Brand Contagion:
The Popularity of New Products in the United States

MARCOS R. FRAZÃO*

This version: November 14, 2020

[Click here for latest version]

Abstract

In this paper, I study the causes of brand sales growth over time and space. I analyze data from a large set of branded retail products sold in different regions in the United States and document a series of stylized facts about their life-cycle. I find that brands typically sell to a small number of locations that tend to be geographically close, and growth usually happens around previously successful markets. Furthermore, I decompose sales into three components: customer base, prices, and quantities per customer. Almost all of the variation in brand sales, both across locations and over time, comes from the first term. The evidence suggests that geography plays a vital role in customer acquisition, but not due to differences in prices. Motivated by these findings, I propose a model in which information about brands' existence spreads geographically, similarly to how contagious diseases spread. Consumers aware of a brand might 'infect' others with that knowledge, and the probability of contagion depends on their location. Additionally, brands have different costs to deliver their goods to different markets. I use the predictions for the correlation of brand sales and customer base across regions to estimate the model using Simulated Methods of Moments and find that the information frictions are more severe between distant locations. Moreover, I use the model to evaluate the welfare gains from reducing information frictions and lowering shipping costs. Finally, I assess the value of brand awareness by computing the expected increase in revenue due to one more informed consumer.

Keywords: Brands, Customer Base, Geography, Contagion, Awareness, Information Frictions

*Department of Economics, Yale University. 28 Hillhouse Avenue, New Haven, CT 06511. Email: marcos.frazao@yale.edu. Website: https://https://sites.google.com/view/marcos-frazao Researcher(s) own analyses calculated (or derived) based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business.The conclusions drawn from the Nielsen data are those of the researcher(s) and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

I. INTRODUCTION

Spatial economics and international trade have blossomed in recent decades by exploring the economic implications of geography. Traditionally, these frameworks consider frictions that make it costly to move goods and people. Recent evidence, however, suggests that distance might also impact economic outcomes by making it harder for information to flow between places. I investigate this hypothesis using data on branded retail products sold in different regions in the United States. Brands allow consumers to identify products that they know. Therefore, analyzing brand data is a natural starting point to evaluate the role of information frictions between producers and final consumers, especially product awareness.

I use the Nielsen Homescan panel data to construct a panel of over 150,000 brands in 44 major regions in the US over a decade. The data uncovers new stylized facts about the growth of brands in the US and its key factors. I show that sales are usually concentrated in a small number of locations that tend to be geographically close. Older brands serve more markets and have higher sales, especially in areas close to their previous top markets.

To explore what drives these patterns, I break sales down into the customer base and sales per customer. Differences in the number of customers explain almost all of the variation in brand sales. Sales per customer are stable for a given brand, both in the cross-section and time-series dimension. Furthermore, I use data on bar-coded products to break sales per customer into price levels and quantities per customer. Both components do not change systematically over space, and as the brand gets older.

These results suggest that trade costs that affect prices cannot explain the dynamic of brand sales - and customer base - in the US. Hence, I take an alternative approach and consider information frictions in the form of product awareness as the main driver for customer acquisition. I propose a parsimonious model in which consumers aware of a brand might'infect' others with this knowledge. Geography is critical, as some regions are more connected than others, which makes contagion more likely. The model generates customer base dynamics similar to the spread of a contagious disease, with the advantage of providing simple characterization for the distribution of the number of customers across brands.

I estimate the model for the year 2016, using a Simulated Method of moments. It can replicate the previously mentioned stylized facts about brand dynamics. Furthermore, I use it to conduct counterfactuals regarding the welfare cost of information frictions and the expected revenue associated with product awareness.

The prominent role of informational frictions found here is in line with the recent literature on international trade. This suggests that the same forces that are at play internationally are also present domestically.

This paper also contributes to the long literature of product and technology adoption following the initial works of Rogers (1962) and Bass (1969) in which consumers' adoption of new products rely on their interactions with other consumers, generating S-shaped adoption curves. The works that followed focused on two aspects of product adoption: the behavioral reasons behind consumer decisions of adopting new products and the mathematical characterization and estimation of these processes. Recent developments in the literature also use data on the particular social network of consumers and find evidence of the importance of their connections in the adoption of certain products.[1]

These advances are important for understanding how particular businesses grow, but there are a couple of considerations to be made when evaluating the aggregate implications of product awareness and contagion. First, we want to consider a broad set of products instead of focusing only on a few, since we want our results to be representative of a large part of the economy. Our choice of data allows us to do so, as it covers most items that final consumers buy. The other concern has to do with the economic environment. Since we want to evaluate counterfactuals we need our model to describe how brands compete in different locations in equilibrium. For this reason, we develop a general equilibrium model, where many brands exist but consumers are not aware of all of them. Brands might have different costs to serve different locations, and for a given product the consumer will choose the least expensive one that they are aware of.

In economics, there is a vast literature that assesses the aggregate effects of product creation and adoption. Romer (1987) provided the initial framework that links product creation and economic growth. Much has been done in this field and, more recently, Perla (2019) evaluates the effects of incorporating a slow diffusion process for new products that can explain patterns observed in the life-cycle of industries.

The fact that firms spend more than $200 billion annually in advertisements only in the US[2] suggests that reaching potential customers is very valuable. By recognizing this, Gourio and Rudanko (2014)

[1]Hauser et al. (2006) provide a literature review on product innovation and adoption in Marketing Science, including a discussion on the current topics that are being researched in the field.

[2]Industry figures from GroupM/Statista.

introduce search frictions for firms to reach consumers, which explains long-term customer relationships and can rationalize patterns in investment volatility by firms. In the context of International Trade Arkolakis (2010) also relies on modeling reaching consumers as a costly activity, which allows him to reconcile the positive correlation between firm entry and market size and the existence of small exporters in different countries.

Campa and Goldberg (2005) document substantial differences between the short-run and long-run transmission of exchange rates to import prices. The information frictions associated with reaching consumers can also explain some important puzzles that are associated with the timing in which economic events take place. By including them in an International Macroeconomics model, Drozd and Nosal (2012) can generate persistent pricing-to-market and quantitatively account for several puzzles in the literature.

By combining both physical trade frictions and information frictions in the form of random search between potential buyers and suppliers, Eaton et al. (2019) can generate the patterns for firm-to-firm trade similar to the ones observed for French manufacturers and their European counterparts, and show how each friction affects their predictions. Using similar search frictions, Lenoir et al. (2018) show that reducing the information frictions associated with reaching buyers abroad can increase the efficiency of exporters in equilibrium.

This paper pushes the understanding of how information frictions in the form of imperfect product awareness by final consumers can explain the dynamics of brands and economic aggregates over a geographic space. In the next section, I describe the data and proceed to introduce the model.

II. Data and Preliminary Evidence

In this work, I use the Nielsen Homescan panel data between 2007 and 2016. It tracks the purchases of about 50,000 panelists per year in the US, with detailed information about households and products. The households' information we are interested in is where they live, what products they buy in a given year, and how much they spend on each product. The geographic unit that is used throughout the paper is the Scantrack® Markets, defined by Nielsen. There are 52 of these regions that cover all major metropolitan areas in the US. During the period evaluated, only the weights of 44 of these regions are designed to represent the whole region's demographics. I restrict my attention to these locations, which account for

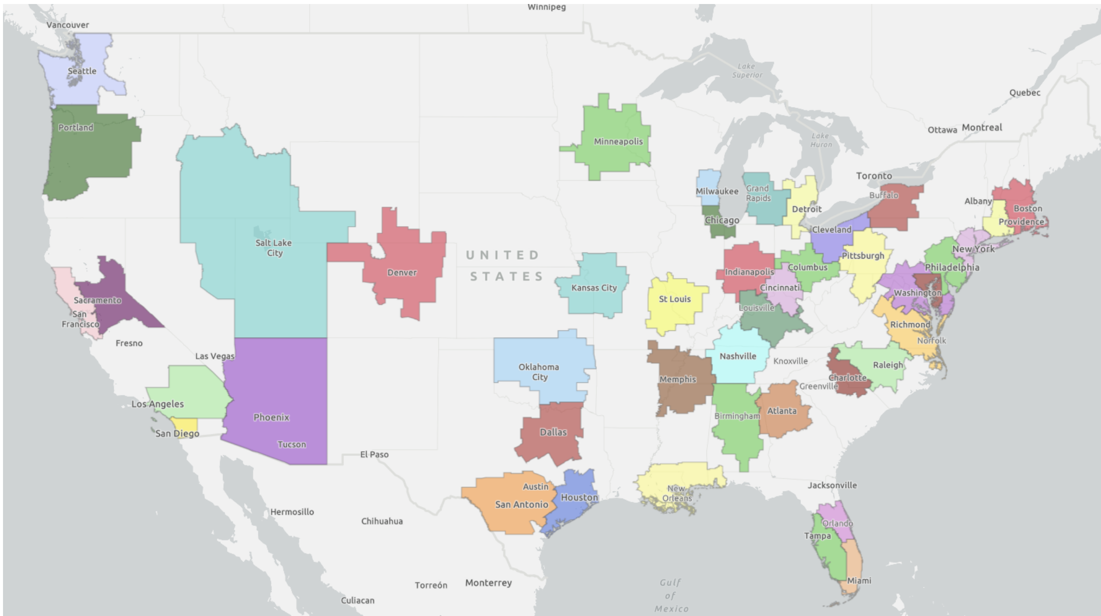


Figure 1: *44 representative Scantrack® Markets.*

roughly 70% of the US population.

Product adoption is typically considered a process that involves several steps. It might include gathering information and evaluating the product quality, but it invariably starts with product awareness, *i.e.* knowledge about a product's existence. Here, as in Kalish (1985), I reduce this process to two stages: awareness and adoption. In our simplified setting, consumers might gain knowledge about a product by interacting with others. Still, they only decide to buy that particular product if it is the least expensive known product within its category.

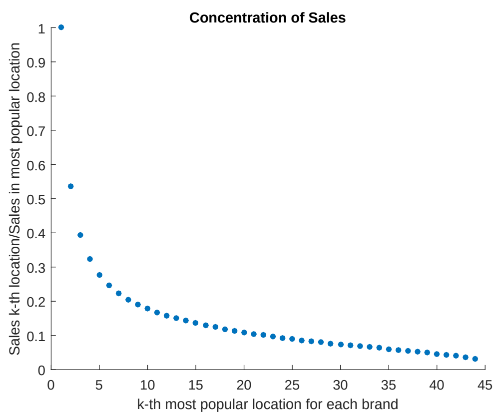
The Nielsen Homescan data presents two possibilities for defining a product: its UPC bar code and its assigned brand code. According to the American Marketing Association dictionary: "A brand is a name, term, design, symbol or any other feature that identifies one seller's good or service as distinct from those of other sellers.". Since I am interested in the effects of product awareness, I naturally focus on the brand code as the product definition, as it presents the necessary aspects for consumers to identify and recall the product. Another reason to avoid using the UPC for this purpose is that many similar products might have different UPCs. For example, a 12 fl oz bottle of soda has a different UPC than the same soda in a 24 fl oz bottle.

