Retail Markups, Misallocation, and Store Variety in the US[*]

JOB MARKET PAPER

Colin Hottman
Department of Economics, Columbia University[†]

November 16, 2014

Abstract

I estimate a structural model of consumer demand and oligopolistic retail competition in order to study three mechanisms through which retailers affect allocative efficiency and consumer welfare. First, variable markups across retail stores within a location induce a misallocation of resources. The deadweight loss from this retail misallocation can be large since a significant fraction of household consumption comes from retail goods. Second, across locations, retail markups may vary with market size. This regional variation plays an important role in recent economic geography models as an agglomeration force. In the limit, models predict that the distortion from variable markups disappears in large markets, although it is an open question, "How Large is Large?" Third, since retail stores are differentiated, differences in the variety of retail stores available to consumers matters for consumer welfare across locations. To quantify the importance of these mechanisms, I estimate my model using retail scanner data with prices and sales at the barcode level from thousands of stores across the US. I find that the deadweight loss and consumption misallocation from variable retail markups are economically significant. I estimate that retail markups are smaller in larger cities, and that markets the size of New York City and Los Angeles are approximately at the undistorted monopolistically competitive limit. My results show that retail store variety significantly impacts the cost of living and could be an important consumption-based agglomeration force.

JEL CLASSIFICATION: D12, D43, D61, L11, L13, L81, R32
KEYWORDS: misallocation, markups, variety, consumers, retail, cities

[*] For the latest version visit www.columbia.edu/~cjh2164/HottmanJMP.pdf. I am grateful to Don Davis, Keshav Dogra, Juan Carlos Hallak, Amit Khandelwal, Michael Riordan, Patrick Sun, Jonathan Vogel, David Weinstein and seminar participants at Columbia University for their advice and helpful comments. All remaining errors are mine. All results are calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

[†] 420 W. 118th Street, MC 3308, New York, NY 10027. Email: cjh2164@columbia.edu. Website: www.columbia.edu/~cjh2164/.

1 Introduction

Retailers have an important economic function. They transport, store, and display thousands of products for consumers to browse and buy. Despite a 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, since 30% of US 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 differences in the variety of retail stores available to consumers matters for consumer welfare across locations. I analyze these three 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 all three mechanisms 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 impact 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 equal, a greater number of retail stores operating in a location will raise consumer welfare. The importance of all three mechanisms (misallocation, markups, 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 aspects of the retail sector 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 had not been able to do this because typically store-level prices are unobservable. In this paper, I use US retail store scanner data where I observe prices and sales at the barcode level from about 16,000 stores from 72 retail chains across 55 metropolitan statistical areas in the US.

Using this 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 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 generalized method of moments (GMM).

An important contribution of this paper is providing the first estimate of the 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. Variable markups across retail stores distort the relative prices faced by consumers and thus the equilibrium share of retail sales across stores. Since more productive[1] retail stores will have higher markups, the equilibrium share of retail sales sold through the relatively productive retail stores is too low relative to the first-best. This misallocation of sales across retail stores makes consumers and society worse off relative to an undistorted equilibrium. Retail misallocation can be the source of a large deadweight loss since 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 structural 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% to 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 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 consumption losses from retail misallocation are about the same magnitude as the losses from producer misallocation in the US due to 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 2014).

A second important contribution is that my framework 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 due to tougher competition. In economic geography models, the

[1] or higher quality.

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, further reducing costs. 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) point out, "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, "How Large is Large?"

My results show that larger cities have significantly lower markups than smaller cities in the US. 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. Additionally, New York City and Los Angeles are found 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 vs metropolitan statistical area) and assumptions about which decision-making unit sets markups (eg. the retail chain or the individual stores).

My framework allows me to provide the first estimates of the consumer gain from having a greater variety of retail stores available. To understand why consumers would gain from a greater variety of stores, consider stores differentiated by location. When there are more stores available, consumers then save on travel costs. Additionally, stores are differentiated by other characteristics, such as store amenities. 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 additional 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 estimates imply that retail store variety has a significant impact 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 (eg. Los Angeles County) relative to counties with populations of 150,000 people (eg. Johnson County, Texas). I show that 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. I find significantly larger differences in my price indices across counties than in prior work focusing on differences in product variety-adjusted price indices across cities (eg.

Handbury and Weinstein forthcoming). 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 3-digits) zip code areas instead of counties. This breaks up Los Angeles County (and other counties) into smaller areas. My results on the gains from retail store variety are 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 large literature studying misallocation across producers (eg. Banerjee and Duflo 2005, Restuccia and Rogerson 2008, Hsieh and Klenow 2009, Bartelsman et al. 2013). 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 2012, Peters 2011, Holmes, Hsu, and Lee forthcoming, Dhingra and Morrow 2013). However, these papers only consider misallocation across producers and ignore retailers in their models. In contrast, I focus on variable markups across retailers as a source of potential misallocation.

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 due to increased competition (eg. 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 2013, and Behrens et al. 2013 in economic geography). Prior work also studies theoretically what happens to markups as markets grow large in the limit (eg. Hart 1979, Guesnerie and Hart 1985, Dhingra and Morrow 2013). In terms of the empirical literature on markups and market size, we have very little direct evidence. Some papers examine indirectly 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 (eg. 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. (2014) 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 on manufacturing data only observe plant level unit values at best, typically use industry 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 US 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 is an important consumption-based agglomeration force (eg. 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 forthcoming, Li 2012, Handbury 2013), but differences in variety-adjusted price indices across cities are relatively small. The gains to consumers from greater restaurant variety in larger cities appears larger (Berry and Waldfogel 2010, Schiff 2012, and Couture 2013). However, measuring restaurant prices and controlling for differences in restaurant quality is difficult. This is the first paper 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. I find larger gains from variety than in prior work.

Lastly, a related paper is Atkin, Faber and Gonzalez-Navarro (2014). They use a similar retail scanner dataset in Mexico and a similar Nested CES demand structure. However, their focus is different. They investigate the welfare impacts of foreign retail entry in Mexico. I focus on retail markups, retail misallocation, and the gains from store variety across US 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 comes from the Kilts retail database from Nielsen and contains barcode-level point-of-sales data from 16,680 stores from 72 retail chains operating in 55 metropolitan statistical areas (MSAs) in the United States.[2] A list of the 55 MSAs is given in the

[2] My results are calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business. Information on availability and access to the data is available at http://research.chicagobooth.edu/nielsen

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 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, film/camera, and grocery food product groups such as carbonated beverages and bread. For the typical city, the observed store-level data contain about a third of all retail grocery, pharmacy, and mass-merchandise sales occuring during this time period. This fraction ranges from about 2/3rds to about 15% across the cities. The data are aggregated to the 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[3]. Since 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. This makes 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. This data provides MSA market shares for the universe of retail chains and thus includes the retail chains not observed in the Kilts scanner data.

Table 1 shows summary statistics on the Kilts retail data. The first thing the table shows is that there is substantially more variation in the number of stores across markets than in the number of retail chains. The 90th percentile market has more than ten times as many stores as the 10th percentile market, but only about two times as many retail chains. This suggests that while sales per store is falling as market size rises, the relationship between market size and sales per chain is not as clear. Furthermore, the competitive

[3] The NAICS codes are: 445110 Supermarkets and Grocery Stores (excluding convenience stores), 446110 Drugstores and Pharmacies, 452112 Discount Department Stores, and 452910 Warehouse Clubs and Supercenters.

model is unlikely to apply to this retail sector, as even the largest market 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 80. 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 1: Sample Statistics

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

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

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

3 Theoretical Framework

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

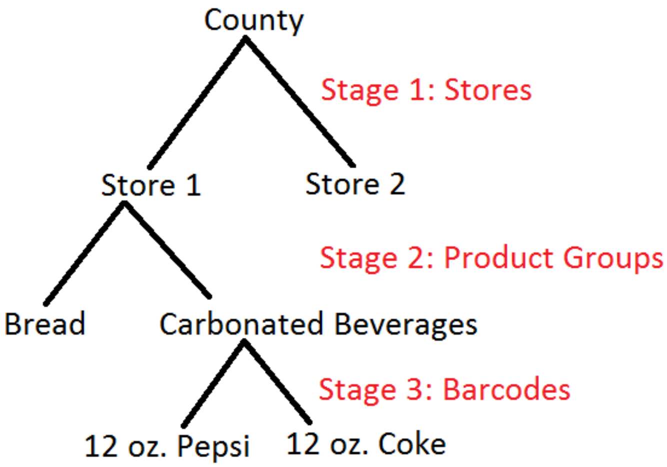
3.1 Market definition

The market definition I use for my benchmark case is the county, so stores only 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 3 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). The county market definition is more disaggregated than using the metropolitan statistical area (MSA), the market definition used in a recent FTC analysis of the impacts 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 which 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 (eg. 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 (eg. 12 oz. Coke) to purchase based on the barcode prices. The demand of the representative consumer will be constant elasticity of substitution (CES) demand at every stage. This is isomorphic to a nested logit model with a population of heterogenous consumers who each choose a single option at each stage (Anderson, de Palma and Thisse 1992).

Figure 1: Multi-stage Budgeting



Two reasons motivate my choice of the nested CES functional form for consumer utility. First, this 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). The constant markup CES model is an important benchmark and the monopolistically competitive limit case in Dhingra and Morrow (2013). The CES model is also used in the literature on consumer gains from variety (Handbury and Weinstein forthcoming, 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

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

where C_{st} is the consumption index of store s at time t; φ_{st} is the quality of store s at time t; R_{ct} is the set of stores in county c at time t; and σ_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

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

where C_{gst} is the consumption index of product group g from store s at time t; φ_{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 σ_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

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

where C_{ust} is the consumption of upc u from store s at time t; φ_{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

σ_{U_g} is the constant elasticity of substitution across upcs within product group g within the store.

Since the utility function is homogeneous of degree one in quality, I will need to choose a normalization of the quality parameters[4]. 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.

While I could choose the same normalization for store quality, I will instead choose a different normalization. I pick the largest drugstore (by sales) which 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 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

$$S_{ust} = \frac{\left(P_{ust}/\varphi_{ust}\right)^{1-\sigma_u}}{\sum_{k \in U_{gst}} \left(P_{kst}/\varphi_{kst}\right)^{1-\sigma_u}}, \qquad \sigma_u > 1,\ \varphi_{kst} > 0 \tag{5}$$

where P_{ust} is the retail price of upc u at store s at time t; φ_{ust} is the quality of barcode 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 σ_U is the constant elasticity of substitution across barcodes in product group g.

