td>
</tr>
<tr>
<td>95</td>
<td>2.6 (2.5, 2.6)</td>
<td>22.8 (16.0, 27.4)</td>
<td>0.22 (0.19, 0.26)</td>
</tr>
<tr>
<td>99</td>
<td>2.6 (2.6, 3.7)</td>
<td>31.7 (25.9, 37.4)</td>
<td>0.36 (0.27, 0.41)</td>
</tr>
</tbody>
</table>

The estimates of $\sigma_u$ can also be compared with the estimated $\sigma_g$ at the middle tier of
demand, shown in Table 3. I estimate $\sigma_g$ to be 4.8 using instrumental variables (IV). This value is lower than $\sigma_u$ for the vast majority of product groups, which implies that barcodes are typically more substitutable within product groups than they are across product groups.

Table 3 also shows the estimation result for $\sigma_s$ in the upper tier of demand. The estimated $\sigma_s$ is 4.5 using IV, which means that a one percent increase in a store’s price index reduces that store’s market share by four and one-half percent. This elasticity is similar to retail store own-price elasticities estimated in the recent industrial organization literature on supermarket competition (e.g., Smith 2004, Richards and Hamilton 2013). The point estimate of $\sigma_s$ is less than $\sigma_g$, which implies that barcodes are more substitutable within stores than across stores. However, I cannot statistically reject that $\sigma_s$ is equal to $\sigma_g$.

\[
\begin{array}{c|cc}
\text{OLS} & \sigma_g & \sigma_s \\
\hline
\text{95\% CI} & \text{1.1 (1.1, 1.1)} & \text{1.5 (1.4, 1.6)} \\
\hline
\text{IV} & \text{4.8 (4.6, 5.0)} & \text{4.5 (4.3, 4.7)} \\
\end{array}
\]

### 5.2 Retail markup estimates

Table 4 shows the retail chain markups implied by the estimated model parameters.\(^9\) The estimated markups are reasonable. The monopolistically competitive markup is 28%.\(^{10}\) The markup of the median retail chain under Bertrand competition is 29%, while it is 33% under Cournot competition. To get a sense of the plausibility of my parameter estimates, my markup estimates can be compared with retail markup estimates obtained using different data and methods. For comparison, in the Census of Retail Trade, the average retail markup is 39% (Faig and Jerez 2005). This figure is broadly comparable to what I estimate. Somewhat higher markups of 45% have been estimated for U.K. supermarkets (Schiraldi et al. 2015).

---

\(^9\)Markups are defined as (price - marginal cost)/marginal cost.

\(^{10}\)Note that my OLS estimate of $\sigma_s$ would imply an absurd monopolistically competitive markup of 200%.
Table 4: Distribution of Markup Estimates

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Markup using Bertrand</th>
<th>Markup using Cournot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.28</td>
<td>0.29</td>
</tr>
<tr>
<td>5</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>10</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>25</td>
<td>0.29</td>
<td>0.30</td>
</tr>
<tr>
<td>50</td>
<td>0.29</td>
<td>0.33</td>
</tr>
<tr>
<td>75</td>
<td>0.30</td>
<td>0.38</td>
</tr>
<tr>
<td>90</td>
<td>0.33</td>
<td>0.51</td>
</tr>
<tr>
<td>95</td>
<td>0.36</td>
<td>0.62</td>
</tr>
<tr>
<td>99</td>
<td>0.44</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Note: Markup = (Price - Marginal Cost)/(Marginal Cost).

Table 5 shows the distribution across counties of the markup of the largest retail chain. For the median county, the largest retail chain under Bertrand has a markup of 34%, while under Cournot competition it has a markup of 52%. There are counties for which the largest retail chains have substantially higher markups. However, the markups of the largest chains are still typically reasonable.

Table 5: Distribution of Markups of each County’s Largest Chain

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Bertrand Markup Largest Chain</th>
<th>Cournot Markup Largest Chain</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.30</td>
<td>0.34</td>
</tr>
<tr>
<td>10</td>
<td>0.30</td>
<td>0.36</td>
</tr>
<tr>
<td>25</td>
<td>0.31</td>
<td>0.41</td>
</tr>
<tr>
<td>50</td>
<td>0.34</td>
<td>0.52</td>
</tr>
<tr>
<td>75</td>
<td>0.38</td>
<td>0.72</td>
</tr>
<tr>
<td>90</td>
<td>0.49</td>
<td>1.22</td>
</tr>
<tr>
<td>95</td>
<td>0.71</td>
<td>2.19</td>
</tr>
</tbody>
</table>

Note: Markup = (P - MC)/MC.