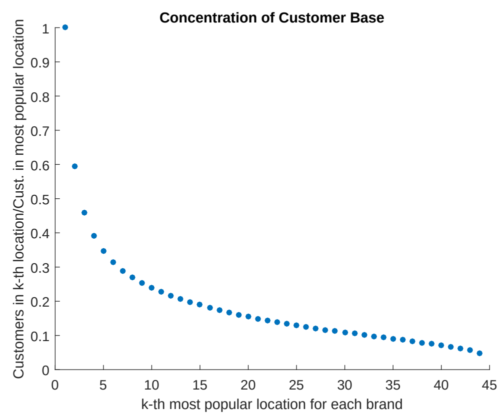
I define the customer base of a brand as the number of households that bought the product at least once during a year. For each year, I compute the brands' customer base in each location and their total sales. The following are summary statistics for all brands and locations in 2016.

Table 1: *Summary Statistics for Brands in 2016*

# of Brands	Avg Number of Locations Served	Avg Total Sales	Avg Customer Base
73,118	11.48	$ 5,372,218	331,077

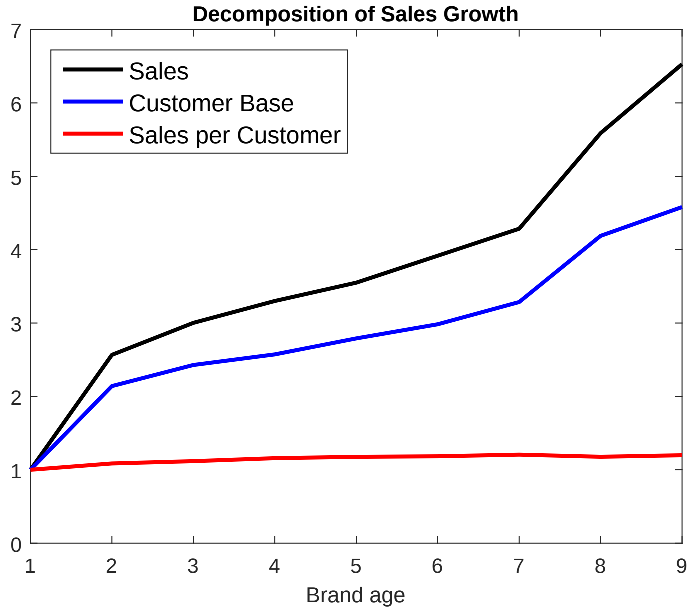
We can see that most brands are present in a few locations, as the average number of locations served is 11. It is also the case that their sales and number of customers are fairly concentrated. To evaluate that, I first compute the sales and number of customers in each brands' top market. Then, I calculate their sales and customers' share in other locations relative to their top market. Finally, I take the average of these measures over all brands, for their first, second, third locations and so on. Zeros are not accounted for in this calculation. If a brand sells to 7 locations during that year, the 0 in their 8th market is not included in the average. The next figure displays the results.





The figures suggest that as we move away from each brand's top markets, their number of customers and sales fall sharply and by comparable rates. One possible reason for this is that the same product might have higher prices in different markets, which would reduce their number of customers and possibly their sales.

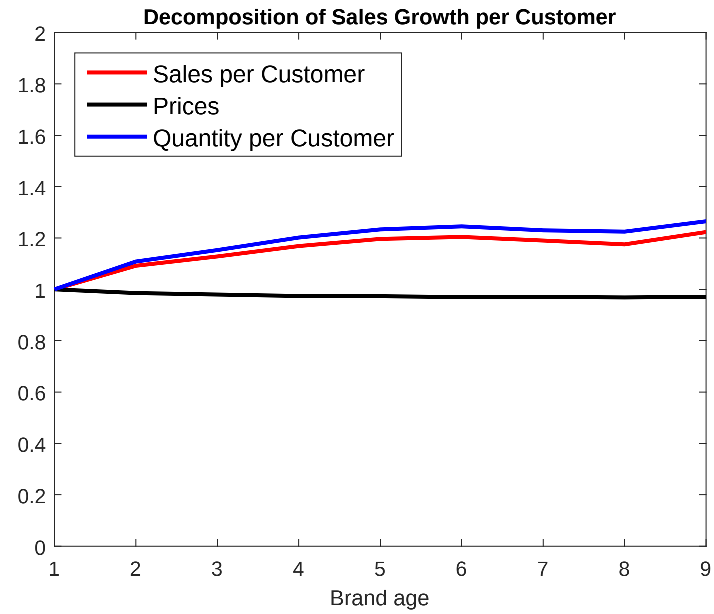
I can also observe the behavior of sales and customer base as brands get older. For the ones that entered after 2007, we can use the first year that they appear in the data to compute their age. I construct an index for their sales, customer base, and sales per customer by normalizing the variables by their initial values. After that, I take the average over all brands in each age group, weighted by their initial sales. In



the next figure, we can see that older brands sell significantly more and to more people than they used to. However, the change in sales per customer is more modest, with an increase of just 20% after eight years. The next step is to break sales per customer down into two components: prices and quantities per customer.

To have a good measure of prices by brand, I construct a price index for each brand in each location that they sell to during a particular year, as well as a nation-wide price index for the brand. I start by constructing the average price for each UPC by computing their sales divided by quantity in each location. After that, I divide their average prices by their nation-wide average price in the first year that they appear in the data set. The brand-level price index in a location is then constructed by averaging all the brand's UPC indices sold there, weighted by their particular sales. However, some corrections need to be made, especially for the inclusion of new UPCs in the data. If a new UPC enters the data and there were already UPCs being sold beforehand, the new UPC's initial index is re-scaled using the price index for that brand in the entry period, excluding the entrant UPC. If no other UPC was being sold at that time, the entrant UPC's initial price index is re-scaled based on the most recent price index of the brand.

To highlight these corrections' importance, suppose that a brand sells a 12 oz can of soda, and its price



doubles every year. By the 3rd year, the price index for the can would be 4. If in that year they launch a 24 oz bottle of the exact same soda with the same price-per-ounce as the can, the UPC price index of the bottle would be 1, and its inclusion would bias the brand-level price-index down. The correction makes the bottle's initial price index to be 4 and avoids this type of bias. Unfortunately, the correction cannot account for potential changes in the quality of new UPCs or unit-price changes associated with different packages.

After computing prices, I construct a quantity per customer index by dividing the brand sales by their customer base times the price index. Again, these measures are normalized by their initial values to see how age affects them. The next figure shows how these components affect sales per consumer. We can see that, on average, prices do not change much as brands get older. The small movement we have observed in sales per customer is accounted for by changes in quantities purchased. But again, these variations are dwarfed by the magnitude of the shift in the customer base.

In order to evaluate how the brands' sales and customer bases evolve over the US geography, it is helpful to restrict our attention to a subset of the data. Ideally, one would take brands from a particular location and track their sales and customers in different places over time. However, in our case, this is not feasible. The UPCs that appear in the data set are re-coded versions of the products real UPCs, so we cannot match them directly with firm data and find what their origin would be. Even if that would be possible, the headquarters of the firm might be different from where production happens. Also, some of these firms have several plants, so they can engage in production in different places across the country. These features make the sole definition of an origin questionable for some brands. To circumvent this, I assign origin by looking at the first location that sells a product from a particular brand. Even though this is aligned with most economics models in which firms' products usually start by selling locally, it is still subject to measurement error due to the sample nature of the data set, and the previously mentioned issues.

With this in mind, I have selected the 168 brands that entered the data in 2008, and the first sale was in Cleveland. The reason for the location choice is that Cleveland is relatively central, surrounded by other representative markets and it is not small. In the next table, we can see the decline in the number of brands that are operating at a given year, and also that their average number of UPCs follows a similar hump-shaped pattern as the brand sales and customers.

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	All
Number of Brands	168	100	82	70	62	59	61	53	45	168
UPCs/Brand	1.44	2.18	3.22	4.00	3.92	3.56	3.26	3.13	2.86	4.07

Figures 6 to 13 in the Appendix show color-coded maps of the US displaying the total sales and customer base of these selected brands in each location for their first 4 years (2008, 2009, 2010, and 2011). There are a few things to note. First, the patterns observed in sales are very similar to the ones observed in customers. Sales are smaller at first and grow over time before they fall. In the beginning, most sales happen in Cleveland and surrounding areas, as well as the two largest markets New York and Los Angeles. As time passes, not only do sales become more spread over other regions, but it seems like areas that are close to those two large markets also experience an unusually large increase in sales and customer acquisition.

To quantify the effects of geography in sales, I construct a measure of similarity between brands' sales in any two places. I start by computing the normalized sales of a brand in each location by dividing the total amount that a brand sells there in a given period by their average sales across all locations that they serve. This procedure generates a vector for each brand that indicates if the amount they sell in a location is above or below the brand's national average, making the observations for brands comparable. Next, I calculate the correlation between the normalized sales of brands in those two places for each pair of locations. If this correlation is high for pairs that are close to each other and low for pairs that are further away, that means that places that a brand has high sales tend to be closer to other places that sales are high, and conversely that places with low sales are also close to other low sales places. This is precisely what we observe by plotting these pairwise correlations and the distance between two locations, as in figure 2.

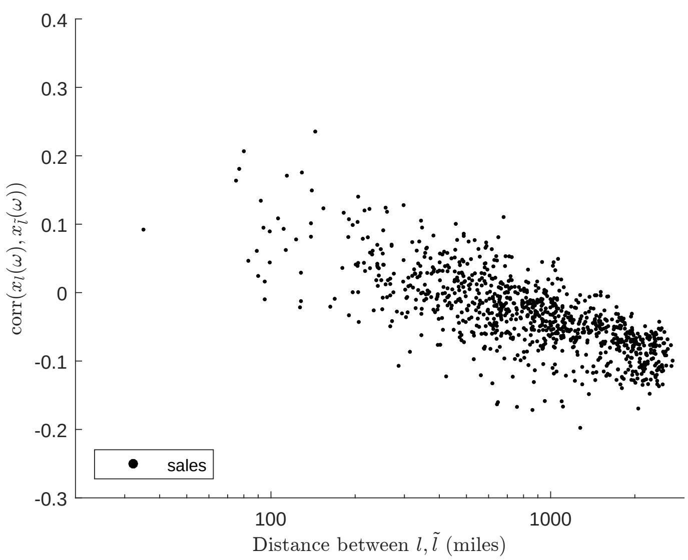


Figure 2: *Correlation of brands relative sales in two locations*

I do the same procedure for all the components of brand sales that we observe: customer base, price index, and the remainder, representing an index for the average quantity individual households buy in a given location. Figure 3 displays the results. We can see that the customer base behaves in the same way as sales. Interestingly, prices and individual quantities bought do not share the same patterns. This suggests that, in our analysis's scope, the geographic pattern that we observe for sales must be explained by something that affects customer acquisition in places that are close to each other, without relying on differences in prices.