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

[4] This will not matter for any of my main results.

$$P_{gst} = \left[\sum_{k \in U_{gst}} \left(\frac{P_{kst}}{\varphi_{kst}} \right)^{1-\sigma_u} \right]^{\frac{1}{1-\sigma_u}} \tag{6}$$

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

$$S_{gst} = \frac{\left(P_{gst} / \varphi_{gst}\right)^{1-\sigma_g}}{\sum_{k \in G_{st}} \left(P_{kst} / \varphi_{kst}\right)^{1-\sigma_g}}, \qquad \sigma_g > 1,\ \varphi_{kst} > 0 \tag{7}$$

where P_{gst} is the product group price index given by equation 6; φ_{gst} is the quality of product group g at store s at time t; G_{st} is the set product groups at store s at time t; and σ_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

$$P_{st} = \left[\sum_{k \in G_{st}} \left(\frac{P_{kst}}{\varphi_{kst}} \right)^{1-\sigma_g} \right]^{\frac{1}{1-\sigma_g}} \tag{8}$$

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

$$S_{sct} = \frac{\left(P_{st} / \varphi_{st}\right)^{1-\sigma_s}}{\sum_{k \in R_{ct}} \left(P_{kt} / \varphi_{kt}\right)^{1-\sigma_s}}, \qquad \sigma_s > 1,\ \varphi_{kst} > 0 \tag{9}$$

where P_{st} is the store price index given by equation 8; φ_{st} is the quality of store s at time t; R_{ct} is the set of stores in county c at time t; and σ_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

$$P_{ct} = \left[\sum_{k \in R_{ct}} \left(\frac{P_{kt}}{\varphi_{kt}} \right)^{1-\sigma_s} \right]^{\frac{1}{1-\sigma_s}} \tag{10}$$

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

$$E_{ust} = S_{ust} S_{gst} S_{sct} E_{ct} \tag{11}$$

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

$$Q_{ust} = \frac{E_{ust}}{P_{ust}} \tag{12}$$

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

$$Q_{ust} = \varphi_{st}^{\sigma_s - 1} \varphi_{gt}^{\sigma_g - 1} \varphi_{ut}^{\sigma_u - 1} E_{ct} P_{ct}^{\sigma_s - 1} P_{st}^{\sigma_g - \sigma_s} P_{gt}^{\sigma_u - \sigma_g} P_{ust}^{-\sigma_u} \tag{13}$$

3.3 Retailer problem

I will define the retail chain as the parent company which owns the retail stores. This is a substantive assumption only when the same parent company owns multiple retail banners (ie. 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 case of each store 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 impacts 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.[5]

[5] See D'Aspremont et al. (1996) for a discusson 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

$$V_{ust}\left(Q_{ust}\right) = z_{ust}Q_{ust}^{1+\delta_g} \tag{14}$$

where Q_{ust} is the total quantity supplied of barcode u by store s; δ_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 for $\delta_g > 0$ 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

$$m_{ust} = \left(1+\delta_g\right) z_{ust}Q_{ust}^{\delta_g} \tag{15}$$

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:

$$\pi_{rct} = \sum_{u\epsilon U_{rct}} \left[P_{ust}Q_{ust} - V_{ust}(Q_{ust})\right] - H_{ct} \tag{16}$$

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

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

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

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

$$P_{ust} = \mu_{rct}m_{ust} \tag{18}$$

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

This markup is given by

$$\mu_{rct} = \frac{\varepsilon_{rct}}{\varepsilon_{rct} - 1}, \tag{19}$$

where ε_{rct} is retail chain r's perceived elasticity of demand in county c at time t and is given by

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

where σ_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 now the retail chain r's perceived elasticity of demand in county c at time t is given by

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

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. 2012, Hottman et al. 2014). 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 σ_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 impacts 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:

$$W_{ct} = \frac{E_{ct}}{P_{ct}} \tag{22}$$

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 into equation 10, I can re-write the county price index as

$$\ln P_{ct} = \ln \tilde{P}_{st} - \frac{1}{\sigma_s - 1} \ln N_{ct} - \frac{1}{\sigma_s - 1} \ln \left[\frac{1}{N_{ct}} \sum_{k \in R_{ct}} \frac{S_{kt}}{\tilde{S}_{st}} \right] - \ln \tilde{\varphi}_{st} \tag{23}$$

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. Since 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 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 σ_s, the elasticity of substitution across stores. As $\sigma_s \rightarrow \infty$, so 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 is a measure of share dispersion and will capture the consumer losses from retail misallocation. Since 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 towards lower productivity (or quality) retail stores costs consumers in terms of welfare. The welfare effects of retail misallocation depends on the elasticity of substitution across stores.

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

4 Structural Estimation

This section explains how I estimate the structural model. First, I explain how to recover the unobserved qualities at a given tier of demand given the elasticity of substitution at that tier. Second, I explain how to recover the unobserved markups and retailer marginal costs given the elasticity of substitution across stores. The rest of this section explains 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 σ_u, 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 σ_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 μ_{rct} for either Bertrand or Cournot competition. Retailer marginal costs can then be computed from the observed retail prices from $m_{usct} = \frac{p_{usct}}{\mu_{rct}}$.

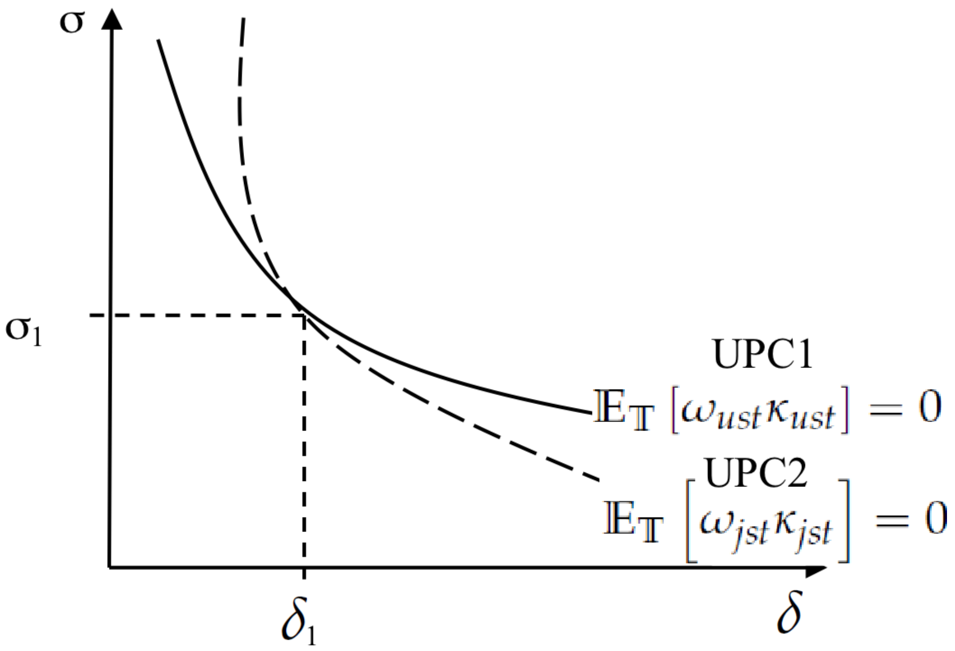
4.2 Estimating the elasticities of substitution

4.2.1 Lowest-tier of demand

Estimation of σ_u in the lowest-tier of demand follows the approach in Broda and Weinstein (2010), based on Feenstra (1994). A similar idea for achieving 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, σ_u and δ_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 (σ_u, δ_g) space which best fits the observed price and share data of that barcode. This can be seen in Figure 2. The orthogonality assumption alone does not provide identification: a higher value of σ_u but a lower value of δ_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.[6] Since the slopes of the demand and supply curves are the same, the intersection of the 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 generalized method of moments (GMM) procedure for estimating the slopes of demand and supply for each product group.

[6] I can reject the null of homoskedasticity in a White test for generalized heteroskedasiticty for the product groups in the data.

Figure 2: Identification


σ
σ1
UPC1
$\mathbb{E}_{\mathbb{T}}\left[\omega_{ust}\kappa_{ust}\right]=0$
UPC2
$\mathbb{E}_{\mathbb{T}}\left[\omega_{jst}\kappa_{jst}\right]=0$
δ1
δ

Start from the demand equation 5, take the time difference and difference relative to another barcode in the same brand, product group, and store. This double-differencing gives

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

where the unobserved error term is $\omega_{ust}=(1-\sigma_U)\left[\triangle^t\ln\varphi_{kst}-\triangle^t\ln\varphi_{ust}\right]$.

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

$$\triangle^{k,t}\ln P_{ust}=\frac{\delta_g}{1+\delta_g}\triangle^{k,t}\ln S_{ut}+\kappa_{ust}, \tag{25}$$

where the unobserved error term is $\kappa_{ust}=\frac{1}{1+\delta_g}\left[\triangle^t\ln z_{usct}-\triangle^t\ln z_{ksct}\right]$.

The orthogonality condition for each barcode is then defined as

$$G(\beta_g)=\mathbb{E}_{\mathbb{T}}\left[x_{ust}(\beta_g)\right]=0 \tag{26}$$

where $\beta_g = \begin{pmatrix} \sigma_U \\ \delta_g \end{pmatrix}$ and $x_{ust} = \omega_{ust}\kappa_{ust}$.

This condition assumes the orthogonality of the idiosyncratic demand and supply shocks at the barcode level, since 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

$$\hat{\beta}_g = \arg\min_{\beta_g} \left\{ G^*(\beta_g)' W G^*(\beta_g) \right\} \tag{27}$$

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 σ_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 this relative to another product group within the same store s to get

$$\Delta^{g,t} \ln S_{gst} = \left(1 - \sigma_g\right) \Delta^{g,t} \ln P_{gst} + \omega_{gst}, \tag{28}$$

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

Ordinary least squares estimation of equation 28 is expected to be biased due to endogeneity, since 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, since barcode supply curves are upward sloping. Estimation of σ_g will therefore use an instrumental variables approach as in Hottman et al. 2014.

Note that the double-differenced CES product group price index can be written as[7]

$$\Delta^{g,t} \ln P_{gst} = \Delta^{g,t} \ln \tilde{P}_{ust} + \frac{1}{1-\sigma_U} \Delta^{g,t} \ln \left[\sum_{u \in U_{gst}} \frac{S_{ust}}{\tilde{S}_{ust}} \right] \tag{29}$$

[7] This is under the normalization that $\tilde{\varphi}_{ust} = 1$.

where 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 are fully captured in this term.

The second term on the right hand side is a term which 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 is unlikely since 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 (eg. 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 is possibly an issue, although if there is such a correlation, the most plausible case is that product groups which gain relative share within the store add more barcodes relative to the other product groups. This 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 upwards (away from zero). Simulations strongly suggest that the effect of an upwards bias in σ_g is to increase the estimated value of σ_s. As I will discuss in the next section, an upwards bias in the substitutability across stores is not a major problem. This 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 towards zero and thus are lower bound estimates.

Keeping this discussion in mind, I estimate σ_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

$$\mathbb{E}\left[\omega_{gst}\Delta^{g,t}\ln\left[\sum_{u\in U_{gst}}\frac{S_{ust}}{\tilde{S}_{ust}}\right]\right]=0 \tag{30}$$

4.2.3 Upper-tier of demand

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

$$\Delta^{s,t}\ln S_{sct}=\left(1-\sigma_s\right)\Delta^{s,t}\ln P_{st}+\omega_{st}, \tag{31}$$

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

As in the middle-tier of demand, estimation of σ_s will use an instrumental variables approach. Note that the store price index can be written as[8]

$$\Delta^{s,t}\ln P_{st}=\frac{1}{1-\sigma_g}\Delta^{s,t}\ln\left[\sum_{g\in G_{st}}\frac{S_{gst}}{\tilde{S}_{gst}}\right]+\Delta^{s,t}\{\frac{1}{N_{Gst}}\sum_{g\in G_{st}}\frac{1}{1-\sigma_U}\ln\left[\sum_{u\in U_{gst}}\frac{S_{ust}}{\tilde{S}_{ust}}\right]\}+\Delta^{s,t}\{\frac{1}{N_{Gst}}\sum_{g\in G_{st}}\ln\tilde{P}_{ust}\} \tag{32}$$

I estimate σ_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

$$\mathbb{E}\left[\omega_{st}\Delta^{s,t}\{\ln\left[\sum_{g\in G_{st}}\frac{S_{gst}}{\tilde{S}_{gst}}\right]+\frac{1}{N_{Gst}}\sum_{g\in G_{st}}\ln\left[\sum_{u\in U_{gst}}\frac{S_{ust}}{\tilde{S}_{ust}}\right]\}\right]=0 \tag{33}$$

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 common across the retail chain quarterly demand and supply shocks within the county (eg. 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, the 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 would induce a negative

[8] This is under the normalization that $\tilde{\varphi}_{gst}=1$.