5.3 Retail markups and city size

With estimated markups in hand, I can now investigate whether markups fall with city size, as some models predict. In this section, I consider how share-weighted average markups under Bertrand and Cournot competition vary with city size. This share-weighted average markup exactly corresponds to the theoretical counterpart in recent economic geography models with variable markups (e.g., Behrens and Murata 2009).

Figure 3 shows the share-weighted average markup by log city population for the Bertrand case. The results show that larger cities have smaller weighted average markups.
The regression line has a slope of -0.018 and is significantly different from zero at the 1% level. The fitted values imply that New York City has a share weighted average markup about 10 percentage points lower than Des Moines. This finding suggests that competition is indeed tougher in larger cities. Because $\sigma_s$ is common across cities, this result is driven by differences in retail chain market shares across cities. The robustness of this result to heterogeneity in $\sigma_s$ is considered at the end of the results section. Furthermore, remember that the monopolistically competitive markup is 0.28. The average markups in New York City and Los Angeles are quite close to this monopolistically competitive limit.

Figure 3: Bertrand Markups by City Size

Figure 4 shows the share-weighted average markup for the case of Cournot competition. Relative to the Bertrand case, the markups are higher, but the pattern across cities is unchanged. Again, larger cities have smaller weighted average markups. The regression line has a slope of -0.083 and is significantly different from zero at the 1% level. These results imply that Des Moines has a share-weighted average markup that is 30 percentage points higher than the markup in New York City. Markup differences across cities are larger in the Cournot case because the retail chain’s perceived elasticity of demand varies more with chain market share than in the Bertrand case. The average markups in New
York City and Los Angeles are further from the monopolistically competitive limit under Cournot than in the Bertrand case but are still fairly close to the limit value of 0.28.

In sum, I find that larger cities have lower markups whether retail competition is Bertrand or Cournot. However, smaller cities have dramatically larger markups under Cournot than Bertrand competition. These results are consistent with the economic geography models that predict that larger cities have tougher competition in spatial equilibrium. I also find that New York City and Los Angeles come quite close to the monopolistically competitive limit outcome in terms of market-share-weighted average markups. Thus, the answer to the question of what market size is sufficiently large turns out to be about the size of New York City.

5.4 Quantifying the losses from retail misallocation

Having estimated the model parameters and retail markups, I can now quantify the misallocation from retailer markup dispersion. The procedure is as follows. Remember that the county price index can be written as
\[
\ln P_{cl} = \ln \tilde{P}_{st} - \frac{1}{\sigma_s - 1} \ln N_{ct} - \frac{1}{\sigma_s - 1} \ln \left[ \frac{1}{N_{ct}} \sum_{k \in R_{ct}} S_{kt} \tilde{S}_{st} \right] - \ln \phi_{st},
\]

where the first term is the log of the geometric mean of store price indices and the third term captures dispersion in market shares across retail stores. To quantify misallocation, we imagine a price regulator forces every retailer to charge the county’s geometric mean markup, which holds the level of markups and the first term in the county price index fixed (the log of the geometric mean of store price indices). The consumer gains from removing the markup distortion are captured in the third term in the county price index.

To find the consumer gain, start by recomputing each store’s price index using the geometric mean markup.\(^{11}\) From equation 9, solve for the new equilibrium market shares for each store. Then use the new equilibrium county price indices from equation 10 to calculate the equivalent variation for consumers. The overall efficiency gain is the equivalent variation of consumers net of compensating retailers for profit changes. To compute the change in profits, use the estimated markups, the geometric mean markup, the observed firm sales, and the firm sales implied by the new market shares to calculate the change in variable profits for the retailers. When computing the change in aggregate consumer welfare below, aggregate utility will be assumed to be a Cobb-Douglas aggregate across counties.

Figure 5 shows the results of this exercise at the city level. The figure plots the total efficiency gains of removing markup dispersion (deadweight loss) for each city versus log city population for the Bertrand case. The Cournot case shows the same pattern. The efficiency gains are smaller in larger cities. The regression line has a slope of -0.005 and is significantly different from zero at the 1% level. These results also show that the deadweight loss from markup dispersion is close to zero for New York City and Los Angeles. These two cities are close to being perfectly efficient.