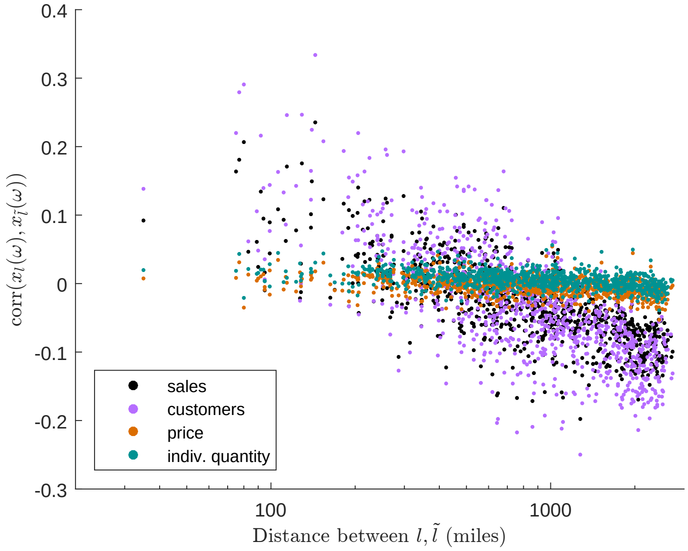


Figure 3: *Correlation of brands relative sales in two locations*

After observing the geographic concentration of sales and customer base, I can compute similar correlations between pairs of locations in different periods. The lagged correlations measure the similarity of the variables in one region today and another tomorrow. The figure suggests that sales and customer base also display geographic persistence, in the sense that a higher level of sales in one location today is associated with higher sales in close areas tomorrow.

One potential explanation for that is what I call contagion, in which consumers that are aware of a brand can inform unaware consumers that are nearby. This mechanism is similar to the one in which Chaney 2014 uses to model French exporter's entry into markets. However, the decisions that the consumer face in that setting are very simplistic, in the sense that if a consumer is aware of a good they demand one unit of it. Here, I provide a framework in the spirit of Eaton et al. (2019), in which information frictions prevent buyers from having access to all sellers, and they choose to purchase the least expensive variety that they are aware of. The model generates predictions about the distribution of brands' customers in different locations that can be brought to the data, and there is effective competition in the sense that not only a brand must be known, but it must also be cheaper than its competitors so that it makes a sale. This competitive setting might, therefore, be more suited for counterfactuals and welfare analysis.

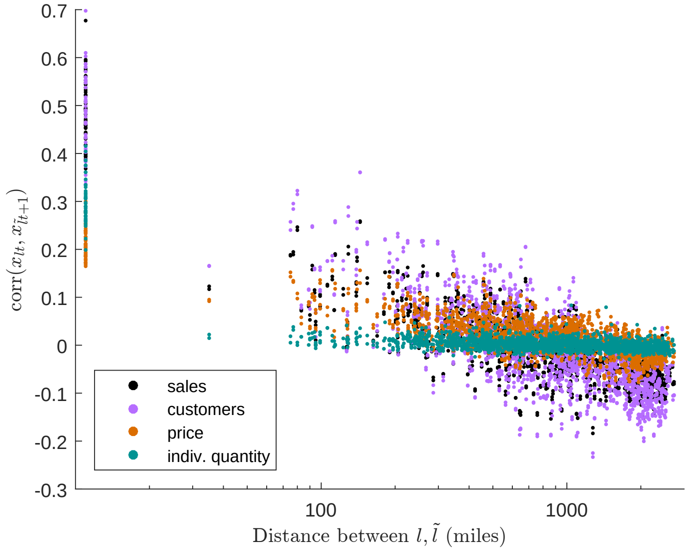


Figure 4: *Correlation of brands relative sales one period ahead*

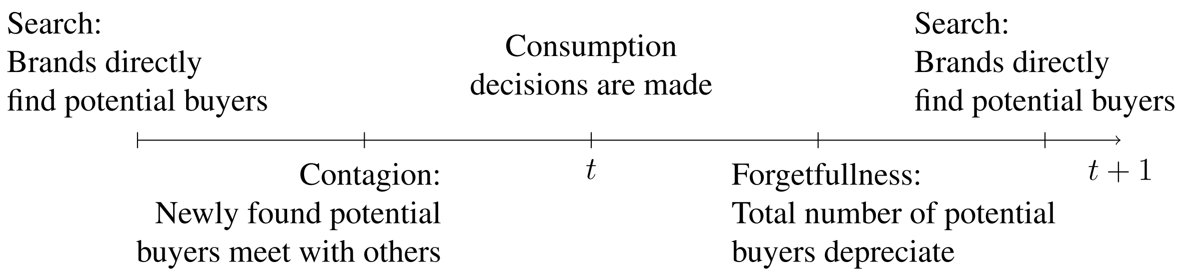
III. Model

The model describes an economy that is composed of discrete number of locations populated by consumers. They buy different varieties of products, and many brands can produce each variety. The key feature of the model is the description of a brand's customer base dynamics. Consumers buy a product if they are aware of the brand and if it is the cheapest among the known brands for that variety.

The evolution of the number of consumers aware of a brand - the *potential* customer base - is governed by a stochastic process. This process describes the brand actively searching for consumers in each region and the possibility of consumers spreading this information to others. If this contagion is more likely among consumers close to each other, the model predicts the patterns of geographic concentration and persistence observed in figures 3 and 4.

The nature of the stochastic process comes from the literature in epidemiology. The particular modeling choices made here generate a closed-form solution for the distribution of brands' potential customers. This way, it is easy to compute the probability that the product is the cheapest one found by the consumer. I start by describing the evolution of the potential customer base. Then, I move to production technology, consumers' preferences, and equilibrium outcomes.

EVOLUTION OF BRAND AWARENESS



When a brand is created, it searches *directly* for consumers in all markets. The number of consumers it reaches this way in location l follows a Poisson distribution, $N_{0,l}^D \sim \text{Poisson}(\lambda_l)$, independent across all regions.

After that, the consumers who have been contacted can inform others who were previously unaware and spread the information about the brand's existence, which is akin to how contagious diseases spread. Let $N_{0,l}$ denote the number of consumers that the brand has matched in its first period in location i. This random variable is

$$N_{0,l} = N_{0,l}^D + \sum_{i=1}^{N_{0,l}^D} n_{0,l}(i) + \sum_{m \neq l} \sum_{i=1}^{N_{0,m}^D} n_{0,ml}(i).$$

The $n(i)$ random variables denote the number of potential consumers that each aware consumer can "infect" locally and in other regions. A high draw for $N_{0,m}^D$ also impacts the expected number of consumers in market l due to contagion. Note that $N_{0,l}$ is a compound random variable since the number of elements in the summations is random.

My goal is to have a simple characterization for the potential customer base of brands with different ages, and dealing with compound random variables can be challenging. For this reason, I use the Generalized Poisson distribution, henceforth GP, which has properties that make the compounding due to contagion straightforward. This distribution was introduced by Consul and Jain (1973) as an extension of the Poisson distribution. It includes an extra parameter that allows for the variance to exceed the mean. If $X \sim GP(\lambda, \phi)$, then $\mathbb{E}(X) = \frac{\lambda}{1-\phi}$ and $\mathbb{V}(X) = \frac{\lambda}{(1-\phi)^3}$. The Poisson distribution is the special case in which $\phi = 0$. Notice that increasing ϕ above 0 increases the mean, but increases the variance by even more. Shoukri and Consul (1987) describe the use of the Generalized Poisson to characterize the total number of people infected by a contagious disease in a setting that is reasonably similar to our framework here. The properties that simplify the characterization of the potential number of buyers are the following[3]

Property 1. If $X \sim GP(\lambda_X, \phi), Y \sim GP(\lambda_Y, \phi)$ are independent, then $X + Y = Z \sim GP(\lambda_X + \lambda_Y, \phi)$.

Property 2. If $X \sim GP(\lambda_X, \phi_X)$ and $\{Y_i\} \sim GP(\phi_Y, \phi_Y)$ is a sequence of *iid* random variables that are also independent from X, then $X + \sum_{i=1}^{X} Y_i \sim GP(\lambda_X, \phi_X + \phi_Y)$.

Property 1 is similar to the convolution property of the Poisson distribution, and it allows to add up independent draws of GP. Property 2 allows for compounding, and it is extremely helpful in the context of contagion. Notice that the number of Y_i random variables in the summation is X, which is a random variable itself.

I assume that the number of consumers that each aware consumer is able to reach in their own location is $n_{0,l}(i) \sim GP(\phi, \phi)$, with $\phi \in [0, 1)$. Recall that $N^D_{0,l}$ is simply Poisson distributed, which means that $N^D_{0,l} \sim GP(\lambda_l, 0)$. The second property implies that $N^D_{0,l} + \sum_{i=1}^{N^D_{0,l}} n_{0,l}(i) \sim GP(\lambda_l, \phi)$. Consumers also contribute to the spread of awareness to other locations. I assume that the total number of consumers that each aware consumer in m finds in location l, is $\sum_{i=1}^{N^D_{0,m}} n_{0,ml}(i) \sim GP(\lambda_m \lambda_{ml}, \phi)$. Noting that the total number of consumers that became aware with information coming from different locations is independent, we can use the first property of the GP to write that

$$N_{0,l} \sim GP\left(\lambda_l + \sum_{m \neq l} \lambda_m \lambda_{ml}, \phi\right).$$

After contagion happens, the brand sells its product to consumers. The average number of potential consumers that a newly created brand has in location l is $\mathbb{E}(N_{0,l}) = [\lambda_l + \sum_{m \neq l} \lambda_m \lambda_{ml}](1 - \phi)^{-1}$. We

[3]The first property is derived in Consul (1989). The proof of the second property consists of an induction argument on the p.m.f of the resulting distribution to show that it is equal to the p.m.f. of $GP(\lambda_X, \phi_X + \phi_Y)$.

can see the influence of the other locations in the expansion of the customers in l, through the contagion parameters λ_{ml}. This is also evident when we compute the covariance of the number of aware consumers in two locations $\text{Cov}(N_{0,l}, N_{0,m}) = [\lambda_l \lambda_{lm} + \lambda_m \lambda_{ml}](1-\phi)^{-2}$. The higher λ_{ml} and λ_{lm} the higher the correlation between location l, m.

As we have seen in the data, the correlation between customers in close regions is higher. This suggests that λ_{lm} is higher if locations l and m are geographically close. Since I want to study the effects of distance on the information frictions, I define $\lambda_{lm} = C_0 \exp(-C_1 \text{distance}_{l,m})$. This choice allows having a simple description of how geography affects the contagion friction and greatly reduces the number of contagion parameters to be estimated.