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 upwards (away from zero). However, an upwards bias in the substitutability across stores is not a significant problem. In that case, my results on the distortions from misallocation, the differences in markups across cities, and the gains from store variety are all biased towards zero and thus are lower bound estimates.

5 Estimation Results

5.1 Model parameters

Table 2 shows the results of estimating σ_u and δ_g for 106 product groups. As expected, OLS estimates of σ_u are much lower than the GMM estimates based on Feenstra (1994). The GMM estimates are reasonably precise. The confidence intervals for σ_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. Since the estimates of δ_g are all less than 1, this implies that marginal cost is inelastic with respect to quantity. This is consistent with the results in Gagnon and López-Salido (2014), who find using different data and methods that supermarket supply curves are relatively flat in the short-run. On the other hand, the median σ_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 7%.[9] 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 a response to price changes.

[9] This is assuming that the barcode has a near-zero market share within its product group.

Table 2: Distribution of σ_u and δ_g Estimates

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

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

The estimates of σ_u can also be compared to the estimated σ_g at the middle-tier of demand, shown in Table 3. I estimate σ_g to be 4.8 using instrumental variables (IV). This is less than σ_u for the vast majority of product groups. This implies that barcodes are typically more substitutable within product groups than they are across product groups.

Table 3 also shows the estimation result for σ_s in the upper-tier of demand. The estimated σ_s is 4.5 using IV. This means that a one percent increase in a store's price index reduces that store's market share by 4.5%. The point estimate of σ_s is less than σ_g, which implies that barcodes are more substitutable within stores than across stores. However, I cannot statistically reject that σ_s is equal to σ_g.

Table 3: σ_g and σ_s

	σ_g (95% CI)	σ_s (95% CI)
OLS	**1.1** (1.1, 1.1)	**1.5** (1.4, 1.6)
IV	**4.8** (4.6, 5.0)	**4.5** (4.3, 4.7)

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

5.2 Markup estimates

Table 4 shows the retail chain markups[10] implied by the estimated model parameters. The estimated markups are reasonable. The monopolistically competitive markup is 28%[11]. To get a sense of the plausability of my parameter estimates, you can compare my markup

[10] Markups are defined as (price - marginal cost)/marginal cost.

[11] Note that my OLS estimate of σ_s would imply an absurd monopolistically competitive markup of 200%.

estimates to retail markup estimates obtained using very different data and methods. For comparison, in the Census of Retail Trade, the average retail markup is 0.39 (Faig and Jerez 2005). This is broadly comparable to what I estimate.

Table 4: Distribution of Markup Estimates

Percentile	Markup using Bertrand	Markup using Cournot
1	0.28	0.29
5	0.29	0.29
10	0.29	0.29
25	0.29	0.30
50	0.29	0.33
75	0.30	0.38
90	0.33	0.51
95	0.36	0.62
99	0.44	0.99

Note: Markup = (Price-Marginal Cost)/(Marginal Cost). Calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

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 that is 21% higher than the median markup. There are counties for which the largest retail chains have substantially higher markups. However, the markups of the largest chains are still reasonable.

Table 5: Distribution of Markups Relative to County Median

Percentile	Bertrand Markup Largest Chain	Cournot Markup Largest Chain
1	0.29	0.32
5	0.30	0.34
10	0.30	0.36
25	0.31	0.41
50	0.34	0.52
75	0.38	0.72
90	0.49	1.22
95	0.71	2.19
99	1.12	4.08

Note: Markup = (P-MC)/MC. Calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

5.3 Quantifying the losses from retail misallocation

Having estimated the model parameters and retail markups, I am now in a position to be able to quantify the misallocation from retailer markup dispersion. The procedure for doing this is as follows. Remember that the county price index can be written as:

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

where the first term is the log 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. This holds the level of markups and the first term in the county price index fixed (log of geometric mean of store price indices). The consumer gains from removing markup distortion are captured in the third term in county price index.

To find the consumer gain, start by recomputing each store's price index using the geometric mean markup. 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 that is difficult to do given the size of the data. From equation 9, solve for the new equilibrium market shares for each store. Then use the new 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. In order to compute the change in aggregate consumer welfare, aggregate utility will be Cobb-Douglas across counties.

Table 6 shows the results of this exercise. In the Bertrand case, consumer welfare rises by 1% after removing markup dispersion. This change in markups benefits consumers by $918 million dollars 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.

Consumer welfare gains and the deadweight loss are larger in the Cournot case. This is 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

	%$\triangle$ Consumer Welfare	Consumer Surplus ($/Year)	Total Surplus ($/Year) (% Total Sales)
Bertrand	1%	$918 million	$302 million (0.3%)
Cournot	4.6%	$4.4 billion	$2.2 billion (2.3%)

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

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 is as follows. We imagine a price regulator forces each retailer to charge the monopolistically competitive markup. This 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 by 2.6% after moving markups to the monopolistically competitive level. This change in markups benefits consumers by $3 billion dollars 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. This is 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 7: Welfare Gains from Moving Markups to Monop. Comp. Limit

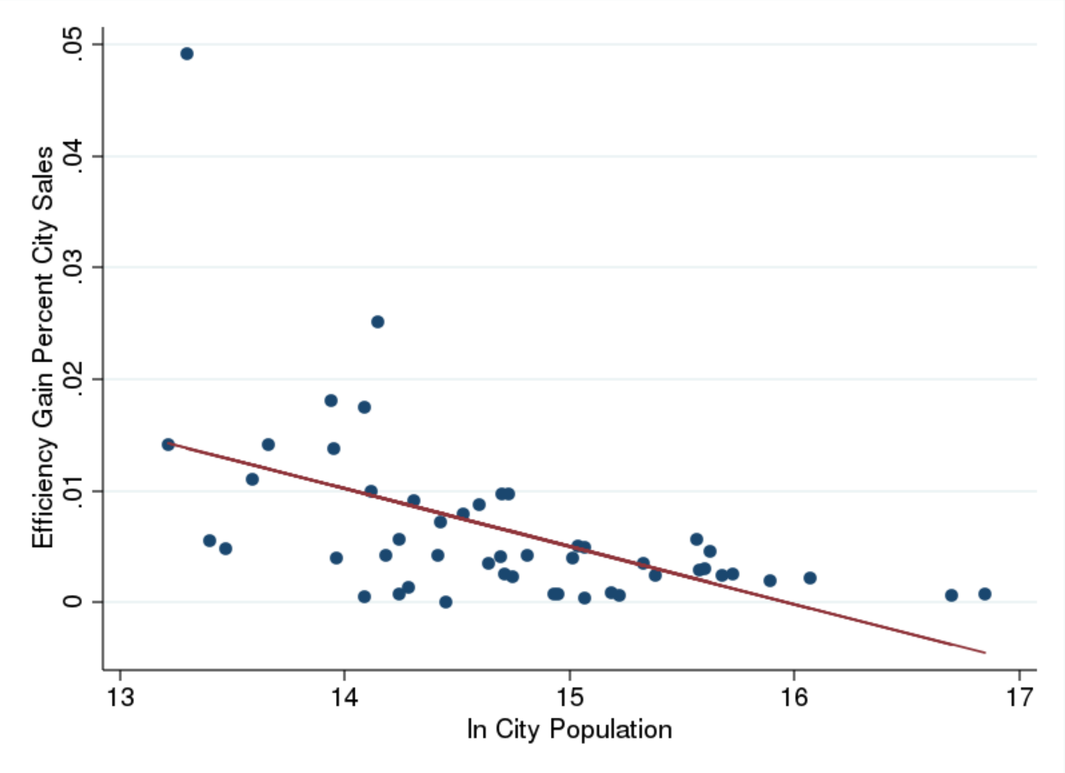
	%$\triangle$ Consumer Welfare	Consumer Surplus ($/Year)	Total Surplus ($/Year) (% Total Sales)
Bertrand	2.6 %	$3 billion	$868 million (0.8 %)
Cournot	11.6 %	$14 billion	$6 billion (5.3 %)

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

The losses from retail misallocation are about the same magnitude as the losses from producer misallocation in the US due to 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 2014).

Figure 3 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 almost 0 for New York City and Los Angeles. These two cities are close to being efficient. The next section of results explores this further by highlighting the reason why the efficiency gains fall with city size.

Figure 3: Misallocation Deadweight Losses by City Size



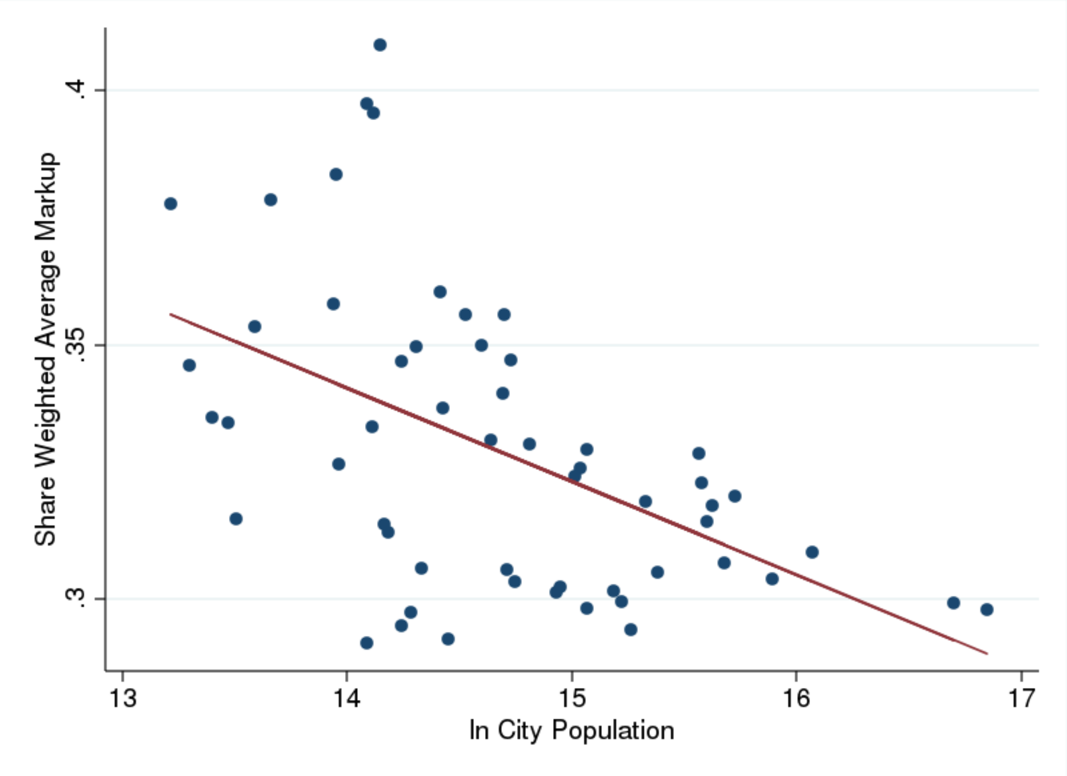
Note: Calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

5.4 Retail markups and city size

With estimated markups in hand, I can now investigate if 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 (eg. Behrens and Murata 2009).