\(^{11}\)For this exercise, I will assume that barcode marginal costs are fixed with respect to quantity. The estimation results suggest that marginal costs rise slowly with output. Allowing barcode marginal costs to change would require solving a fixed point problem for barcode prices and quantities, which is difficult given the size of the data.
Table 6 shows the aggregate welfare gains from removing markup dispersion. In the Bertrand case, consumer welfare rises 1% after removing markup dispersion. This change in markups benefits consumers by $918 million per year. For comparison, the total population in these 55 MSAs is about 187 million people. Even after compensating retailers for lost profits, the net benefit to consumers and thus the deadweight loss from misallocation is $302 million per year. This deadweight loss is equal to 0.3% of total sales. The losses from retail misallocation are about the same magnitude as the losses from producer misallocation in the United States as a result of either financial frictions (Gilchrist, Sim, and Zakrajsek 2013), job creation/destruction frictions (Hopenhayn and Rogerson 1993), or consumer packaged goods producers’ markups (Hottman, Redding, and Weinstein 2016).

Consumer welfare gains and the deadweight loss are larger in the Cournot case, because under Cournot competition, markups are higher and there is greater markup dispersion. In that case, consumer welfare rises by 4.6%, or $4.4 billion per year. The deadweight loss is also larger, at $2.2 billion per year or 2.3% of total sales.

Table 6: Welfare Gains from Removing Markup Dispersion

<table>
<thead>
<tr>
<th>Case</th>
<th>%△ Consumer Welfare</th>
<th>Consumer Surplus ($/Year)</th>
<th>Total Surplus ($/Year) (% Total Sales)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bertrand</td>
<td>1%</td>
<td>$918 million</td>
<td>$302 million (0.3%)</td>
</tr>
<tr>
<td>Cournot</td>
<td>4.6%</td>
<td>$4.4 billion</td>
<td>$2.2 billion (2.3%)</td>
</tr>
</tbody>
</table>
We can also quantify the monopoly distortion, in addition to the losses from misallocation. This monopoly distortion arises because when holding wages fixed in counterfactual exercises, I am implicitly assuming perfectly elastic labor supply. The level of (average) retail markups therefore distorts equilibrium allocations through a labor wedge. The procedure for quantifying this monopoly distortion is as follows. Imagine a price regulator forces each retailer to charge the monopolistically competitive markup. This requirement is equivalent to setting the chain’s market share equal to zero in 20. Then, recompute each store’s price index using the monopolistically competitive markup. As before, I will assume that barcode marginal costs are fixed with respect to quantity. From equation 9, solve for new equilibrium market shares for each store. Next, use the estimated markups, the monopolistically competitive markup, the observed firm sales, and the firm sales implied by the new market shares to calculate the change in variable profits for the retailers. Then use the county price indices from equation 10 to calculate the equivalent variation for consumers.

Table 7 shows the results of this exercise. In the Bertrand case, consumer welfare rises 2.6% after moving markups to the monopolistically competitive level. This change in markups benefits consumers by $3 billion per year. Remember, the total population in these 55 MSAs is about 187 million people. Even after compensating retailers for lost profits, the net benefit to consumers and thus the total efficiency gain is $868 million per year. This efficiency gain is equal to 0.8% of total sales.

Consumer welfare gains and the efficiency gains are larger in the Cournot case, because under Cournot competition, markups are higher and there is greater markup dispersion. In that case, consumer welfare rises by 11.6%, or $14 billion per year. The efficiency gain is also larger, at $6 billion per year, or 5.3% of total sales.

<table>
<thead>
<tr>
<th></th>
<th>%△ Consumer Welfare</th>
<th>Consumer Surplus ($/Year)</th>
<th>Total Surplus ($/Year) (% Total Sales)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bertrand</td>
<td>2.6 %</td>
<td>$3 billion</td>
<td>$868 million (0.8 %)</td>
</tr>
<tr>
<td>Cournot</td>
<td>11.6 %</td>
<td>$14 billion</td>
<td>$6 billion (5.3 %)</td>
</tr>
</tbody>
</table>