Before the next period, some of the consumers aware of the product forget about its existence. I consider a survival process similar to a binomial survival process for a random variable with a Poisson distribution. Say that $R \sim \text{Poisson}(\lambda)$ describes the number of people that know about a brand. If each of these individuals is independent and equally likely to forget about the brand existence, and the probability of it happening at the individual level is δ_b, then the distribution of the individuals that remember the brand conditional on X is $R|X \sim \text{Bin}(X; (1-\delta_b))$. The unconditional distribution of the consumers that remember is $R \sim \text{Poisson}((1-\delta_b)\lambda)$. Unfortunately, this survival process does not preserve the form of General Poisson distribution. However, the Quasi-Binomial distribution of type-II introduced by Consul and Mittal (1975) does. If $X \sim GP(\lambda, \phi)$ and $R|X \sim QBDII(X; (1-\delta_b), \phi/\lambda)$, then $R \sim GP((1-\delta_b)\lambda, \phi)$. Since I use this operation quite frequently, I define the following operator

$$\text{If } X \sim GP(\lambda, \phi), \text{ then } \chi(X)|X := QBDII(X; (1-\delta_b), \phi/\lambda).$$

Evaluating $\chi(X)$ before knowing the value of X implies that $\chi(X) \sim GP((1-\delta_b)\lambda, \phi)$. Therefore, the total number of potential customers that did not forget about the brand is

$$\chi(N_{0,l}) \sim GP\left((1-\delta_b)\left[\lambda_l + \sum_{m \neq l} \lambda_m \lambda_{ml}\right], \phi\right).$$

After that, the brand directly searches for consumers as before. The number of consumers that they find is independently distributed $N^D_{1,l} \sim \text{Poisson}(\lambda_l)$, and contagion happens as in the previous period. Since the number of newly found potential consumers is independent from the ones that remember the brand, we can write

$$
\begin{aligned}
N_{1,l} &= \chi(N_{0,l}) + N^D_{1,l} + \sum_{i=1}^{N^D_{1,l}} n_{1,l}(i) + \sum_{m \neq l} \sum_{i=1}^{N^D_{1,m}} n_{1,ml}(i) \\
N_{1,l} &\sim GP\left(\left[\lambda_l + \sum_{m \neq l} \lambda_m \lambda_{ml}\right] + (1-\delta_b)\left[\lambda_l + \sum_{m \neq l} \lambda_m \lambda_{ml}\right], \phi\right).
\end{aligned}
$$

In general, for a brand with age a, we have that

$$
N_{a,l} \sim GP\left(\sum_{i=0}^{a} (1-\delta_b)^i \left[\lambda_l + \sum_{m \neq l} \lambda_m \lambda_{ml}\right], \phi\right) := GP(\Lambda_{l,a}, \phi).
$$

where $\Lambda_{l,a} := \sum_{i=0}^{a} (1-\delta_b)^i \left[\lambda_l + \sum_{m \neq l} \lambda_m \lambda_{ml}\right]$. Given the properties of the GP, the mean and variance of the potential customer base is

$$
\mathbb{E}(N_{la}) = \frac{\Lambda_{la}}{1-\phi}, \quad \mathbb{V}(N_{la}) = \frac{\Lambda_{la}}{(1-\phi)^3}.
$$

The average number of potential customers grow with age, and so does the variance. The assumption that there is a positive probability that consumers forget about the brand guarantee that the process is stationary.

This process is related to epidemiologic metapopulation models[4]. There, infected people in one location might spread the disease to others in different places. The main difference here is that consumers are only contagious when the brand directly found them. This assumption limits the intertemporal effects that contagion might have on the evolution of the customer base. However, it allows for a closed-form

[4]See Sattenspiel (2009) for examples.

characterization of the distribution of all brands' potential customer base given their age. This helps keep track of how many brands each *consumer* is aware of, which is a key market condition that governs the probability that a brand actually sells its product.

To evaluate the *effective* number of customers for each brand, we move to the description of the costs that the brand faces and how consumers choose which products they buy.

TECHNOLOGY

The data doesn't assign an origin to a brand. Consequently, it is harder to consider the effects of geography on the costs of delivering goods to different locations, as it is not possible to model the production cost plus the shipping costs explicitly. However, the model circumvents this problem by considering that brands are assigned a vector of productivities to deliver their good in different locations. This setting allows for the interpretation that regions with low costs are close to the production site, and the ones with high costs are hard for the brand to reach. This interpretation is also consistent with several plants producing a single brand's goods, which is the case for many well established retail products. The data also suggests that differences in brand prices across regions are not essential to describing their sales' geographic patterns. Therefore, I choose to simplify the cost structure of brands and allow for a richer characterization of awareness evolution.

A brand can deliver the good that they produce in every location $l = 1, \ldots, \mathcal{L}$, but at different costs. Each brand has a linear productivity in each location z_l. The measure of brands of all ages that have their productivity vector greater or equal to $\mathbf{z}$ is[5]

$$M(\mathbf{z}) = \left(\sum_{l=1}^{\mathcal{L}} \frac{z_l}{T_l^{1/\theta}} \right)^{-\theta}.$$

This implies that for $z_l > 0$, the measure of brands of all cohorts that have productivity higher than z_l is $M_l(z_l) = T_l z_l^{-\theta}$, and the correlation between the productivities of a brand in any two locations is $1/\theta$. The

[5]One way to interpret the description of the measure of brands is to consider that a measure ϵ^{-1} of brands draw $\mathbf{z}$ from a multivariate Pareto distribution as in Mardia (1962) with CDF $1 - \left(\sum_{l=1}^{\mathcal{L}} \frac{z_l}{(\epsilon T_l)^{1/\theta}} - \mathcal{L} + 1 \right)^{-\theta}$, and to take the limiting case as $\epsilon \to 0$.

measure of brands that have delivery cost below c is then $\mu_l(c) = T_l c^\theta$. After every period a fraction of δ_e of brands of all ages cease to exist and are replaced by entrant brands. This implies that the measure of brands with age a that have unit costs below c in location l is $\mu_l^a(c) = \delta_e(1-\delta_e)^a \mu_l(c) = \delta_e(1-\delta_e)^a T_l c^\theta$.

The parameters T_l govern how costly it is, on average, to deliver a good to a particular location. In the data, regions might have different price distributions for various reasons, such as rents or distance from where production happens. In the model, T_l can capture these cost differences. In equilibrium, however, the price index of a location also depends on how well information about brands circulates.

CONSUMERS

All consumers have the same utility function. They value consumption in a single period according to a CES aggregator over the quantity that they consume of each variety j.

$$U(\mathbf{q}) = \sum_{t=0}^{\infty} \beta^t \left[\int_0^1 q_t(j)^{\frac{\sigma-1}{\sigma}} dj \right]^{\frac{\sigma}{\sigma-1}}.$$

There is perfect substitution within a given variety j. Consumers in l are endowed with $\bar{X}_l/L_l$ units of cash every period. Their budget constraint is $\int_0^1 p_t(j)q_t(j)dj = \bar{X}_l/L_l$, where j indexes the varieties. They are aware of many brands that produce each variety. They are also indifferent about which brand produces the good. Therefore, their decision process consists of looking at all the costs from brands they know as quotes and buying from the cheapest one. Now, I describe the distribution of the number of brands consumers know.

Consumers are equally likely to be hit by any type of shock that either makes them aware or unaware of a brand. Therefore, the intensity that consumers in l are matched with a brand with age a is simply the average number of consumers reached by brands of that cohort, divided by the local population: $\frac{\Lambda_{l,a}}{L_l(1-\phi)}$. Hence, the distribution of the number of quotes that consumers have with cost below c is Poisson distributed, with the following parameter

$$\begin{aligned}\rho_l(c) &= \sum_{a=0}^{\infty}\int_0^c \frac{\Lambda_{l,a}}{L_l(1-\phi)}d\mu_l^a(c) = \frac{\delta_e T_l}{L_l}\left(\sum_{a=0}^{\infty}\frac{(1-\delta_e)^a\Lambda_{l,a}}{1-\phi}\right)c^\theta \\ &= \nu_l c^\theta.\end{aligned}$$

This implies that the probability that a buyer encounters no quotes below c for a given variety is $\exp(-\rho_l(c))$. The effective price paid by the consumer is the lowest quote they can find. The price distribution in l is given by integrating over all varieties in the unit interval [0,1]:

$$G_l(p) = 1 - \exp\left(-\nu_l p^\theta\right).$$

This is the distribution of quotes from brands of all ages. To study brand sales evolution, I derive the quote distribution of brands with a particular age from evaluating the probability that a consumer buys from a brand from that cohort. The distribution of quotes that a buyer in l receives from brands with age a also follows a Poisson distribution, this time with parameter

$$\rho_l^a(c) = \int_0^c \lambda_l^a d\mu_l^a(c) = \frac{\delta_e T_l}{L_l}\left(\frac{(1-\delta_e)^a\Lambda_{l,a}}{1-\phi}\right)c^\theta := \nu_{l,a}c^\theta.$$

Hence, the distribution of the lowest quotes coming from brands with age a in location l is

$$G_{l,a}(c) = 1 - \exp\left(-\nu_{la}c^\theta\right).$$

The probability that a variety is bought from a brand with age a can be found by solving the following integral $\pi_{la} = \int_0^\infty \Pi_{a'\neq a}[1 - G_{la'}(c)]dG_{la}(c)$, which implies that

$$\pi_{la} = \frac{\nu_{la}}{\nu_l} = \frac{(1-\delta_e)^a\Lambda_{l,a}}{\sum_{a'=0}^{\infty}(1-\delta_e)^{a'}\Lambda_{l,a'}}.$$

This equation shows two forces at play that affect the probability that consumers buy from a given cohort. On the one hand, as brands get older, they become more well-known, as shown by the Λ_{la} term that increases with age. However, as time passes, a brand is also more likely to exit, as we can see on the $(1-\delta_e)^a$ term. Those two forces generate a hump-shaped curve for the total customer base and sales of particular cohorts, similar to what we observe in the data.

PRICE INDEX AND EQUILIBRIUM

To characterize the equilibrium, I start by constructing the price index, which imposes a necessary parameter restriction for its existence. Second, I describe each brand's effective customer base, which is needed to compute its sales, and discuss the market clearing conditions in detail.

For the CES utility, the price index is given by $P_l = \left[\int_0^\infty p^{1-\sigma} dG_l(p)\right]^{\frac{1}{1-\sigma}}$, where p is the effective price paid by consumers. As mentioned above, the price paid is the lowest quote that consumers can find, which implies that $G_l(p) = 1 - \exp(-\nu_l p^\theta)$. Solving the integral, we find that

$$P_l = \nu_l^{-\frac{1}{\theta}} \left[\Gamma\left(1 - \frac{\sigma - 1}{\theta}\right)\right]^{\frac{1}{1-\sigma}}$$

This equation implies that we need to impose the parameter restriction that $\theta > \sigma - 1$, for the equilibrium to be well-defined.