Figure 4 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 has a share weighted average markup about 10 percentage points lower than Des Moines. This suggests that competition is indeed tougher in larger cities. Since σ_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 σ_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.

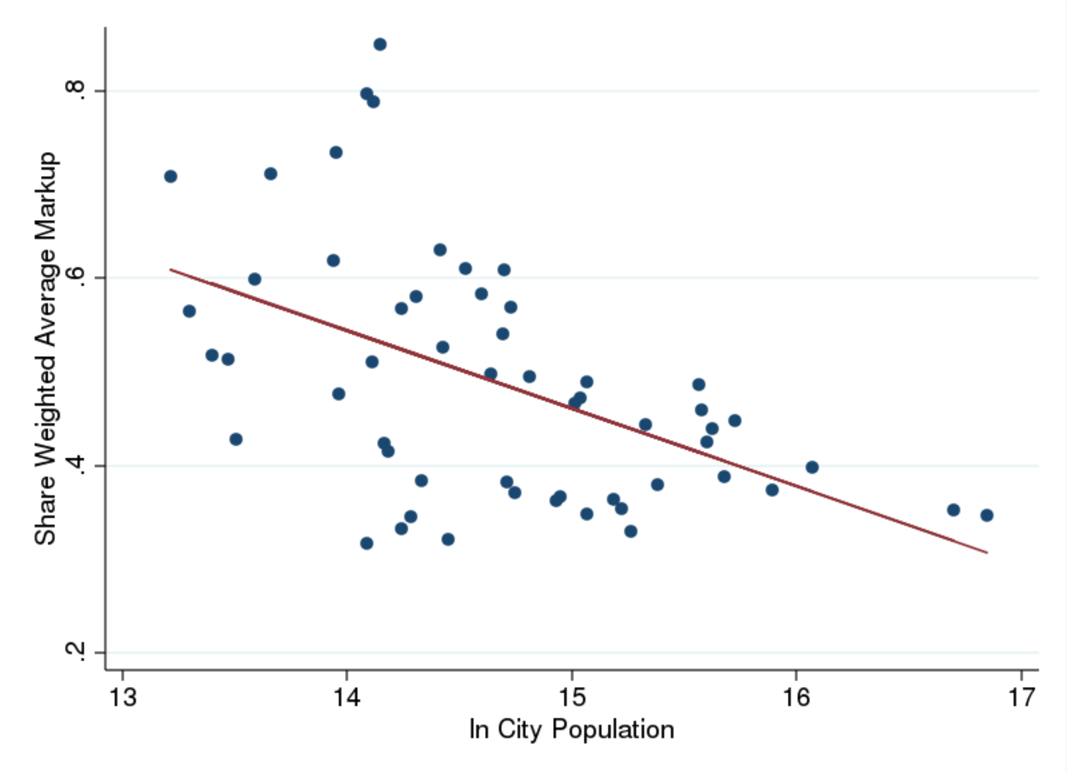
Figure 4: Bertrand Markups by City Size



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

Figure 5 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. 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.

Figure 5: Cournot Markups by City Size



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

To sum up the results of this section, I find that larger cities have lower markups. This is the case whether retail competition is Bertrand or Cournot. The difference between the two types of retail competition is that 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 and the deadweight losses from misallocation. Thus, the answer to the question of "How large is large?" turns out to be "about the size of New York City".

5.5 Gains from retail store variety

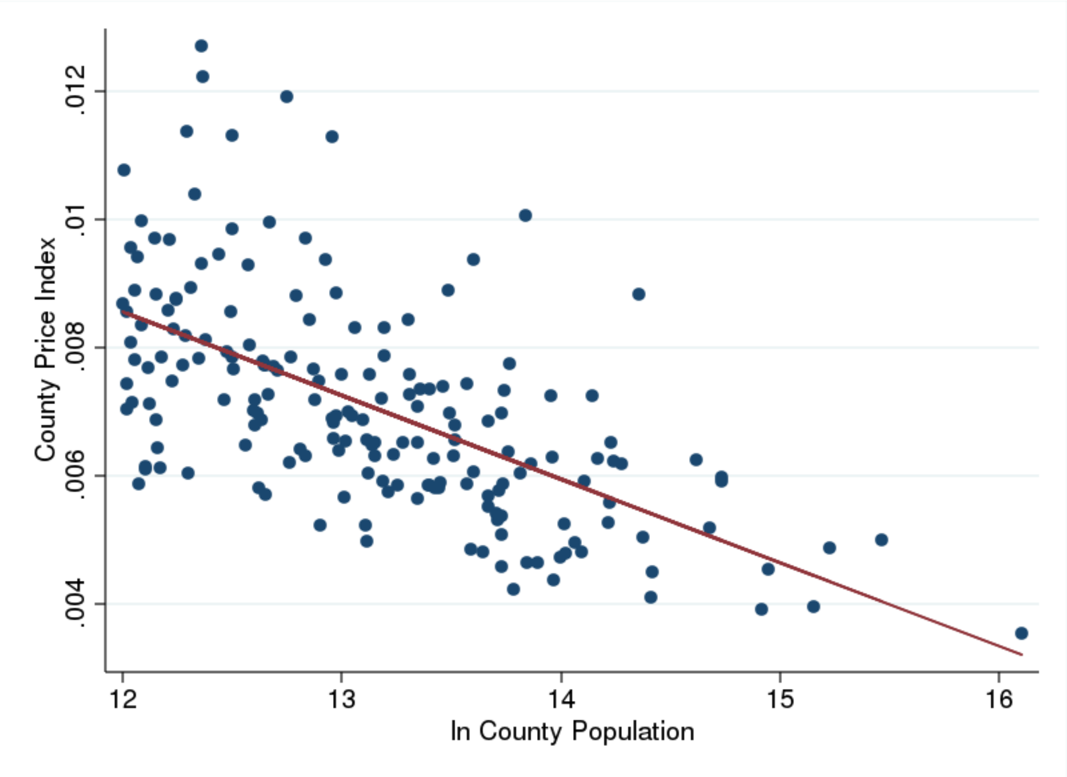
Next, I will consider the gains from retail store variety. The magnitude of differences in gains from store variety across counties depends on differences in the number of stores across counties. We saw earlier in the data section that there are large differences in the number of stores across US counties.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

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

The first term captures the average product variety adjusted store price index. The second term captures the gains from retail store variety. Since the estimated 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 is evidence for important consumption-based agglomeration forces in the US.

Figure 6: County Price Index by County Size

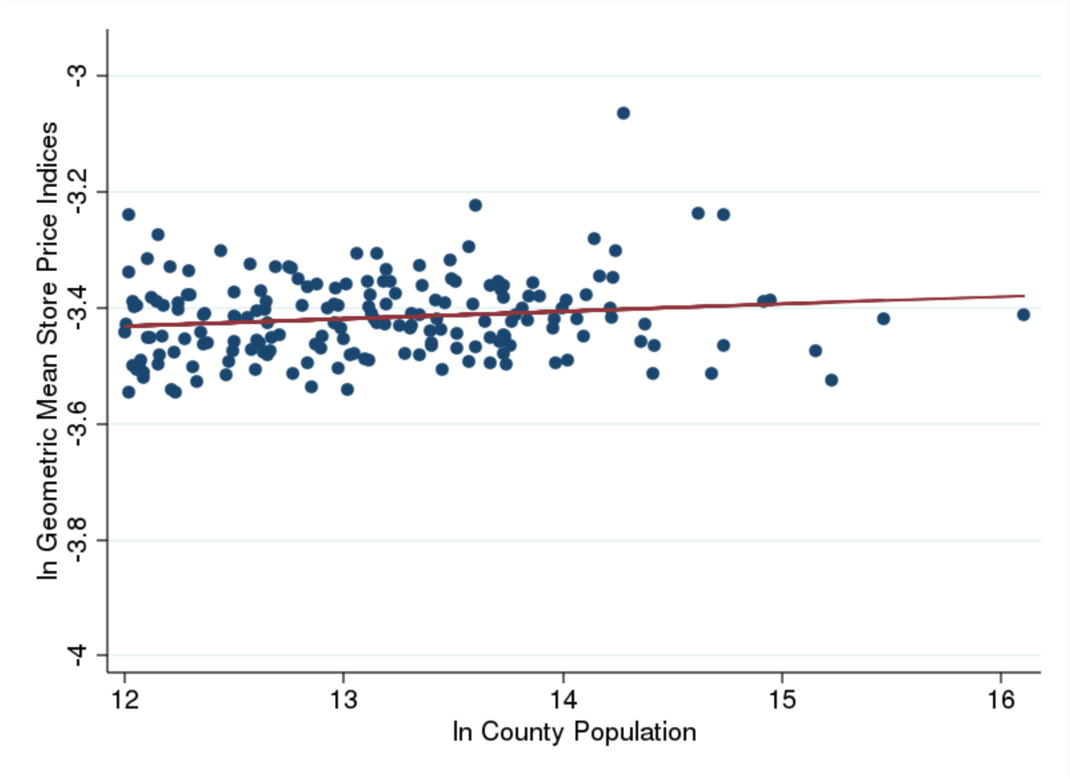


Note: Calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

Since the county price index reflects both retail store variety and differences in tradable product variety across counties, I will decompose it into the 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 shows that even accounting for product variety at the store level, the average store in the largest counties has a higher price index than 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.

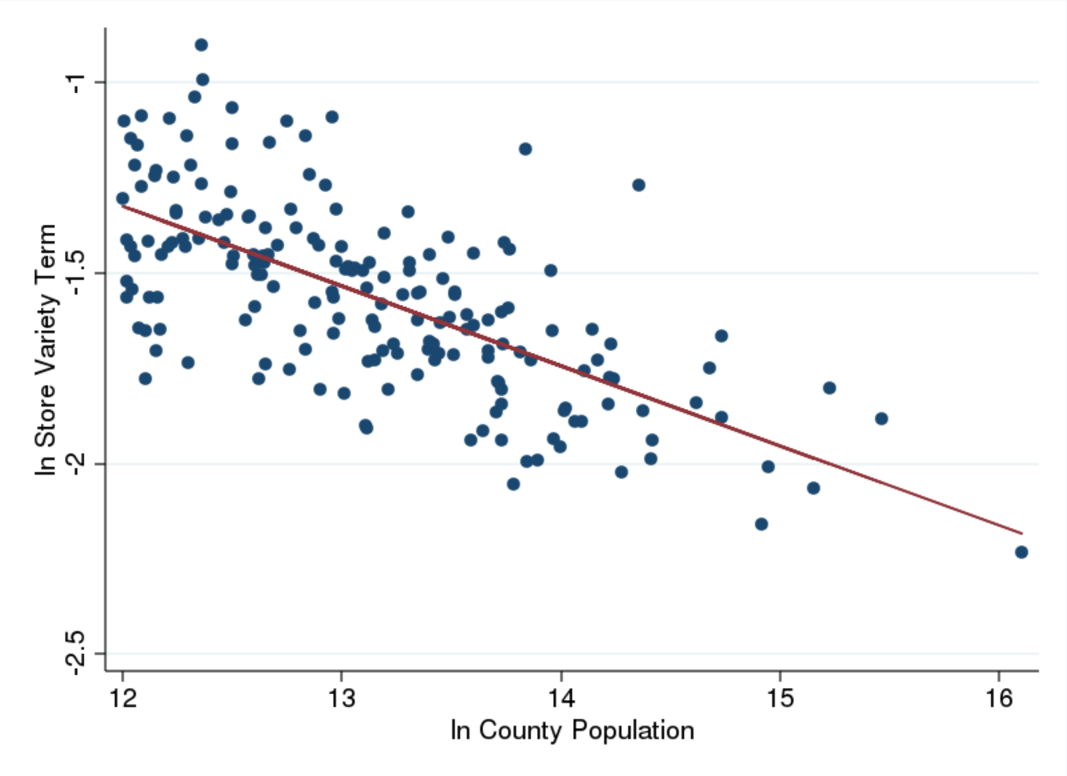
Figure 7: Average Store Price Index by County Size



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

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.