5.5 Gains from retail store variety

Next, I will consider the gains from retail store variety. The magnitude of differences in gains from store variety across counties depends on differences in the number of stores across counties. We saw previously in the data section that there are large differences in the number of stores across U.S. counties. This fact suggests that we should expect that
gains from store variety will be important. Remember that the county price index can be written as:

\[ \ln P_{ct} = \ln \bar{P}_{st} - \frac{1}{\sigma_s - 1} \ln N_{ct} - \frac{1}{\sigma_s - 1} \ln \left[ \frac{1}{N_{ct}} \sum_{k \in R_{ct}} \frac{S_{kt}}{\bar{S}_{st}} \right] - \ln \bar{\phi}_{st}. \]

The first term captures the average product-variety-adjusted store price index. The second term captures the gains from retail store variety. Because the elasticity of substitution across stores is estimated to be less than infinite, I expect there will be consumer gains from greater retail variety in larger counties.

Figure 10 shows the overall county price indices (in levels) plotted by county population. The figure shows that county price indices dramatically fall with county size. The regression line has a slope of -0.0013 and is significantly different from zero at the 1% level. Los Angeles County, the largest county, has a price index half that of counties with a population of 150,000 people. This finding is suggestive evidence for potentially important consumption-based agglomeration forces in the United States.

Because the county price index reflects both retail store variety and differences in tradable product variety across counties, I will decompose it into multiple terms to isolate the
effects of retail store variety. First, I will investigate differences in the first term of the welfare equation, which reflects differences in tradable product variety for the (geometric) average store in the county. Then I will investigate specifically the consumer gains from store variety.

Figure 7 shows the log of the geometric mean store price index by log county population. The figure shows that the average store price index is larger for larger counties. The regression line has a slope of 0.013 and is significantly different from zero at the 5% level. This result shows that even accounting for product variety at the store level, the average store in the largest counties has a higher price index than the average store in smaller counties. Therefore, the negative relationship between the county price index and county size is not driven by differences in tradable product variety across counties.

![Figure 7: Average Store Price Index by County Size](image)

Figure 8 shows the gains from store variety by county size. The results show that the county price index falls dramatically with county size because of large differences in retail store variety. The regression line has a slope of -0.21 and is significantly different from zero at the 1% level. These results suggest that non-tradable services, in this case retail services, are necessary to generate large consumption-based agglomeration forces.
It may not be surprising that large counties contain a lot more stores. For example, Los Angeles County is substantially larger than other U.S. counties. To consider the robustness of the variety result, Figure 9 shows the retail store variety adjusted price index using truncated (first three digit) zip codes instead of counties. There are about 2.5 times more three-digit zip code areas than there were counties. The results show that the zip code price index falls dramatically with zip code population, just as in the county results. The regression line has a slope of -0.005 and is significantly different from zero at the 1% level.
5.6 Robustness

In this section, I investigate the robustness of my results with regard to three changes. First, I estimate different $\sigma$'s in different locations and compare results. Second, I consider what happens if I define the market as the MSA instead of the county. Finally, I allow each store to set its own markup instead of coordinating pricing at the retail chain level.

5.6.1 Heterogeneity in $\sigma$

In the earlier results I relied on a common estimate of $\sigma$ across locations. In this section, I investigate the robustness of my markup and price index results to allowing heterogeneity in $\sigma$ across locations. The misallocation results are qualitatively unchanged after relaxing the assumption of a common $\sigma$.

Table 8 reports results for estimating different $\sigma$ parameters for different portions of the city size distribution. Splitting the set of cities in half and estimating a different $\sigma$ for each half produces estimates not too far on either side of the base $\sigma$ estimate of 4.5. Larger cities are estimated to have a higher elasticity of substitution across stores. The two estimates are statistically different at the 5% level.
Table 8: $\sigma_s$ Estimates by City Size Dist.

<table>
<thead>
<tr>
<th>City Size Dist.</th>
<th>1st Half of Cities by Size</th>
<th>2nd Half of Cities by Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV Estimate $\sigma_s$ (95% CI)</td>
<td>3.9 (3.5, 4.3)</td>
<td>4.7 (4.5, 5.0)</td>
</tr>
</tbody>
</table>

Table 8 also shows results after splitting the data into quartiles by city size. The results are very similar to those found by splitting the sample in half. In the point estimates, larger cities have a higher elasticity of substitution across stores. However, the confidence intervals across quartiles almost overlap for all quartiles. The benchmark $\sigma_s$ estimate of 4.5 is contained in three out of four quartile confidence intervals and is almost contained in the remaining quartile confidence interval.