The CES preferences imply that a consumer spends $c_l^{1-\sigma} P_l^{\sigma-1} \frac{\bar{X}_l}{L_l}$ in a variety with cost c_l. To find the total sales of a particular brand, we need to characterize the total number of consumers that effectively buy their product. Let $B_{la}(c_l)$ denote the random variable that describes the customer base of a brand with age a and that has cost c_l. As previously discussed, age matters for the distribution of their potential customer base N_{la}, but each potential consumer has a probability $\exp(-\nu_l c_l^\theta)$ of actually buying the product. This implies that, given a realization of N_{la}, the number of buyers follows a binomial distribution with parameters N_{la} and $\exp(-\nu_l c_l^\theta)$. We can use the total law of expectations to show that

$$\mathbb{E}[B_{la}(c_l)] = \mathbb{E}[\mathbb{E}[B_{la}(c_l)|N_{la}]] = \mathbb{E}[\exp(-\nu_l c^\theta) N_{la}] = \frac{\exp(-\nu_l c_l^\theta)\Lambda_{la}}{1-\phi}.$$

We can see that on average older and efficient brands have more customers. However, randomness still plays an essential role in the outcomes of specific brands. First, even a brand with high productivity might be unlucky to match consumers who have found cheaper alternatives. As time passes, this effect is mitigated by the expected increase in their potential customer base. But the trajectory of consumers who are aware of the brand is also random, and some brands will be more successful in finding consumers. In particular, if a brand directly finds consumers in one location, they might 'infect' others randomly. If the contagion parameters are larger for close regions, this luck can also spread to neighboring places. This process generates the pattern we observe in the data: the customer base of a brand is geographically concentrated and persistent.

In the data, we observe the same geographic pattern for sales. We can see this in the model by writing the brand sales as $x_{la}(c_l) = c_l^{1-\sigma} P_l^{\sigma-1} \frac{\bar{X}_l}{L_l} B_{la}(c_l)$. Therefore, the expected sales of a brand is

$$\mathbb{E}[x_{la}(c_l)] = c_l^{1-\sigma} P_l^{\sigma-1} \frac{\bar{X}_l}{L_l} \frac{\exp(-\nu_l c_l^{\theta}) \Lambda_{la}}{1 - \Phi_{la}}.$$

In equilibrium, the fact that the demand of each brand is met by supply also implies that

$$\sum_{a=0}^{\infty} \int_0^{\infty} \mathbb{E}[x_{la}(c)] d\mu_{al}(c) = \bar{X}_l.$$

Other Model Predictions

The model generates other predictions that can be brought to the data. Going back to the definition of the effective customer base of a brand $B_{la}(c)$, we know that $B_{la}(c)|N_{la} \sim Bin(N_{la}, \exp(-\nu_l c^{\theta}))$. This implies that a brand that has N_{la} potential customers has a probability $(1 - \exp(-\nu_l c^{\theta}))^{N_{la}}$ of not selling to any of them. Since $N_{la} \sim GP(\Lambda_{la}, \Phi_a)$, we can use the p.m.f. of the Generalized Poisson distribution to numerically compute the probability that a brand of age a and cost c finds at least one customer in l as

$$
\begin{aligned}
\mathbb{P}(B_{la}(c) > 0) &= 1 - \mathbb{P}(B_{la}(c) = 0) \\
&= 1 - \sum_{k=0}^{\infty} (1 - \exp(-\nu_l c^{\theta}))^k \frac{\Lambda_{la}(\Lambda_{la} + k\phi)^{k-1}}{k!} e^{-\Lambda_{la} - k\phi}.
\end{aligned}
$$

We can then integrate over all brands costs to find the mass of brands that operate in l with age a

$$
F_{la} = \int_0^{\infty} \mathbb{P}(B_{la}(c) > 0) d\mu_{la}(c).
$$

Furthermore, we can add them to find the total number of brands that serve each location as $F_l = \sum_a F_{la}$.

One way to compute the average number of buyers among brands that operate in a market is to add up all the customers that brands find in a given location and divide by the number of brands that are selling in that market:

$$
R_{la} = \int_0^{\infty} \mathbb{E}[B_{la}(c)] d\mu_{la}(c) = \int_0^{\infty} \frac{\exp(-\nu_l c^{\theta})\Lambda_{la}}{1 - \Phi_{la}} d\mu_{la}(c) = \frac{\Lambda_{la}}{1 - \Phi_l} \frac{\delta_e(1-\delta_e)^a T_l w^{-\theta}}{\nu_l}.
$$

The average number of buyers among brands that actively sell to location l is then

$$
\bar{b}_{la} = \frac{R_{la}}{\tilde{F}_{la}}, \text{ and } \bar{b}_l = \frac{\sum_a R_{la}}{\sum_a F_{la}}.
$$

These moments are targeted in the estimation algorithm and are helpful to identify the probability that a consumer forgets about the brand δ_b.

Furthermore, there are closed-form solutions for the variance of customer base, but the solution for the covariances is much involved. These moments are crucial for the identification of the contagion parameters. To proceed with the estimation, I rely on the Simulated Method of Moments. For thousands of brands, I simulate draws for productivities, the evolution of the potential customer base, and whether consumers effectively buy the brands' products. I then use the model-generated data to compute the correlation between the customer base and brands' sales in different locations. I describe the algorithm in more detail in the next section.

IV. Estimation

In this section I describe the algorithm that estimates the model. The first thing to notice is that the parameter β does not affect the equilibrium allocations, and in our context is only relevant for welfare analysis. When performing those exercises we choose a range for β that is compatible with the literature on intertemporal discounting.

I start by recovering an estimate for the elasticity of substitution, σ. Under the model assumptions, if a consumer buys a good with price p, the quantity that they acquire is the following

$$q_l(p) = p^{-\sigma} P_l^{\sigma-1} \frac{\bar{X}_l}{L_l}.$$

The equation states that the quantity demanded depends on a location-specific factor, including the local price index and individual spending, and the price. Since the measures that I have for prices and quantities are indices computed for each brand, I regress the quantities sold by a brand in a location on the price paid there, including location and brand fixed effects.

$$\ln(q_l(\omega)) = -\hat{\sigma} \ln(p_l(\omega)) + I_l + I_\omega + \epsilon_{\omega,l},$$

where ω denotes an individual brand, $q_l(\omega)$, their quantity index and $p_l(\omega)$ their price index. So $\hat{\sigma}$ is the estimate for the elasticity of substitution that is consistent with the patterns of local demand observed in the data

After that, I estimate the remaining parameters of the model, using a Simulated Method of moments. Let $\Theta = (\theta, \delta_e, \delta_b, \boldsymbol{\lambda}, \boldsymbol{\phi})$. For every Θ I pick a vector of T_l that rationalizes the difference in prices observed in the data. That is

$$T_l(\Theta) = \left(\bar{P}_l \Gamma\left(\theta - \frac{\sigma - 1}{\theta}\right)^{\frac{1}{1-\hat{\sigma}}}\right)^{-\theta} \bar{\nu}_l^{-1},$$

where $\bar{\nu}_l := \nu_l / T_l$ and $\bar{P}_l$ is the average of the normalized prices of brands in location l. The parameter T_l defines the average draw of costs in each location and directly affects the price level in a region. A higher T_l increases the brand probability of having low costs in that location ex-ante. But it also influences the competition landscape since it affects the distribution of costs and the probability that the brand effectively sells its product to a potential consumer.

I separate the moments that are targeted by the algorithm into two groups: $m_1(\Theta)$ and $m_2(\Theta)$. The first group represents moments that do not require simulation and can be computed directly. They are the number of brands that sell in each location, their average customer base, and average sales $F_l, \bar{b}_l, \bar{x}_l$. For the year of 2016, I can assign age for brands with $a = 0, \ldots, 8$, so I also target the number of brands, average customer base and average sales by cohort $(F_{la}, \bar{b}_{la}, \bar{x}_{la})_{a=0}^{8}$. The evolution of the number of brands is important to identify the parameter δ_e, which determines the exogenous exit rate of brands. The evolution of the average customer base and sales provide information about the depreciation of customer base δ_b.

The second set of parameters requires simulation. I choose a large number of brands to simulate $\bar{K}$. Then I draw the productivity vectors from the following CDF: $1 - \left(\sum_{l=1}^{\mathcal{L}} \frac{z_l}{(K^{-1}T_l)^{1/\theta}} - \mathcal{L} + 1\right)^{-\theta}$. This distribution is convenient as it has closed-form solutions for the marginal distributions $F(z_l)$, and the conditional distributions $F_l(z_l|(\bar{\mathbf{z}}))$, where $\bar{\mathbf{z}}$ is any subset of $\mathbf{z}$. This allows me to draw $\mathbf{z}$ sequentially: $\bar{z}_1 \sim F_1(z_1)$, $\bar{z}_2 \sim F_2(z_2|\bar{z}_1), \ldots, \bar{z}_{\mathcal{L}} \sim F_{\mathcal{L}}(z_{\mathcal{L}}|\bar{z}_1, \ldots, \bar{z}_{\mathcal{L}-1})$. After that, I simulate the evolution of brand awareness by sequentially following the steps of customer acquisition described in the model: direct search, contagion based on the previous draws and then the number of the customers that forget. I divide the $\bar{K}$ number of brands in cohorts with size proportional to the model implications for brand survival, that is $1 - \delta_e$ to the power of the cohort, and compute their trajectories until the current year. This way, the distribution of brands age is similar to the model predictions.

With the values of productivities and potential customer base in hand, I compute the probability that the consumer buy a product from each brand as $\exp(-\nu_l c_l^{\theta})$. I use the brand-location specific probability of actually selling the good and the number of potential customers to draw the number of actual customers. After that, I compute brand sales by multiplying the customer base by the amount spent by consumer $c_l^{1-\sigma} P_l^{\sigma-1} \frac{\bar{X}_l}{L_l}$. This way, I can compute the same normalized correlations of brand sales and brand customer base as shown in figure 2. I also compute age-specific correlations both in the model and in their data. Those moments are the key elements that guide the choice of the contagion parameters, since the greater the contagion parameters between two regions, the larger their current correlation of sales and customers.

Finally, the algorithm searches for the set $\hat{\Theta}$ that solves the following problem

$$\hat{\Theta} = \text{argmin}_{\Theta} ||m(\Theta) - \bar{m}||, \quad \text{s.t. } \Theta > 0, \quad \phi < 1, \quad \theta > \hat{\sigma} - 1,$$

where the model implied moments are $m(\Theta) = [m_1(\Theta), m_2(\Theta)]$, and $\bar{m}$ are their data counterparts.

I have intentionally not included the lagged correlations of figure 3 as a moment to be targeted. This way, I can evaluate how well the model performs in describing the dynamics of the geographic spread of brands by contrasting the predictions of the estimated model with these moments.

Now I discuss the results of the estimation and the counterfactuals.

V. RESULTS AND COUNTERFACTUALS

The following table summarizes the estimated parameters, and the moments that are associated with their identification.