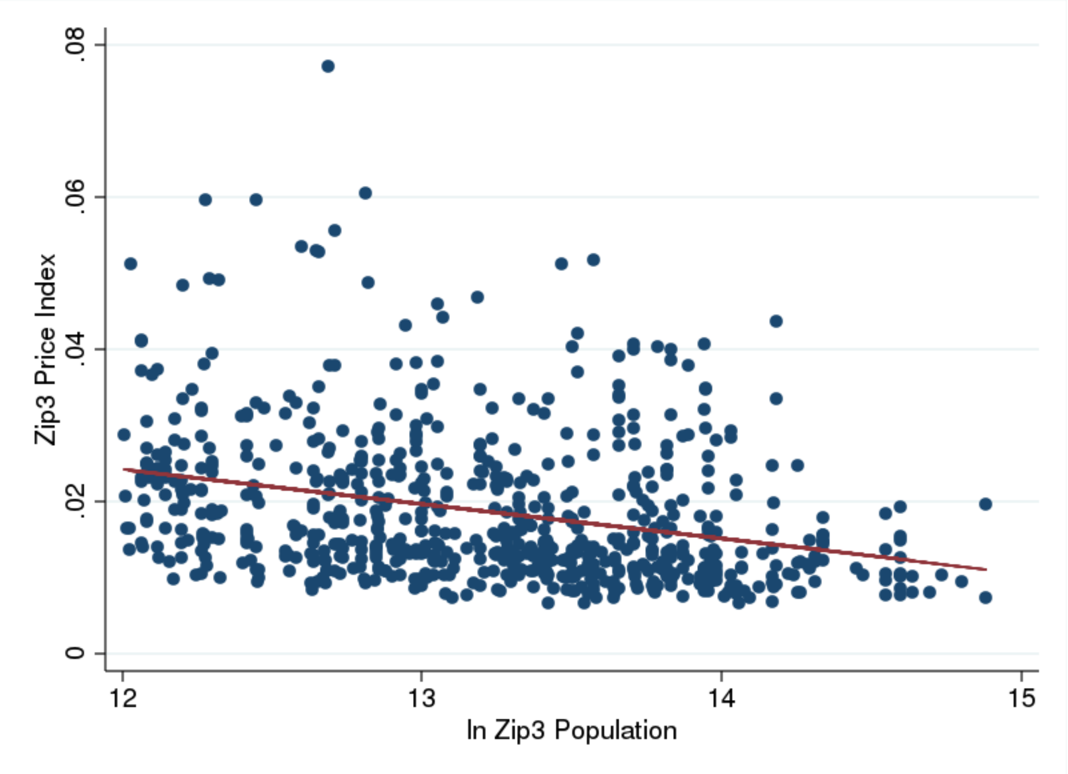
Figure 8: Store Variety Term by County Size



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

A concern one might have with this store variety result is that it may not be surprising that large counties contain a lot more stores. The largest county in particular, Los Angeles County, is very large. To consider the robustness of the variety result, Figure 9 shows the retail store variety adjusted price index using truncated (first 3 digit) zip codes instead of counties. There are about 2.5 times more 3 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.

Figure 9: Store Variety Term by County Size



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

5.6 Robustness

In this section, I investigate the robustness of my results with regards to three changes. First I estimate different σ_s's in different locations and compare results. Second, I consider what happens if I define the market as the Metropolitan Statistical Area instead of the county. Finally, I conclude by allowing each store to set its own markup instead of coordinating pricing at the retail chain level.

5.6.1 Heterogeneity in σ_s

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

Table 8 reports results for estimating different σ_s parameters for different portions of the city size distribution. Splitting the set of cities in half and estimating a different σ_s

for each half produces estimates not too far on either side of the base σ_s 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: σ_s Estimates by City Size Dist.

	1st Half of Cities by Size	2nd Half of Cities by Size
IV Estimate σ_s (95% CI)	**3.9** (3.5, 4.3)	**4.7** (4.5, 5.0)

City Size Dist.	1st Quartile	2nd Quartile	3rd Quartile	4th Quartile
IV Estimate σ_s (95% CI)	**3.9** (3.4, 4.5)	**3.9** (3.4, 4.4)	**4.5** (4.2, 4.9)	**4.8** (4.5, 5.1)

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

Table 8 also shows results after splitting the data into quartiles by city size. The results are very similar to 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 σ_s estimate of 4.5 is contained in 3 out 4 quartile confidence intervals, and almost contained in the remaining quartile confidence interval.

Table 9 shows results for estimating a different σ_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 σ_s equals 4.5 for the majority of the cities. I conclude that the differences in the estimated σ_s parameters across cities are mostly due to lack of precision from estimating on small sample sizes, since I find no correlation between the estimated σ_s parameters and either city size or density (not shown).

Table 9: σ_s Estimates by MSA

Percentile	σ_s
10	**3.6** (2.2, 4.2)
25	**3.8** (2.6, 5.1)
50	**4.7** (3.1, 6.0)
75	**5.7** (4.1, 7.9)
90	**7.1** (4.9, 10.1)

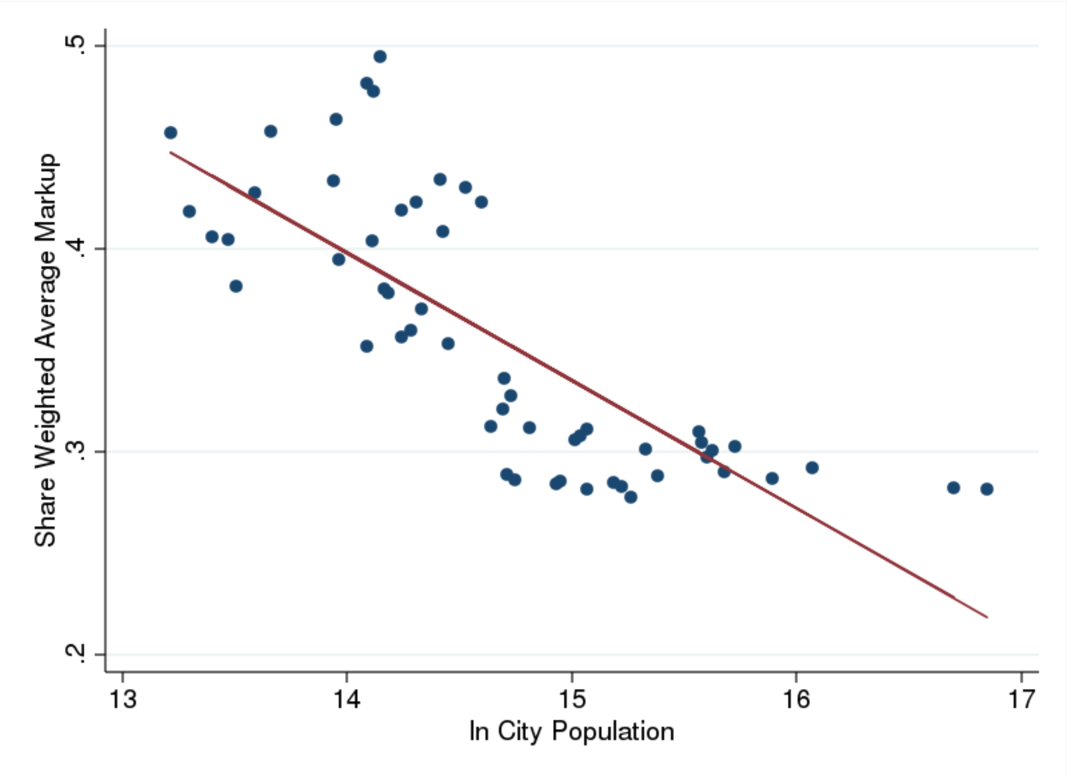
Note: Calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

Having estimated different σ_s parameters across locations, I will next see how these differences matter for the markup results discussed earlier. I will first consider the two σ_s

parameters estimated by splitting the cities into two groups by size. I will then consider the set of city-level σ_s estimates.

Figure 10 shows the relationship between the Bertrand markups and city size when using the σ_s estimates from each half of the city size distribution. Since 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 σ_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.

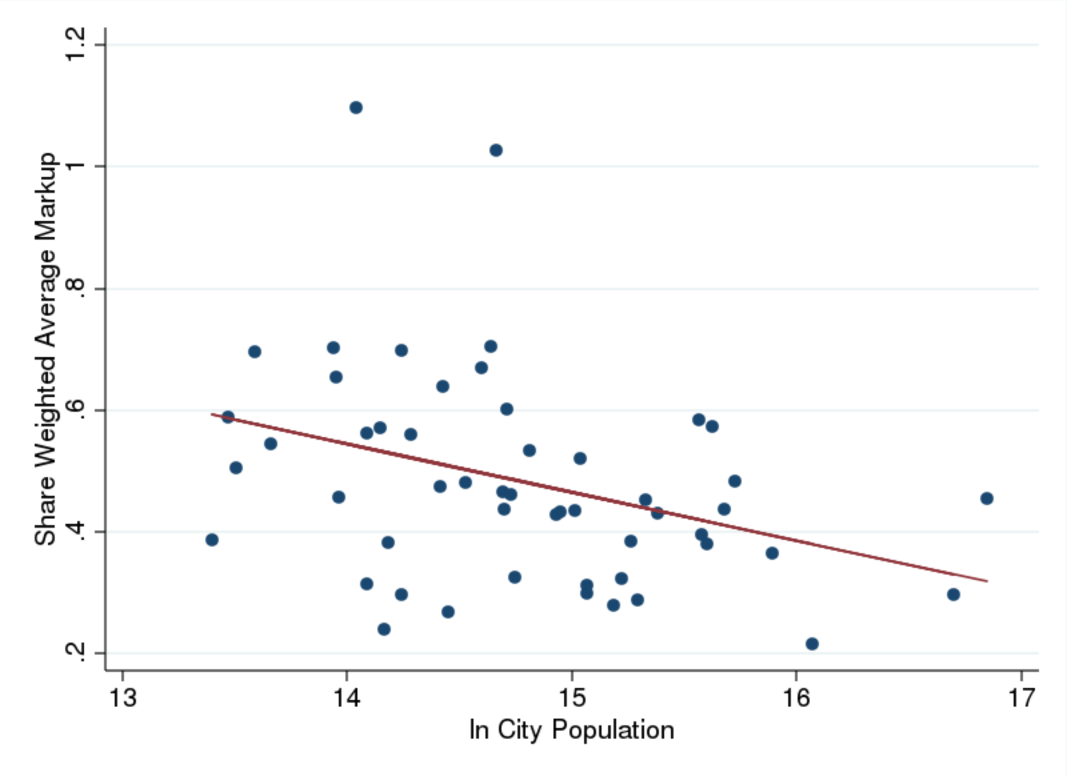
Figure 10: Bertrand Markups for Two σ_s Case



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

Figure 11 shows the relationship between the Cournot markups and city size when I estimate a different σ_s for every city. Larger cities are still estimated to have statistically and economically significantly lower markups in the Cournot case. New York 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 σ_s case. However, the large range of σ_s estimates adds a lot of noise and the Betrand markup slope is no longer statistically significant.