Table 9 shows results for estimating a different $\sigma_s$ for every city. The median estimate is very close to the benchmark estimate of 4.5. The confidence intervals overlap for nearly all the percentiles shown. I cannot reject that the true $\sigma_s$ equals 4.5 for the majority of the cities. Because I find no correlation between the estimated $\sigma_s$ parameters and either city size or density (not shown), I conclude that the differences in the estimated $\sigma_s$ parameters across cities are mostly due to lack of precision from estimating on small sample sizes.

Table 9: $\sigma_s$ Estimates by MSA

<table>
<thead>
<tr>
<th>Percentile</th>
<th>$\sigma_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>3.6 (2.2, 4.2)</td>
</tr>
<tr>
<td>25</td>
<td>3.8 (2.6, 5.1)</td>
</tr>
<tr>
<td>50</td>
<td>4.7 (3.1, 6.0)</td>
</tr>
<tr>
<td>75</td>
<td>5.7 (4.1, 7.9)</td>
</tr>
<tr>
<td>90</td>
<td>7.1 (4.9, 10.1)</td>
</tr>
</tbody>
</table>

Having estimated different $\sigma_s$ parameters across locations, I will next see how these differences matter for the markup results discussed previously. I will first consider the two $\sigma_s$ parameters estimated by splitting the cities into two groups by size and will then consider the set of city-level $\sigma_s$ estimates.

Figure 10 shows the relationship between the Bertrand markups and city size when using the $\sigma_s$ estimates from each half of the city size distribution. Because the point estimates imply that larger cities have a higher elasticity of substitution across stores, these results only reinforce the prior result that larger cities have lower markups. These two $\sigma_s$’s imply that even monopolistically competitive retail chains (with near-zero market shares)
charge lower markups in larger cities. The monopolistically competitive markup in the smaller half of cities is about 0.35, while the monopolistically competitive markup in the large half of cities is 0.26. With these estimates, the share-weighted average markup in Des Moines is now estimated to be about 17 percentage points higher than in New York City.

Figure 10: Bertrand Markups for Two $\sigma_s$ Case

Figure 11 shows the relationship between the Cournot markups and city size when I estimate a different $\sigma_s$ for each city. Larger cities are still estimated to have statistically and economically significantly lower markups in the Cournot case. New York City, for example, still has a markup that is about 30 percentage points lower than Des Moines. In the Bertrand case, the slope of the best fit line is nearly the same as in the benchmark one $\sigma_s$ case. However, the large range of $\sigma_s$ estimates adds a lot of noise, and the Betrand markup slope is no longer statistically significant.
Next, consider the robustness of the variation in county price indices to heterogeneity in $\sigma_s$. Figure 12 plots the county price indices using the estimate of $\sigma_s$ from each half of the city size distribution. Despite differences in the gains from store variety across small versus large cities, the results are nearly unchanged from the base case considered earlier. The largest county (Los Angeles County) still has a price index that is half of counties with populations of 150,000 people. This relationship is statistically significant at the 1% level.
Figure 12 plots the county price indices using the estimates of $\sigma_s$ for each city. There is significantly more variation in county price indices in this case relative to the base case, driven by heterogeneity in the gains from retail store variety across cities. However, the relationship between the price indices and county size remains robust. The regression slope is still negative and statistically significant at the 1% level. The results still imply that counties with a population of 150,000 people have a price index twice as high as the price index for Los Angeles County.
5.6.2 MSA market definition

In the earlier results the relevant market is the county. I now consider the robustness of the markup results if the market is instead defined as the MSA. I also construct retail store variety-adjusted city price indices using the MSA as the relevant market and compare these price indices with city size.

Figure 14 plots the Bertrand markups estimated using MSA retail chain market shares versus city size. The results are very similar to the base case. Larger cities have lower retail markups, and this relationship is statistically significant at the 1% level. Des Moines is estimated to have markups higher by about 15 percentage points relative to New York City. The results for the Cournot markups share a similar pattern, with larger variation in markups across cities.
Figure 15 plots city-level price indices by city size. Similar to before, larger cities have lower variety-adjusted price indices. The regression slope is the same as before: -0.001. This relationship is statistically significant at the 1% level. New York City still has a price index that is about half the level of the price index in Des Moines. Surprisingly, using a larger market definition (MSA versus county) did not result in larger differences in price indices across locations.
5.6.3 Store versus chain market share