The estimated model predicts a crucial role for distance in the strength of contagion. To see that,

Parameter	Value	Target
σ	0.1257	regression
θ	4.1753	$corr(x_{la}, x_{ma})$
δ_e	0.1560	F_{la}
δ_b	0.2925	$\mathbb{E}(b_{la})$
ϕ	0.2072	$\mathbb{V}(b_{la})$
C_0	0.5316	$corr(b_{la}, b_{ma})$
C_1	0.0015	$corr(b_{la}, b_{ma})$
λ_l	Various	$\mathbb{E}(b_{l0})$

compare the two closest regions, D.C. and Baltimore, with the ones that are furthest away, Seattle and Miami. For the cities on the opposite side of the country, contagion is very unlikely. The contagion parameter between them is $C_0 * exp(-C_1 2730) = 0.0089$. The contagion parameter between the two cities in the middle of the east coast, on the other hand, is quite large $C_0 * exp(-C_1 35) = 0.5044$. On average, an informed consumer in D.C. spreads information to more than 0.5 consumers in Baltimore. This is more than 50 times the contagion probability between the two distant regions.

To further investigate how geography affects the outcome of brands in the model, I return to figures 2 and 3. They describe how close regions tend to be more similar concerning sales and number of customers. I use the model simulated data to compute the same correlations and plot both the data and model variables for sales and customer base.

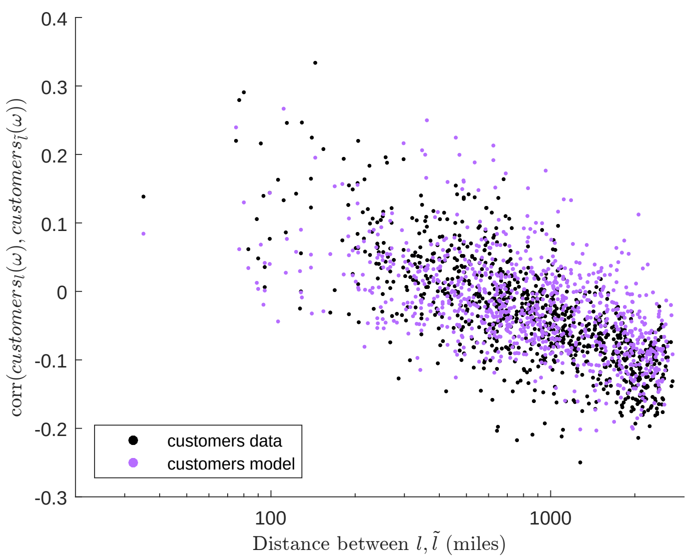


Figure 5: *Correlation of brands relative customer base: data x model*

We can see that the model performs quantitatively well in replicating the geographic concentration patterns observed in the data for sales and customer base. Furthermore, the model predictions for prices and individual quantities are not geographically correlated. The model also replicates the increasing trajectory for average sales and customer base, although with higher levels.

I also evaluate the model predictions with respect to the geographic persistence, by computing the

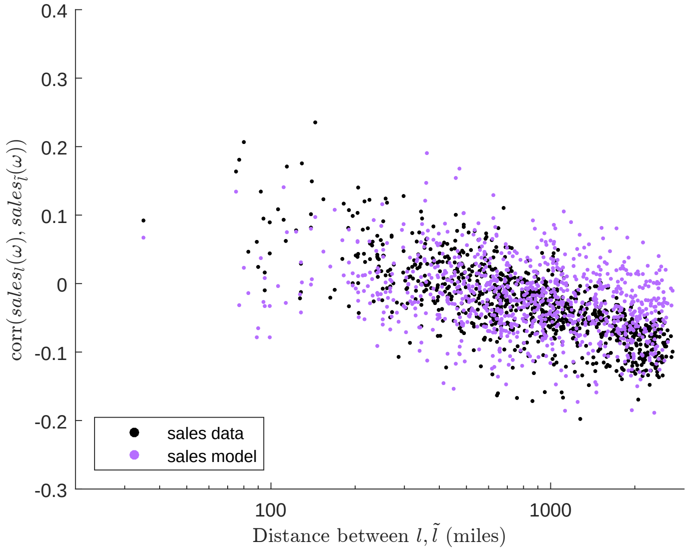


Figure 6: *Correlation of brands relative sales: data x model*

model-predicted correlations for sales and customer base as in figure 4.

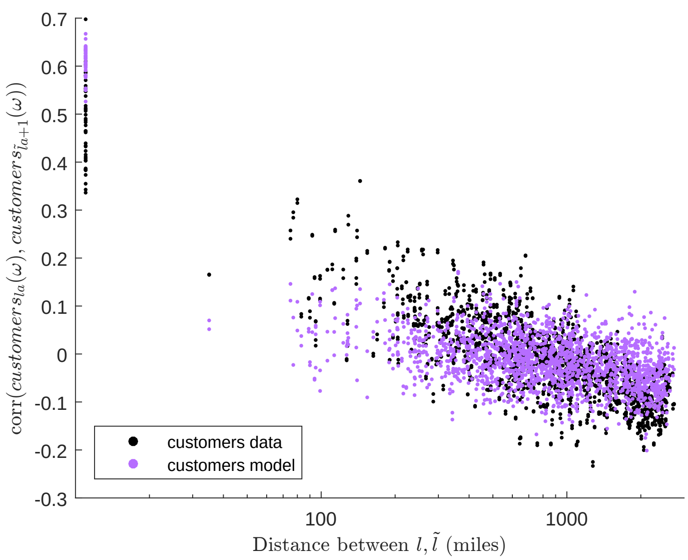


Figure 7: *Correlation of brands relative customers one period ahead: data x model*

We can see that the model generates geographic persistence, but to a lesser degree than observed in the

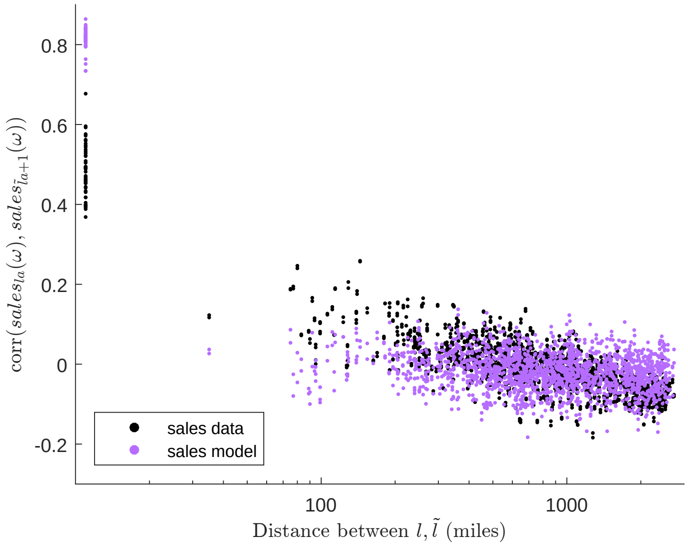


Figure 8: *Correlation of brands relative sales one period ahead: data x model*

data. One reason behind that is the strong assumption imposed on the contagion process that only newly found potential customers might spread the information to others. As discussed before, this assumption reduces the influence that the current customer base has in acquiring customers in the future. Since contagion is more likely to happen in closer regions, this assumption reduces the degree of geographic persistence. Now, I discuss the counterfactuals, using the estimated model as the benchmark.

First, I evaluate the implications of reducing information frictions. I use the model to address the welfare gains of eliminating the role of geography in contagion. In the benchmark, contagion parameters are defined as $\lambda_{lm} = C_0 exp(C_1 \text{distance}_{l,m})$. I assume that in the counterfactual economy, distance does not matter. There, the contagion parameter between any two locations is $\tilde{\lambda}_{lm} = C_0 = 0.5316$. Notice that I compare two stationary economies, meaning that the counterfactual economy *always* had $\tilde{\lambda}_{lm}$ governing the evolution of brand awareness. The computation of new outcomes is straightforward since all equations derived before still hold for the new economy.

The welfare for consumers with CES utility function can expressed by their real income, which is $W_l = \bar{X}_l / P_l L_l$. Since the income and populations remain the same, our measure of welfare gain is $\Delta\tilde{W}_l = \tilde{W}_l / W_l - 1 = P_l / \tilde{P}_l - 1$. The price level in the new economy for all locations is

$$\tilde{P}_l = \tilde{\nu}_l^{-1/\theta} \left[\Gamma \left(1 - \frac{\sigma - 1}{\theta} \right) \right]^{\frac{1}{1-\sigma}},$$

where $\tilde{\nu}_l = \frac{\delta_e T_l}{L_l} \left(\sum_{a=0}^{\infty} \frac{(1-\delta_e)^a \tilde{\Lambda}_{l,a}}{1-\phi} \right)$. The welfare gains are simply $\Delta \tilde{W}_l = \left(\frac{\sum_{a=0}^{\infty} (1-\delta_e)^a \tilde{\Lambda}_{la}}{\sum_{a=0}^{\infty} (1-\delta_e)^a \Lambda_{la}} \right)^{1/\theta} - 1$. The average of the welfare gains across all 44 locations is 32.5%, which is a large number. In the counterfactual economy, information about the existence of brands circulates more. Therefore, consumers are more likely to find brands with lower prices, and consequentially increase consumption. Next, I evaluate the welfare gains of eliminating the effects of geography on costs.

In models where shipping costs among locations are explicit, this exercise can be conducted by reducing them to zero. Consider the case of a firm that has a cost c to produce and sell their good locally but face an iceberg cost of τ to deliver it to another location. In this case, the effect of geography is eliminated by setting $\tau = 1$. This makes selling to other sites as expensive as selling domestically. The same logic is applied in this exercise.

The productivity vector $\mathbf{z}$ accounts for differences in the cost of selling to different locations. We can consider the situation where all brands have their costs reduced to their lowest z_l^{-1} in all areas. This way, each brand might sell its good to all locations at its lowest cost, similar to what eliminating shipping costs would do.

Again, I consider that the counterfactual economy is stationary, meaning that all brands faced the alternative distribution of costs since their beginning. As in the previous exercise, to calculate the welfare, I need to compute the new price indices P'_l. However, this is not straightforward anymore. Consider the productivity vector of a brand $z = (z_1, \ldots, z_{\mathcal{L}})$. In the counterfactual economy, its productivity vector would be $z' = (\overline{z}, \ldots, \overline{z})$, where $\overline{z} = \max(z_1, \ldots, z_{\mathcal{L}})$. The distribution of the maximum entry of the productivity vector does not have the same properties of the original marginal distribution. I briefly describe how I compute the new price distribution numerically.

First, I follow the same steps as the estimation procedure to draw the productivity vector for millions of brands. After that, I find the highest productivity entry for each brand. I assign a brand's cost in every location as the inverse of its highest value $\overline{z}_l$. Then, I estimate the density of the new cost distribution using a standard kernel function. After that, I multiply the estimated density function by the measure of brands simulated $\bar{K}$. The result is $d\mu'(c)$ analogous to $d\mu_l(c)$ in the original model. Here, $\mu'(c)$ is the measure of brands with the *lowest* cost below c. Notice that the brand's cost now does not depend on the location they sell to, but their ability to reach customers is still location-specific.