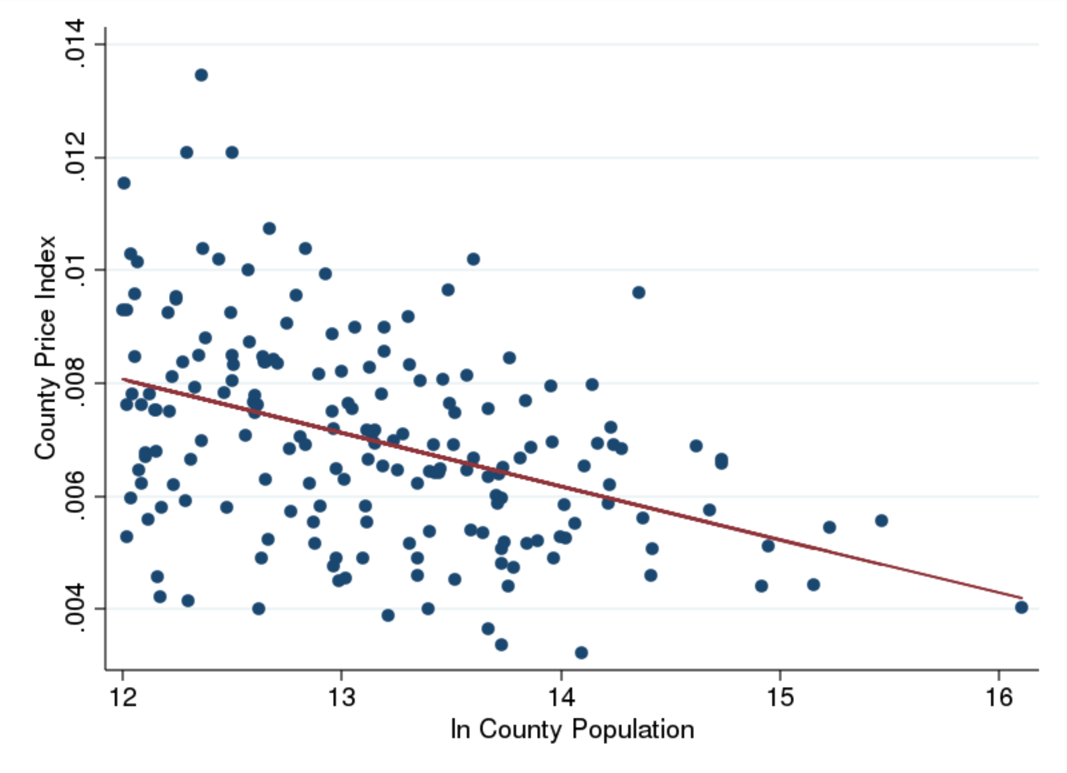
Figure 11: Cournot Markups for MSA σ_s Case



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

Next, consider the robustness of the variation in county price indices to heterogeneity in σ_s. Figure 12 plots the county price indices using the estimate of σ_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.

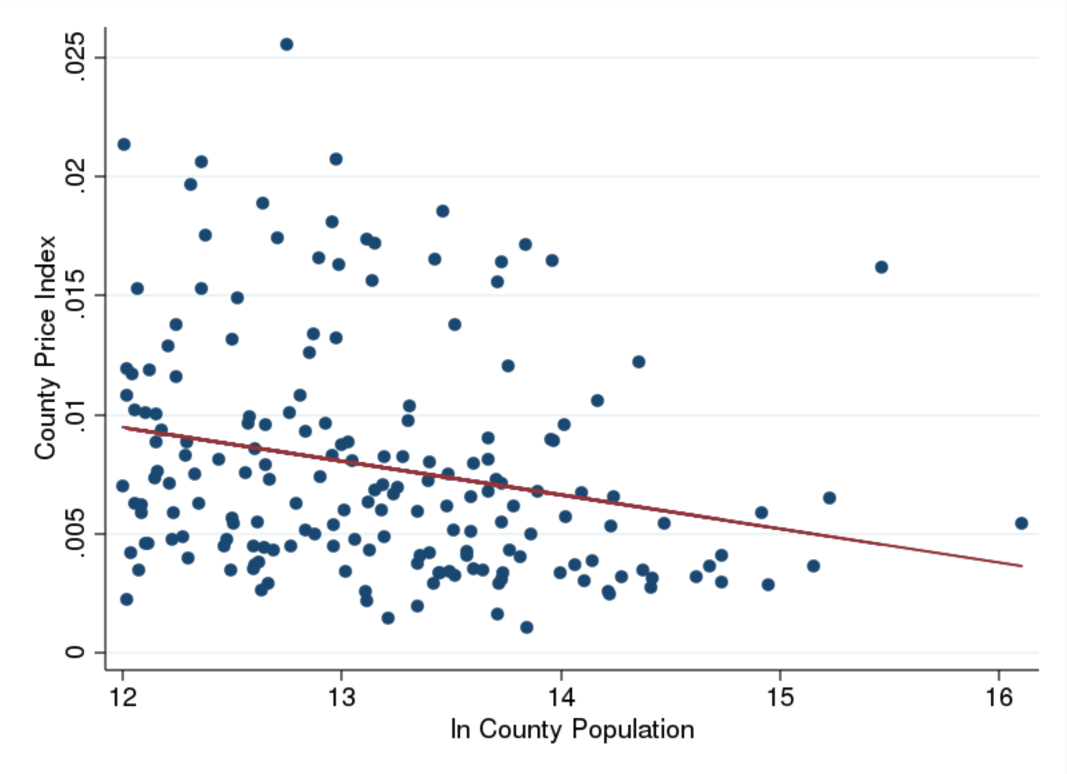
Figure 12: County Price Indices for Two σ_s Case



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

Figure 13 plots the county price indices using the estimates of σ_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 population of 150,000 people have a price index twice as high as the price index for Los Angeles County.

Figure 13: County Price Indices for MSA σ_s Case



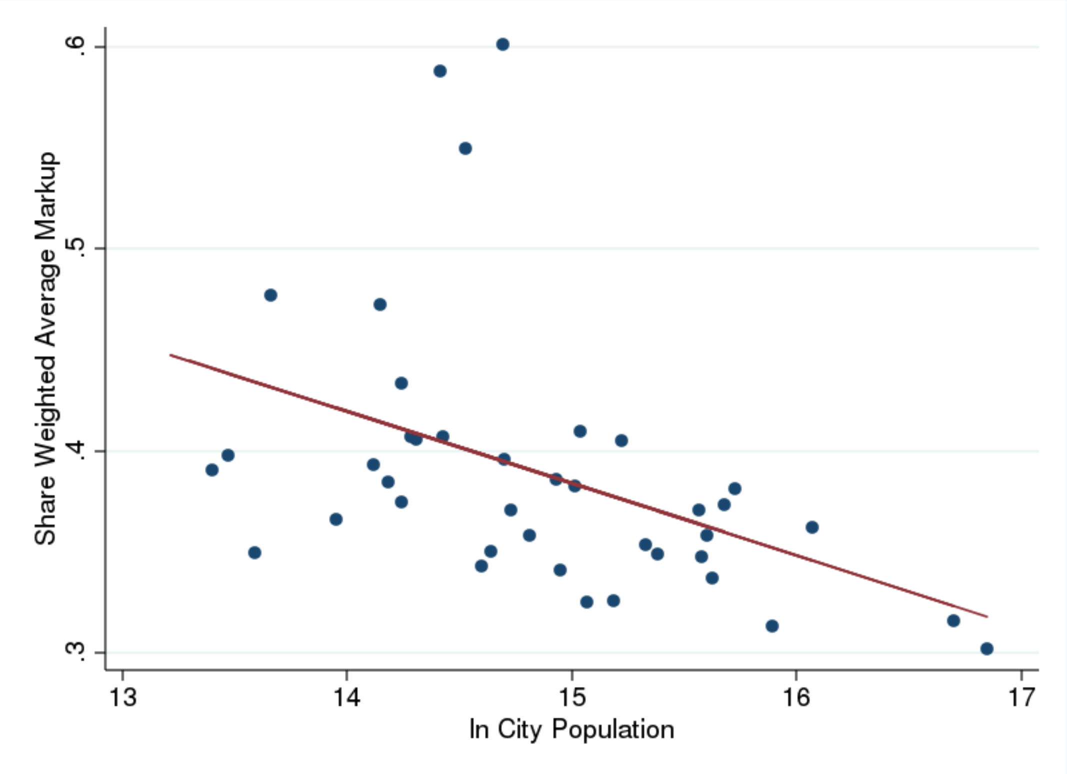
Note: Calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

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 instead the market is defined as the metropolitan statistical area. 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. This relationship is statistically significant at the 1% level. Des Moines is estimated to have about 15 percentage points higher markups relative to New York. The results for the Cournot markups share a similar pattern, with larger variation in markups across cities.

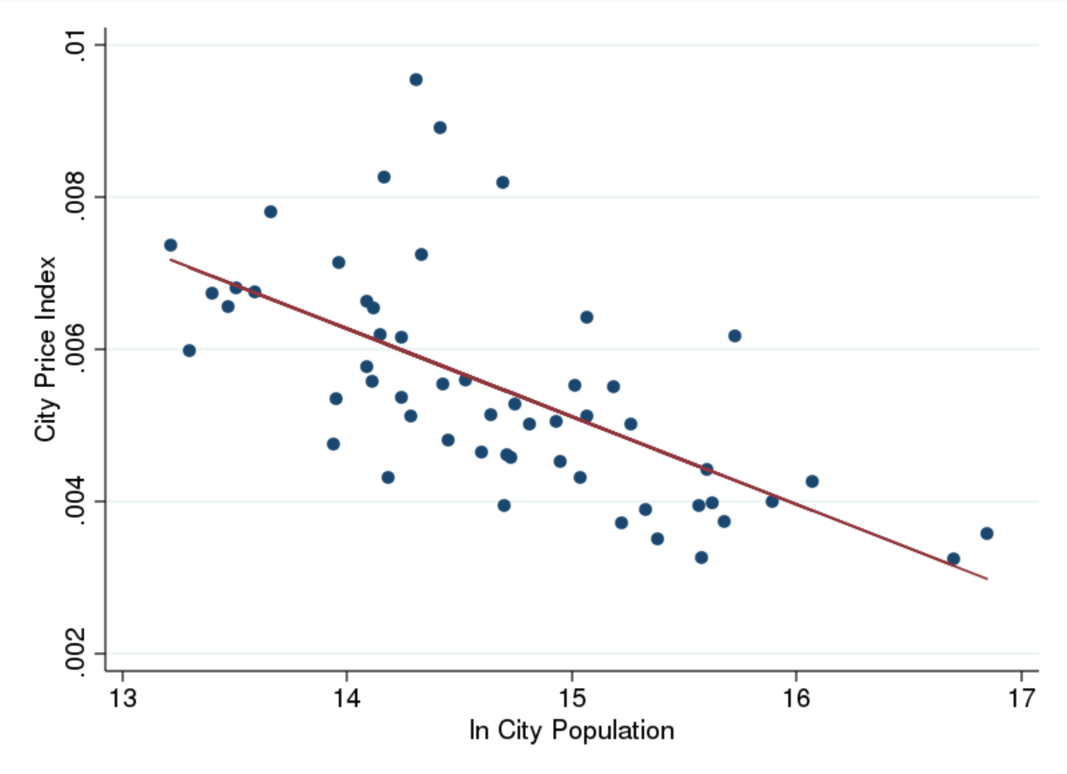
Figure 14: Bertrand Markups using MSA Markets



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

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 still has a price index which is about a half the level of the price index in Des Moines. Surprisingly, using a larger market definition (MSA vs county) did not result in larger differences in price indices across locations.