In this section, I investigate whether markups still fall with city size if I allow each store to set its own markups instead of having constant markups at the chain level. Figure 16 plots the share-weighted average Cournot markup in this case by city size. The finding that larger cities have lower markups is robust, although the difference across cities is smaller. In this case, Des Moines has about a 6 percentage point higher markup relative to New York City, compared with a difference of 30 percentage points in the base case. The estimated Bertrand markups show the same pattern in that larger cities have lower markups. However, the difference in Bertrand markups across cities is only a few percentage points in this case.
6 Conclusion

I provide a unified framework to study three aspects of the retail sector: average markup variation with city size, consumption misallocation from dispersion in retail markups, and the consumer gains from retail store variety. I use detailed retail store scanner data to structurally estimate a model of consumer demand and oligopolistic retail competition for 55 MSAs in the United States. I use counterfactual exercises and decompositions of the consumer price index to quantify the importance of each of the three retail mechanisms.

I find that New York City has a lower share-weighted average markup by 10 to 30 percentage points relative to Des Moines, depending on the nature of competition. Cournot competition features larger markup differences across cities than Bertrand competition. Additionally, New York City and Los Angeles are found to be approximately at the undistorted monopolistically competitive limit in terms of markups and the deadweight loss from misallocation. These findings are robust to different market definitions (county versus MSA) and assumptions about which decision-making unit sets markups (e.g., the retail chain or the individual stores).
My estimates show that losses from retail misallocation are economically significant. Misallocation losses are between 1% and 4.6% of aggregate packaged goods consumption, depending on the nature of competition. The value to consumers of this lost consumption is $918 million to $4.4 billion per year. The deadweight loss from retail misallocation is $302 million to $2.2 billion per year. These deadweight losses represent between 0.3% and 2.3% of total yearly sales. The consumption losses from retail misallocation are about the same magnitude as the losses from producer misallocation in the United States.

My estimates imply that retail store variety significantly effects the cost of living and could be an important consumption-based agglomeration force. Retail store variety-adjusted county price indices are 50% lower in the largest counties (e.g., Los Angeles County) relative to counties with populations of 150,000 people (e.g., Johnson County, Texas). This result is driven by differences in the number of available retail stores and not by differences in available product variety within stores across counties. These results are robust to constructing price indices using truncated three-digit zip code areas instead of counties.

My results have important implications for policy. Any policies that can prevent or reduce concentration (e.g., merger policy) may be welfare improving, particularly in the smallest cities. Concentration may in the short run be reduced by loosening policies that prevent entry, such as the policies used by some localities to prevent entry of big box retailers. However, allowing entry of big box retailers, in-so-far as this entry leads incumbent retailers to exit, may cost consumers in terms of reduced retail store variety. Policymakers should pay careful attention to potential trade-offs of this type.

My results also raise questions for future work to address. For example, I find significant differences in consumer price indices across locations in the United States, suggesting a potential role for consumption-based agglomeration forces. However, we know that agglomeration forces must be counteracted by congestion costs in such a way as to prevent everyone from moving to the largest cities. I leave it to future work to investigate whether congestion costs such as differences in housing costs (e.g., rents) across U.S. cities are large enough to counteract the agglomeration forces suggested by this paper, or whether we must look elsewhere for the dispersion forces that explain the observed equilibrium population distribution across U.S. cities.
References


A Appendix A:

A.1 Derivation of Equations (18)-(20)

The first-order condition with respect to the price of a given UPC is:

\[ Q_{ust} + \sum_{k \in U_{ct}} [P_{kst} \frac{\partial Q_{kst}}{\partial P_{ust}} - \frac{\partial V_{kst}(Q_{kst})}{\partial Q_{kst}} \frac{\partial Q_{kst}}{\partial P_{ust}}] = 0. \]  (34)

Using equation (13) and the condition that UPC supply equals demand gives

\[ \frac{\partial Q_{kst}}{\partial P_{ust}} = (\sigma_s - 1) \frac{Q_{kst}}{P_{ct}} \frac{\partial P_{ct}}{\partial P_{ust}} + (\sigma_g - \sigma_s) \frac{Q_{kst}}{P_{st}} \frac{\partial P_{st}}{\partial P_{ust}} + (\sigma_u - \sigma_g) \frac{Q_{gst}}{P_{gust}} \frac{\partial P_{gust}}{\partial P_{ust}} - \sigma_u \frac{Q_{kst}}{P_{ust}} \frac{\partial P_{kst}}{\partial P_{ust}}. \]