To compute the price distribution, I calculate the intensity that consumers in l find quotes below c under the new cost distribution

$$\rho_l'(c) = \left[\sum_{a=0}^{\infty} \frac{\Lambda_{la}\delta_e(1-\delta_e)^a}{L_l(1-\phi)} \right] \int_0^c d\mu'(c).$$

The new price distribution is $G_l'(c) = 1 - \exp(-\rho_l'(c))$, and I can compute the new price indices as

$$P_l' = \left[\int_0^{\infty} p^{1-\sigma} dG_l'(p) \right]^{\frac{1}{1-\sigma}}.$$

Now, I can compute the welfare gains in each location $\Delta W_l'$. The average of the welfare gains across all locations is 57%. This is also a considerable improvement in consumption associated with reducing the costs of brands that consumers are already aware of.

The interpretation of the last counterfactual as the reduction of shipping costs requires an important caveat. Say that the model had a considerably larger number of locations, *i.e.* take $\mathcal{L}$ to be 1000. When the largest productivity is selected, the odds that any brand achieves a high number is substantial. Therefore, the cost reduction associated with geography is also conflated with increased expected productivity associated with more draws.

Nevertheless, the counterfactuals show that significant welfare gains are associated with reducing the impacts geography has on contagion and reducing the costs of brands that consumers already know.

VI. CONCLUSION

This paper studies how brand sales evolve over time and space in the United States. I show evidence of a significant role for geography in the growth of brands. Unlike traditional spatial economics and trade models, these differences are not associated with changes in costs and prices.

To reconcile the data with economic modeling, I posit that geography can affect a brand's customer base through informational frictions that are not directly associated with changes in prices. I provide a parsimonious model of the geographic spread of brand awareness that relies on contagion: customers aware of a brand might infect unaware consumers with their knowledge. The model can replicate the stylized facts about how brand sales and customer base evolve. Furthermore, the estimates show that the proposed information frictions are more severe between distant locations. The model can also contrast the welfare costs associated with reducing the role of geography in prices and the flow of information. Finally, it can also evaluate how much expected revenue a single aware customer generates by considering the spread of this knowledge and the probability that consumers reached will effectively buy the product.

VII. Appendix - Tables and Figures

Table 2: *Summary Statistics for Locations in 2016*

Location	# of Brands	Sales ($MM)	Location	# of Brands	Sales ($MM)
New York	28,771	33,967	Portland	17,664	7,161
Los Angeles	25,569	25,574	Orlando	17,952	7,024
Philadelphia	24,922	15,605	Richmond	17,879	7,009
Boston	22,731	14,328	Salt Lake City	16,317	6,528
Chicago	25,238	14,212	Sacramento	19,482	6,521
Washington, D.C.	23,094	13,158	Hartford	16,195	5,889
Dallas	22,151	12,890	Birmingham	16,317	5,886
San Francisco	20,017	12,406	Cincinnati	17,613	5,857
Miami	20,279	11,534	Indianapolis	16,398	5,630
Houston	21,592	11,207	Oklahoma City	15,025	5,421
Tampa	21,541	11,134	Nashville	15,933	5,336
Phoenix	21,615	10,933	Baltimore	16,893	5,230
Atlanta	21,050	10,862	St. Louis	17,696	5,217
Detroit	21,687	10,480	San Diego	15,591	5,054
Seattle	20,254	8,935	Grand Rapids	16,043	4,901
Denver	19,429	8,557	Columbus	17,492	4,827
Cleveland	20,047	8,502	Charlotte	17,453	4,793
Raleigh	18,840	8,290	Kansas City	16,236	4,707
San Antonio	18,908	7,884	Louisville	16,303	4,637
Pittsburgh	17,976	7,380	Buffalo	17,407	4,611
Minneapolis	19,695	7,247	Milwaukee	15,535	4,246
New Orleans	16,928	7,198	Memphis	13,514	4,014