Figure 15: City Price Indices

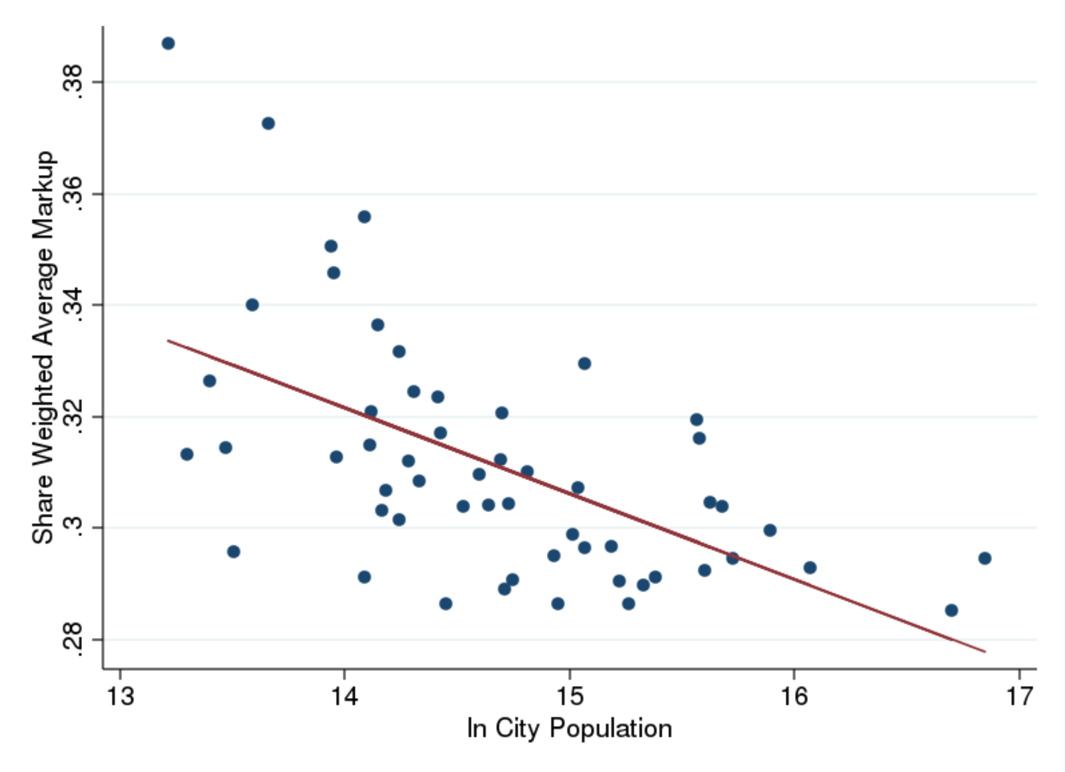


Note: Calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

5.6.3 Store vs 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 aross cities is smaller. In this case, Des Moines has about a 6 percentage point higher markup relative to New York, compared to 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. Howver, the difference in Bertrand markups across cities is only a few percentage points in this case.

Figure 16: Cournot Markups using Store Market Shares



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

6 Conclusion

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

My estimates show that losses from retail misallocation are economically significant. Misallocation losses are between 1% to 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 US.

I also 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 vs metropolitan statistical area) and assumptions about which decision-making unit sets markups (eg. the retail chain or the individual stores).

My estimates imply that retail store variety significantly impacts 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 (eg. Los Angeles County) relative to counties with populations of 150,000 people (eg. 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 3-digit zip code areas instead of counties.

My results have important implications for policy. Any policy change that can reduce concentration (eg. loosening policies that prevent entry, such as the policies used by some localities to prevent entry of big box retailers), may be welfare improving, particularly in the smallest cities. Policies that prevent further concentration (eg. merger policy) are also important. On the other hand, allowing entry of big box retailers, in so far as this 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 much larger differences in consumer prices indices across locations in the US than prior work in the literature. This suggests that consumption-based agglomeration forces may be even more important than economists have previously realized. 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 (eg. rents) across US 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 US cities.

References

[1] Anderson, Simon P., de Palma, André, and Jacques-François Thisse, *Discrete Choice Theory of Product Differentiation*, Cambridge MA: MIT Press, 1992.

[2] Atkeson, Andrew, and Ariel Burstein, "Pricing to Market, Trade Costs, and International Relative Prices", American Economic Review, Vol. 98, No. 5, 2008, pp. 1998-2031.

[3] Atkin, David, Faber, Benjamin, and Marco Gonzalez-Navarro, "Retail Globalization and Household Welfare: Evidence from Mexico", Working Paper, 2014.

[4] Badinger, Harald, "Market Size, Trade, Competition, and Productivity: Evidence from OECD Manufacturing Industries", *Applied Economics*, Vol. 39, No. 17, pp. 2143-2157, 2007.

[5] Baldwin, Richard, and Toshihiro Okubo, "Heterogenous Firms, Agglomeration, and Economic Geography: Spatial Selection and Sorting", *Journal of Economic Geography*, Vol. 6, 2006, pp. 323-346.

[6] Banerjee, Abhijit, and Esther Duflo, "Growth Theory through the Lens of Development Economics", in Aghion, Philippe, and Steven Durlauf (eds), *Handbook of Economic Growth*, Vol. 1A, North-Holland, 2005, pp. 473-552.

[7] Bartelsman, Eric, Haltiwanger, John, and Stefano Scarpetta, "Cross-Country Differences in Productivity: The Role of Allocation and Selection", *American Economic Review*, Vol. 103, No. 1, 2013 (February), pp. 305-334.

[8] Behrens, Kristian, and Yasusada Murata, "City Size and the Henry George Theorem Under Monopolistic Competition", *Journal of Urban Economics*, Vol. 65, No.2, 2009, pp. 228-235.

[9] Behrens, Kristian, and Frédéric Robert-Nicoud, "Survival of the Fittest in Cities: Urbanization and Inequality", Working Paper, 2013.

[10] Behrens, Kristian, Mion, Giordano, Murata, Yasusada, and Jens Südekum, "Spatial Frictions", Working Paper, 2013.

[11] Bellone, Flora, Musso, Patrick, Nesta, Lionel, and Frederic Warzynski, "International Trade and Firm-level Markups when Location and Quality Matter", Working Paper, 2014.

[12] Berry, Steven, and Joel Waldfogel, "Product Quality and Market Size", *Journal of Industrial Economics*, Vol. 58, No. 1, pp. 1-31, 2010

[13] Broda, Christian, and David E. Weinstein, "Product Creation and Destruction: Evidence and Price Implications", *American Economic Review*, Vol. 100, 2010, pp. 691-723.

[14] Campbell, Jeffrey, and Hugo Hopenhayn, "Market Size Matters", *Journal of Industrial Economics*, Vol. 53, 2005, pp. 1-25.

[15] Burstein, Ariel, and Christian Hellwig, "Prices and Market Shares in a Menu Cost Model", *NBER Working Paper*, 13455, 2007.

[16] Campbell, Jeffrey, "Competition in Large Markets", *NBER Working Paper,* 11847, 2005.

[17] Combes, Pierre-Philippe, Duranton, Gilles, Gobillon, Laurent, Puga, Diego, and Sébastien Roux, "The Productivity Advantages of Large Cities: Distinguishing Agglomeration from Firm Selection", *Econometrica*, Vol. 80, 2012, pp. 2543-2594.

[18] Combes, Pierre-Philippe, and Miren Lafourcade, "Competition, Market Access, and Economic Geography: Structural Estimation and Predictions for France", *Regional Science and Urban Economics*, Vol. 41, 2011, pp. 508-524.

[19] Couture, Victor, "Valuing the Consumption Benefits of Urban Density", Working Paper, 2013.

[20] D'Aspremont, Claude, Ferreira, Rodolphe Dos Santos, and Louis-André Gérard-Varet, "On the Dixit-Stiglitz Model of Monopolistic Competition", *American Economic Review*, Vol. 86, No. 3, 1996 (June), pp. 623-629.

[21] Dhingra, Swati, and John Morrow, "Monopolistic Competition and Optimum Product Diversity Under Firm Heterogeneity", Working Paper, 2013.

[22] Dunne, Timothy, Klimek, Shawn, Roberts, Mark, and Daniel Xu, "The Dynamics of Market Structure and Market Size in Two Health Service Industries", in Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts, eds., *Producer Dynamics: New Evidence from Micro Data*, University of Chicago Press, 2009.

[23] Edmond, Chris, Midrigan, Virgiliu, and Daniel Yi Xu, "Competition, Markups, and the Gains from International Trade", *NBER Working Paper*, 18041, 2012.

[24] Epifani, Paolo and Gino Gancia, "Trade, Markup Heterogeneity, and Misallocations", *Journal of International Economics*, Vol. 83, No. 1, 2011, pp. 1-13.

[25] Faig, Miquel, and Belén Jerez, "A Theory of Commerce", *Journal of Economic Theory*, Vol. 122, 2005, pp. 60-99.

[26] Feenstra, Robert, "New Product Varieties and the Measurement of International Prices", *American Economic Review*, Vol. 84, No. 1, 1994 (March), pp. 157-177.

[27] Feenstra, Robert, "Restoring the Product Variety and Pro-Competitive Gains from Trade with Heterogeneous Firms and Bounded Productivity", *NBER Working Paper*, 19833, 2014.

[28] Gagnon, Etienne, and David López-Salido, "Small Price Responses to Large Demand Shocks", *Federal Reserve Board Finance and Economics Discussion Series Working Paper*, 2014.

[29] Gallego, Guillermo, Huh, Woonghee Tim, Kang, Wanmo, and Robert Phillips, "Price Competition with the Attraction Demand Model: Existence of Unique Equilibrium and Its Stability", *Manufacturing and Service Operations Management*, Vol. 8, No. 4, 2006 (Fall), pp. 359-375.

[30] Gilchrist, Simon, Sim, Jae W., and Egon Zakrajsek, "Misallocation and Financial Market Frictions: Some Direct Evidence from the Dispersion in Borrowing Costs", *Review of Economic Dynamics*, Vol. 16, No. 1, 2013 (January), pp. 159-176.