Rewrite \( \frac{\partial Q_{kst}}{\partial P_{ust}} \) as

\[ \frac{\partial Q_{kst}}{\partial P_{ust}} = (\sigma_s - 1) \left( \frac{\partial P_{ct}}{\partial P_{ct}} \right) \left( \frac{\partial P_{st}}{\partial P_{ct}} \right) \left( \frac{\partial P_{gust}}{\partial P_{gust}} \right) \frac{Q_{kst}}{P_{ust}} + (\sigma_g - \sigma_s) \frac{Q_{gst}}{P_{gust}} \frac{\partial Q_{gst}}{\partial P_{gust}} + (\sigma_u - \sigma_g) \frac{Q_{ust}}{P_{gust}} \frac{\partial Q_{ust}}{\partial P_{gust}} - \sigma_u \frac{Q_{kst}}{P_{gust}} \frac{\partial Q_{kst}}{\partial P_{gust}}. \]

Use the property of CES that \( \frac{\partial P_{gust}}{\partial P_{gust}} = S_{gust} \) to solve for the elasticities to give:

\[ \frac{\partial Q_{kst}}{\partial P_{ust}} = (\sigma_s - 1) S_{scst} S_{gst} S_{ust} \frac{Q_{kst}}{P_{ust}} + (\sigma_g - \sigma_s) S_{gust} S_{ust} \frac{Q_{gst}}{P_{ust}} + (\sigma_u - \sigma_g) S_{ust} \frac{Q_{kst}}{P_{ust}} - \sigma_u \frac{Q_{kst}}{P_{ust}}. \]  (35)

If we now substitute equation (35) into equation (34) and divide both sides by \( Q_{ust} \), we get

\[ 1 + \sum_{k \in U_{cst}} (\sigma_s - 1) S_{scst} S_{gst} S_{ust} \frac{P_{kst} Q_{kst}}{Q_{ust}} + \sum_{k \in U_{cst}} (\sigma_g - \sigma_s) S_{gst} S_{ust} \frac{P_{kst} Q_{kst}}{Q_{ust}} \]

\[ + \sum_{k \in U_{gst}} (\sigma_u - \sigma_g) S_{ust} \frac{P_{kst} Q_{kst}}{Q_{ust}} - \sigma_u - \sum_{k \in U_{cst}} (\sigma_s - 1) S_{scst} S_{gst} S_{ust} \frac{\partial V_{kst}(Q_{kst})}{\partial Q_{kst}} \frac{Q_{kst}}{P_{kst} Q_{kst}} \]

\[ - \sum_{k \in U_{gst}} (\sigma_g - \sigma_s) S_{gst} S_{ust} \frac{\partial V_{kst}(Q_{kst})}{\partial Q_{kst}} \frac{Q_{kst}}{P_{kst} Q_{kst}} - \sum_{k \in U_{gst}} (\sigma_u - \sigma_g) S_{ust} \frac{\partial V_{kst}(Q_{kst})}{\partial Q_{kst}} \frac{Q_{kst}}{P_{kst} Q_{kst}} + \sigma_u \frac{\partial V_{kst}(Q_{kst})}{\partial Q_{kst}} \frac{Q_{kst}}{P_{kst} Q_{kst}} = 0. \]

Note the different sets over which the summations occur. This property is a result of the weak separability from the multi-stage budgeting.

We define the markup at the retail chain or UPC level as \( \mu_k \equiv P_k / \frac{\partial V_k(Q_k)}{\partial Q_k} \).

Since \( S_{ust} \frac{1}{\frac{1}{\frac{P_{kst} Q_{kst}}{Q_{ust}}} - \frac{1}{\frac{Q_{kst}}{P_{kst} Q_{kst}}} = 1 \) and analogously for the upper tiers, we can rewrite the previous equation as

\[ 1 + (\sigma_s - 1) S_{rcst} + (\sigma_g - \sigma_s) + (\sigma_u - \sigma_g) - \sigma_u - (\sigma_s - 1) S_{rcst} \sum_k \frac{\partial V_k(Q_k)}{\partial Q_k} \frac{Q_{kst}}{P_{kst} Q_{kst}} \]