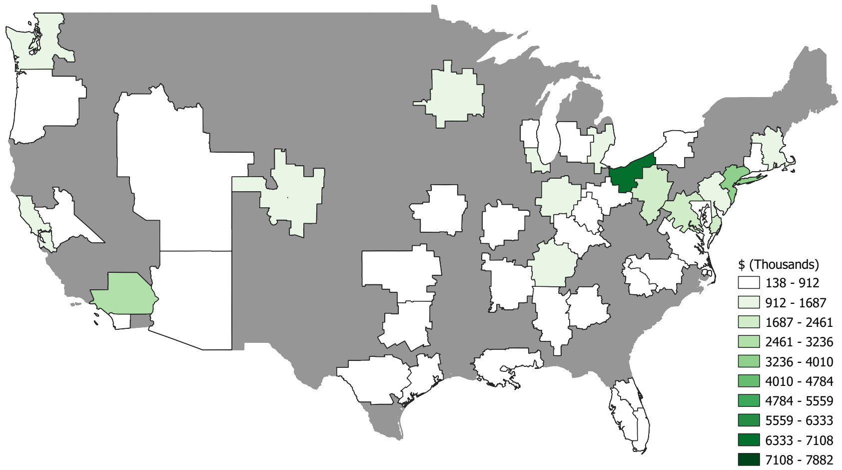


Figure 9: *Sales of Brands that started in Cleveland - 2008*

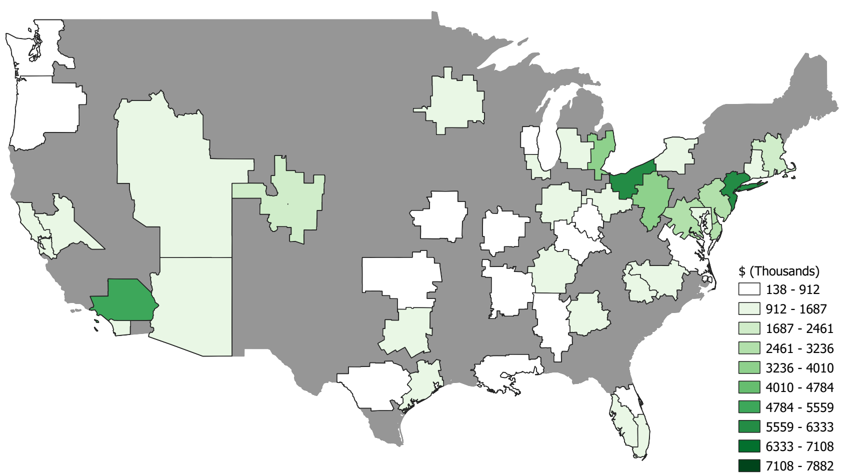


Figure 10: *Sales of Brands that started in Cleveland - 2009*

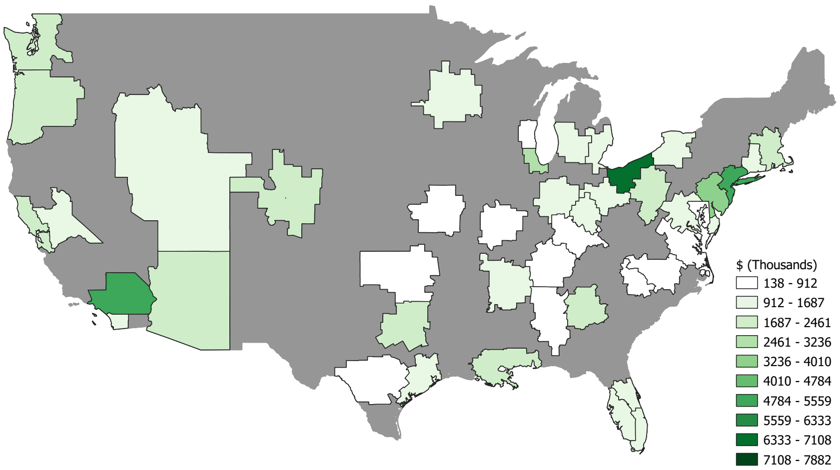


Figure 11: *Sales of Brands that started in Cleveland - 2010*

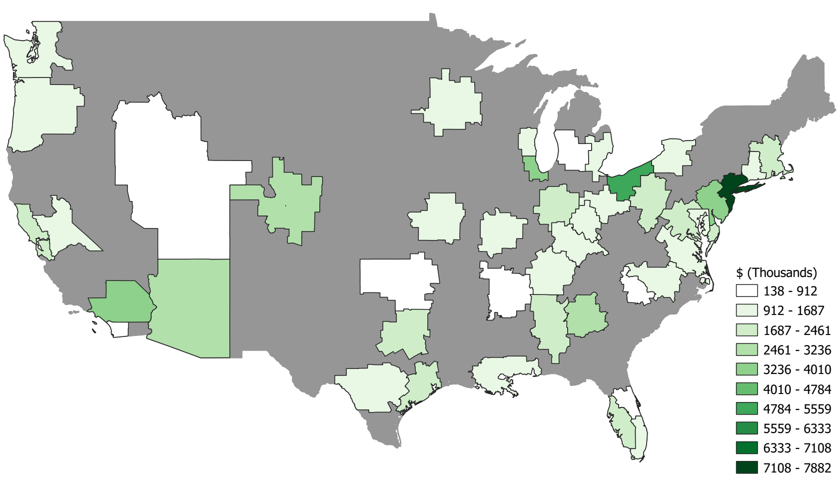


Figure 12: *Sales of Brands that started in Cleveland - 2011*

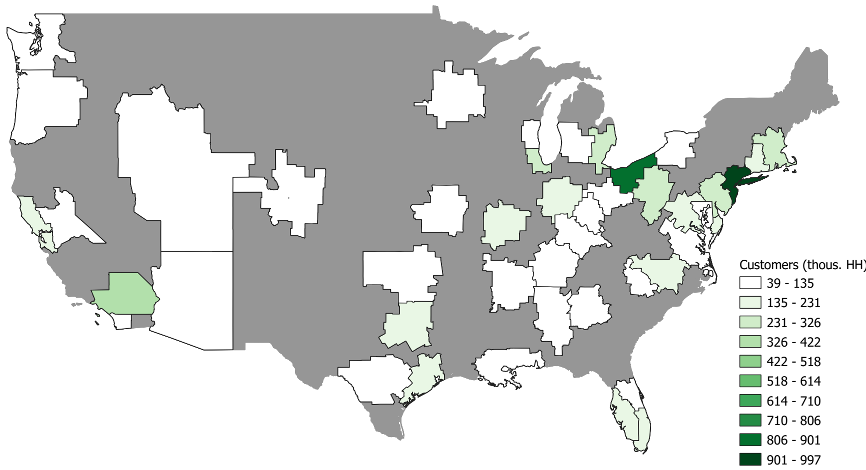


Figure 13: *Customers of Brands that started in Cleveland - 2008*

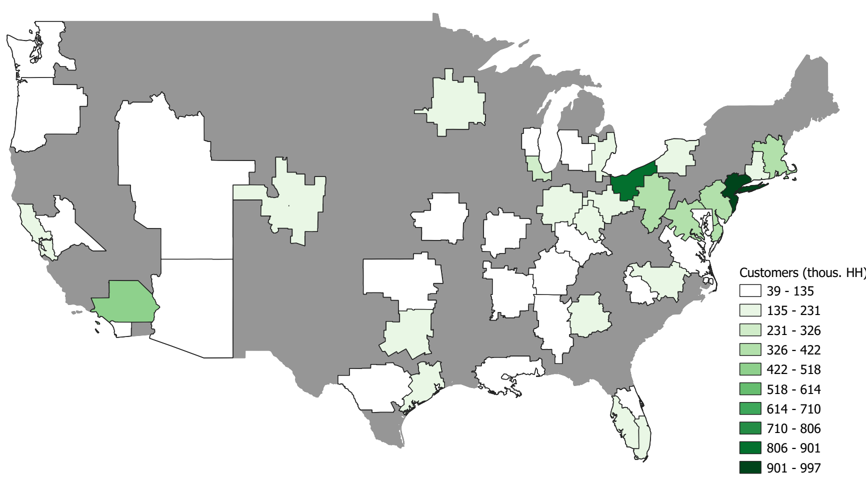


Figure 14: *Customers of Brands that started in Cleveland - 2009*

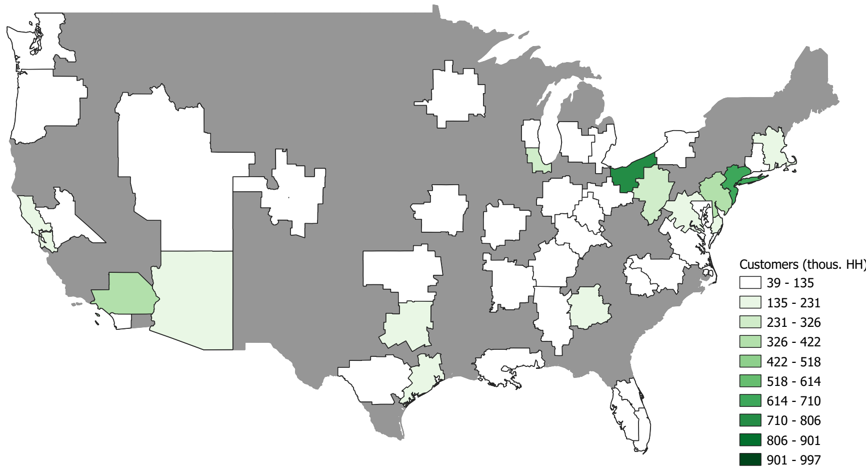


Figure 15: *Customers of Brands that started in Cleveland - 2010*

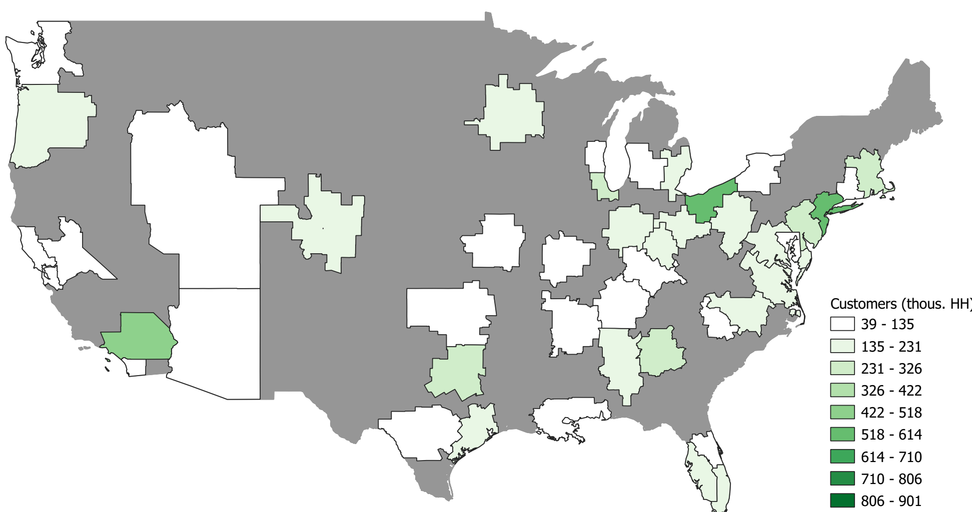


Figure 16: *Customers of Brands that started in Cleveland - 2011*

REFERENCES

Albornoz, F., Calvo Pardo, H., Corcos, G., and Ornelas, E. (2012). Sequential exporting. *Journal of International Economics*, 88(1):17–31.

Argente, D., Lee, M., and Moreira, S. (2018). How do firms grow? the life cycle of products matters. 2018 Meeting Papers 1174, Society for Economic Dynamics.

Arkolakis, C. (2010). Market Penetration Costs and the New Consumers Margin in International Trade. *Journal of Political Economy*, 118(6):1151–1199.

Bagwell, K. (2005). The economic analysis of advertising. *forthcoming, Handbook of Industrial Organization*, 3.

Bailey, M., Cao, R. R., Kuchler, T., Stroebel, J., and Wong, A. (2017). Measuring social connectedness. Working Paper 23608, National Bureau of Economic Research.

Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science*, 15(5):215–227.

Bronnenberg, B. and Albuquerque, P. (2003). Geography and marketing strategy in consumer packaged goods. *Advances in Strategic Management*, 20.

Bronnenberg, B., Dube, J.-P., Gentzkow, M., and Shapiro, J. (2014). Do pharmacists buy bayer? informed shoppers and the brand premium.

Bronnenberg, B. J., Dhar, S. K., and Dubé, J. H. (2009). Brand history, geography, and the persistence of brand shares. *Journal of Political Economy*, 117(1):87–115.

Bronnenberg, B. J., Dubé, J.-P. H., and Gentzkow, M. (2012). The evolution of brand preferences: Evidence from consumer migration. *American Economic Review*, 102(6):2472–2508.

Campa, J. and Goldberg, L. (2005). Exchange rate pass-through into import prices. *The Review of Economics and Statistics*, 87(4):679–690.

Caplin, A., Dean, M., and Leahy, J. (2017). Rationally inattentive behavior: Characterizing and generalizing shannon entropy. NBER Working Papers 23652, National Bureau of Economic Research, Inc.

Caplin, A., Leahy, J., and Matějka, F. (2015). Social Learning and Selective Attention. NBER Working Papers 21001, National Bureau of Economic Research, Inc.

Chandrasekaran, D. and Tellis, G. (2007). A critical review of marketing research on diffusion of new products. *Review of Marketing Research*, 3.

Chaney, T. (2014). The network structure of international trade. *American Economic Review*, 104(11):3600–3634.

Chevalier, J. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *J. Marketing Res*, pages 345–354.

Consul, P. C. (1989). *Generalized Poisson Distributions: Applications and Properties*. Marcel Dekker, Inc., New York.

Consul, P. C. and Jain, G. C. (1973). A generalization of the poisson distribution. *Technometrics*, 15(4):791–799.

Consul, P. C. and Mittal, S. P. (1975). A new urn model with predetermined strategy. *Biometrische Zeitschrift*, 17(2):67–75.

Dinlersoz, E. and Yorukoglu, M. (2008). Informative advertising by heterogeneous firms. *Information Economics and Policy*, 20(2):168–191.

Draganska, M. and Klapper, D. (2011). Choice set heterogeneity and the role of advertising: An analysis with micro and macro data. *Journal of Marketing Research*, 48(4):653–669.

Drozd, L. A. and Nosal, J. B. (2012). Understanding International Prices: Customers as Capital. *American Economic Review*, 102(1):364–395.

Eaton, J., Kortum, S., and Kramarz, F. (2011). An anatomy of international trade: Evidence from french firms. *Econometrica*, 79(5):1453–1498.

Eaton, J., Kramarz, F., and Kortum, S. (2019). Firm-to-Firm Trade: Exports, Imports, and the Labor Market. 2019 Meeting Papers 702, Society for Economic Dynamics.

Eslava, M., Tybout, J., Jinkins, D., Krizan, C., and Eaton, J. (2015). A Search and Learning Model of Export Dynamics. 2015 Meeting Papers 1535, Society for Economic Dynamics.

Evenett, S. and Venables, A. (2002). Export growth in developing countries: Market entry and bilateral trade flows.

Fitzgerald, D., Haller, S., and Yedid-Levi, Y. (2016). How exporters grow. 2016 Meeting Papers 499, Society for Economic Dynamics.

Gourio, F. and Rudanko, L. (2014). Customer Capital. *Review of Economic Studies*, 81(3):1102–1136.

Hauser, J., Tellis, G. J., and Griffin, A. (2006). Research on innovation: A review and agenda for "marketing science". *Marketing Science*, 25(6):687–717.

Hillberry, R. and Hummels, D. (2008). Trade responses to geographic frictions: A decomposition using micro-data. *European Economic Review*, 52(3):527–550.

Kalish, S. (1985). A new product adoption model with price, advertising, and uncertainty. *Management Science*, 31(12):1569–1585.

Lenoir, C., Mejean, I., and Martin, J. (2018). Search Frictions in International Good Markets. 2018 Meeting Papers 878, Society for Economic Dynamics.

Lim, K. (2017). Firm-to-firm Trade in Sticky Production Networks. 2017 Meeting Papers 280, Society for Economic Dynamics.

Mahajan, V. and Peterson, R. A. (1979). Integrating time and space in technological substitution models. *Technological Forecasting and Social Change*, 14(3):231 – 241.

Mardia, K. V. (1962). Multivariate pareto distributions. *Ann. Math. Statist.*, 33(3):1008–1015.

Morales, E., Sheu, G., and Zahler, A. (2019). Extended Gravity. *Review of Economic Studies*, 86(6):2668–2712.

Oakes, J., Andrade, K., Biyoow, I., and Cowan, L. (2015). Twenty years of neighborhood effect research: An assessment. *Current Epidemiology Reports*, 2.

Perla, J. (2019). A model of product awareness and industry life cycles. Working paper.

Piveteau, P. (2015). An empirical dynamic model of trade with consumer accumulation. Technical report, Mimeo.

Rasouli, S. and Timmermans, H. (2013). Influence of social networks on latent choice of electric cars: A mixed logit specification using experimental design data. *Networks and Spatial Economics*, 16.

Rogers, E. M. (1962). Diffusion of Innovations. New York: The Free Press of Glencoe, 1962. *Social Forces*, 41(4):415–416.

Romer, P. M. (1987). Growth Based on Increasing Returns Due to Specialization. *American Economic Review*, 77(2):56–62.

Sattenspiel, L. (2009). *The Geographic Spread of Infectious Diseases: Models and Applications*. Princeton University Press, Princeton.

Shoukri, M. M. and Consul, P. C. (1987). *Some Chance Mechanisms Generating the Generalized Poisson Probability Models*, pages 259–268. Springer Netherlands, Dordrecht.

Sovinsky, M. (2008). Limited information and advertising in the u.s. personal computer industry. *Econometrica*, 76(5):1017–1074.