[31] Glaeser, Edward L., Kolko, Jed, and Albert Saiz, "Consumer City", *Journal of Economic Geography*, Vol. 1, 2001, pp. 27-50.

[32] Guesnerie, Roger, and Oliver Hart, "Welfare Losses Due to Imperfect Competition: Asymptotic Results for Cournot Nash Equilibria with and without Free Entry", *International Economic Review*, Vol. 26, No. 3, 1985 (October), pp. 525-545.

[33] Handbury, Jessie, "Are Poor Cities Cheap for Everyone? Non-Homotheticity and the Cost of Living Across US Cities", Working Paper, 2013.

[34] Handbury, Jessie, and David E. Weinstein, "Goods Prices and Product Availability in Cities", *Review of Economic Studies*, forthcoming.

[35] Hanner, Daniel, Hosken, Daniel, Olson, Luke M. and Loren K. Smith, "Dynamics in a Mature Industry: Entry, Exit, and Growth of Big-Box Retailers", *FTC Bureau of Economics Working Paper,* No. 308, 2011.

[36] Hart, Oliver, "Monopolistic Competition in a Large Economy with Differentiated Commodities", *Review of Economic Studies*, Vol. 46, No. 1, 1979 (January), pp. 1-30.

[37] Helpman, Elhanan, "The Size of Regions", in Pines, David, Sadka, Efraim, and Itzhak Zilcha, eds., *Topics in Public Economics*, Cambridge: Cambridge University Press, 1998, pp, 33-54.

[38] Holmes, Thomas J., and John Stevens, "Geographic Concentration and Establishment Scale", *Review of Economics and Statistics*, Vol. 84, 2002, pp. 682-690.

[39] Holmes, Thomas J., Hsu, Wen-Tai, and Sanghoon Lee, "Allocative Efficiency, Mark-ups, and the Welfare Gains from Trade", *Journal of International Economics*, forthcoming.

[40] Hopenhayn, Hugo, and Richard Rogerson, "Job Turnover and Policy Evaluation: A General Equilibrium Analysis", *Journal of Political Economy*, Vol. 101, No. 5, 1993, pp. 915-938.

[41] Hosken, Daniel, Olson, Luke M., and Loren K. Smith, "Do Retail Mergers Affect Competition? Evidence from Grocery Retailing", *FTC Bureau of Economics Working Paper,* No. 313, 2012.

[42] Hottman, Colin, Redding, Stephen J., and David E. Weinstein, "What is Firm Heterogeneity in Trade Models? The Role of Quality, Scope, Markups, and Cost", *NBER Working paper*, 20436, 2014.

[43] Hsieh, Chang-Tai, and Peter Klenow, "Misallocation and Manufacturing TFP in China and India", *Quarterly Journal of Economics*, Vol. 124, No. 4, 2009 (November), pp. 1403-1448.

[44] Leontief, Wassily, "Ein Versuch Zur Statistichen Analyse von Angebot und Nachfrage", *Weltwirtschaftliches Archiv*, Vol. 30, No. 1, 1929, pp. 1-53.

[45] Lewbel, Arthur, "Using Heteroskedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models", *Journal of Business and Economic Statistics*, Vol. 30, 2012, pp. 67-80.

[46] Li, Nicholas, "Store Variety and Consumer Welfare in Canada-US Retail", Working Paper, 2012.

[47] Lu, Yi, Tao, Zhigang, and Linhui Yu, "The Markup Effect of Agglomeration", Working Paper, 2014.

[48] Manning, Alan, "The Plant Size-Place Effect: Agglomeration and Monopsony in Labour Markets", *Journal of Economic Geography*, Vol. 10, 2010, pp. 717-744.

[49] Melitz, Marc, "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity", *Econometrica*, Vol. 71, 2003, pp. 1695-1725.

[50] Melitz, Marc, and Gianmarco Ottaviano, "Market Size, Trade, and Productivity", *Review of Economic Studies*, Vol. 75, 2008, pp. 295-316.

[51] Peters, Michael, "Heterogeneous Markups and Endogenous Misallocation", Working Paper, 2011.

[52] Krugman, Paul, "Increasing Returns and Economic Geography", *Journal of Political Economy*, Vol. 99, No. 3, 1991 (June), pp. 483-499.

[53] Ottaviano, Gianmarco, Tabuchi, Takatoshi, and Jacques-François Thisse, "Agglomeration and Trade Revisited", *International Economic Review*, Vol. 43, No. 2, 2002 (May), pp. 409-435.

[54] Restuccia, Diego, and Richard Rogerson, "Policy Distortions and Aggregate Productivity with Heterogenous Establishments", *Review of Economic Dynamics*, Vol. 11, No. 4, 2008, pp. 707-720.

[55] Rigobon, Roberto, "Identification Through Heteroskedasticity", *Review of Economics and Statistics*, Vol. 85, No. 4, 2003 (November), pp. 777-792.

[56] Schiff, Nathan, "Cities and Product Variety", Working Paper, 2012.

[57] Syverson, Chad, "Prices, Spatial Competition, and Heterogeneous Producers: An Empirical Test", *Journal of Industrial Economics*, Vol. 55, No. 2, 2007, pp. 197-222.

[58] Zhao, Liqiu, "Markups and Agglomeration: Price Competition Versus Externalities", Working Paper, 2011.

A Appendix A:

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

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

$$Q_{ust} + \sum_{k\epsilon U_{rct}} [P_{kst}\frac{\partial Q_{kst}}{\partial P_{ust}} - \frac{\partial V_{kst}(Q_{kst})}{\partial Q_{kst}}\frac{\partial Q_{kst}}{\partial P_{ust}}] = 0 \tag{34}$$

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

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

Rewrite $\frac{\partial Q_{kst}}{\partial P_{ust}}$ as

$$\frac{\partial Q_{kst}}{\partial P_{ust}} = (\sigma_s - 1)\left(\frac{\partial P_{ct}}{\partial P_{st}}\frac{P_{st}}{P_{ct}}\right)\left(\frac{\partial P_{st}}{\partial P_{gt}}\frac{P_{gt}}{P_{st}}\right)\left(\frac{\partial P_{gt}}{\partial P_{ut}}\frac{P_{ut}}{P_{gt}}\right)\frac{Q_{kst}}{P_{ust}} + (\sigma_g - \sigma_s)\left(\frac{\partial P_{st}}{\partial P_{gt}}\frac{P_{gt}}{P_{st}}\right)\left(\frac{\partial P_{gt}}{\partial P_{ut}}\frac{P_{ut}}{P_{gt}}\right)\frac{Q_{kst}}{P_{ust}}$$
$$+ (\sigma_u - \sigma_g)\left(\frac{\partial P_{gt}}{\partial P_{ut}}\frac{P_{ut}}{P_{gt}}\right)\frac{Q_{kst}}{P_{ust}} - \sigma_u\frac{Q_{kst}}{P_{ust}}1_{\{u=k\}}$$

Use the property of CES that $\frac{dP_{jt}}{dP_{kt}}\frac{P_{kt}}{P_{jt}} = S_{kt}$ to solve for the elasticities to give:

$$\frac{\partial Q_{kst}}{\partial P_{ust}} = (\sigma_s - 1)\,S_{sct}S_{gst}S_{ust}\frac{Q_{kst}}{P_{ust}} + (\sigma_g - \sigma_s)\,S_{gst}S_{ust}\frac{Q_{kst}}{P_{ut}} + (\sigma_u - \sigma_g)\,S_{ust}\frac{Q_{kst}}{P_{ut}} - \sigma_u\frac{Q_{kst}}{P_{ust}}1_{\{u=k\}} \tag{35}$$

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

$$1 + \textstyle\sum_{k\epsilon U_{rct}}(\sigma_s - 1)\,S_{sct}S_{gst}S_{ust}\frac{P_{kst}Q_{kst}}{P_{ust}Q_{ust}} + \sum_{k\epsilon U_{st}}(\sigma_g - \sigma_s)\,S_{gst}S_{ust}\frac{P_{kst}Q_{kst}}{P_{ust}Q_{ust}}$$
$$+ \textstyle\sum_{k\epsilon U_{gst}}(\sigma_u - \sigma_g)\,S_{ust}\frac{P_{kst}Q_{kst}}{P_{ust}Q_{ust}} - \sigma_u - \sum_{k\epsilon U_{rct}}(\sigma_s - 1)\,S_{sct}S_{gst}S_{ust}\frac{\frac{\partial V_{kst}(Q_{kst})}{\partial Q_{kst}}Q_{kst}}{P_{ust}Q_{ust}}$$
$$- \textstyle\sum_{k\epsilon U_{st}}(\sigma_g - \sigma_s)\,S_{gst}S_{ust}\frac{\frac{\partial V_{kst}(Q_{kst})}{\partial Q_{kst}}Q_{kst}}{P_{ust}Q_{ust}} - \sum_{k\epsilon U_{gst}}(\sigma_u - \sigma_g)\,S_{ust}\frac{\frac{\partial V_{kst}(Q_{kst})}{\partial Q_{kst}}Q_{kst}}{P_{ust}Q_{ust}} + \sigma_u\frac{\frac{\partial V_{kst}(Q_{kst})}{\partial Q_{kst}}}{P_{ust}} =$$

0.

Note the different sets over which the summations occur. This 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 $S_{ust}\frac{1}{P_{ust}Q_{ust}} = \frac{1}{\sum_{k\epsilon U_{gst}} P_{kst}Q_{kst}}$ and therefore $\sum_{k\epsilon U_{gst}} S_{ust}\frac{P_{kst}Q_{kst}}{P_{ust}Q_{ust}} = 1$, and analogously for the upper tiers, we can rewrite the previous equation as

$$1 + (\sigma_s - 1)\,S_{rct} + (\sigma_g - \sigma_s) + (\sigma_u - \sigma_g) - \sigma_u - (\sigma_s - 1)\,S_{rct}\frac{\sum_k \frac{\partial V_{kst}(Q_{kst})}{\partial Q_{kst}}Q_{kst}}{\sum_k P_{kst}Q_{kst}}$$
$$- (\sigma_g - \sigma_s)\,\frac{\sum_k \frac{\partial V_{kst}(Q_{kst})}{\partial Q_{kst}}Q_{kst}}{\sum_k P_{kst}Q_{kst}} - (\sigma_u - \sigma_g)\,\frac{\sum_k \frac{\partial V_{kst}(Q_{kst})}{\partial Q_{kst}}Q_{kst}}{\sum_k P_{kst}Q_{kst}} + \sigma_u\frac{1}{\mu_{ust}} = 0.$$

Because we assume that σ_u, σ_g, and σ_s is the same for all u, g, and s within the retail chain, μ_{ust} is the only u-specific term in this expression. Hence, μ_{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 μ_{rct} by

$$1 + (\sigma_s - 1) S_{rct} + \left(\sigma_g - \sigma_s\right) + \left(\sigma_u - \sigma_g\right) - \sigma_u - (\sigma_s - 1) S_{rct} \tfrac{1}{\mu_{rct}}$$
$$- \left(\sigma_g - \sigma_s\right) \tfrac{1}{\mu_{rct}} - \left(\sigma_u - \sigma_g\right) \tfrac{1}{\mu_{rct}} + \sigma_u \tfrac{1}{\mu_{rct}} = 0$$

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

A.2 List of 55 Metropolitan Statistical Areas in 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