\[ - (\sigma_g - \sigma_s) \sum_k \frac{\partial V_k(Q_k)}{\partial Q_k} \frac{Q_{kst}}{P_{kst} Q_{kst}} - (\sigma_u - \sigma_g) \sum_k \frac{\partial V_k(Q_k)}{\partial Q_k} \frac{Q_{kst}}{P_{kst} Q_{kst}} + \sigma_u \mu_{ust} = 0. \]
Because we assume that $\sigma_u, \sigma_g, \text{ and } \sigma_s$ is the same for all $u, g, \text{ and } s$ within the retail chain, $\mu_{ust}$ is the only $u$-specific term in this expression. Hence, $\mu_{ust}$ must be constant for all $u$ produced by retail chain $r$ in time $t$; in other words, markups only vary at the retail chain level. Together these two results ensure the same markup across all UPCs supplied by the chain.

We can now solve for $\mu_{rcr}$ by

$$1 + (\sigma_s - 1) S_{rcr} + (\sigma_g - \sigma_g) + (\sigma_u - \sigma_g) - \sigma_u - (\sigma_s - 1) S_{rcr} \frac{1}{\mu_{rcr}}$$

$$- (\sigma_g - \sigma_s) \frac{1}{\mu_{rcr}} - (\sigma_u - \sigma_g) \frac{1}{\mu_{rcr}} + \sigma_u \frac{1}{\mu_{rcr}} = 0$$

$$\Rightarrow \mu_{rcr} = \frac{\sigma_s - (\sigma_s - 1) S_{rcr}}{\sigma_s - (\sigma_s - 1) S_{rcr} - 1}.$$  

### A.2 List of 55 MSAs in the data

1. Albany-Schenectady-Troy, NY
2. Albuquerque, NM
3. Atlanta-Sandy Springs-Roswell, GA
4. Austin-Round Rock, TX
5. Birmingham-Hoover, AL
6. Boise City, ID
7. Boston-Cambridge-Newton, MA-NH
8. Buffalo-Cheektowaga-Niagara Falls, NY
9. Charleston, WV
10. Charleston-North Charleston, SC
11. Charlotte-Concord-Gastonia, NC-SC
12. Chicago-Naperville-Elgin, IL-IN-WI
13. Cleveland-Elyria, OH
14. Columbus, OH
15. Dallas-Fort Worth-Arlington, TX
16. Denver-Aurora-Lakewood, CO
17. Des Moines-West Des Moines, IA
18. Detroit-Warren-Dearborn, MI
19. Durham-Chapel Hill, NC
20. Grand Rapids-Wyoming, MI
21. Greensboro-High Point, NC
22. Greenville-Anderson-Mauldin, SC
23. Harrisburg-Carlisle, PA
24. Houston-The Woodlands-Sugar Land, TX
25. Huntington-Ashland, WV-KY-OH
26. Jacksonville, FL
27. Little Rock-North Little Rock-Conway, AR
28. Los Angeles-Long Beach-Anaheim, CA
29. Louisville/Jefferson County, KY-IN
30. Memphis, TN-MS-AR
31. Miami-Fort Lauderdale-West Palm Beach, FL
32. Milwaukee-Waukesha-West Allis, WI
33. Minneapolis-St. Paul-Bloomington, MN-WI
34. Nashville-Davidson–Murfreesboro–Franklin, TN
35. New Haven-Milford, CT
36. New Orleans-Metairie, LA
37. New York-Newark-Jersey City, NY-NJ-PA
38. Orlando-Kissimmee-Sanford, FL
39. Philadelphia-Camden-Wilmington, PA-NJ-DE-MD
40. Phoenix-Mesa-Scottsdale, AZ
41. Pittsburgh, PA
42. Portland-Vancouver-Hillsboro, OR-WA
43. Providence-Warwick, RI-MA
44. Raleigh, NC
45. Richmond, VA
46. Sacramento–Roseville–Arden-Arcade, CA
47. St. Louis, MO-IL
48. Salt Lake City, UT
49. San Antonio-New Braunfels, TX
50. San Diego-Carlsbad, CA
51. San Francisco-Oakland-Hayward, CA
52. Sioux Falls, SD
53. Tampa-St. Petersburg-Clearwater, FL
54. Virginia Beach-Norfolk-Newport News, VA-NC
55. Washington-Arlington-Alexandria, DC-VA-MD-WV