An Alternative Theory of the Plant Size Distribution, with Geography and Intra- and International Trade

Thomas J. Holmes

University of Minnesota, Federal Reserve Bank of Minneapolis,
and National Bureau of Economic Research

John J. Stevens

Board of Governors of the Federal Reserve System

Abstract

There is wide variation in the sizes of manufacturing plants, even within the most narrowly defined industry classifications. Standard theories attribute such size differences to productivity differences. This paper develops an alternative theory in which industries are made up of large plants producing standardized goods and small plants making custom or specialty goods. It uses confidential Census data to estimate the parameters of the model. The model fits the data well. In particular, the predictions of the model regarding the effect of a surge of imports from China are consistent with what happened over the period 1997–2007.

The research presented here was funded by the National Science Foundation under Grant SES 0551062. We thank Brian Adams, Steve Schmeiser, and Julia Thornton Snider for their research assistance. We are grateful for detailed comments of the referees and the editor, and the comments of many seminar participants. We have benefited from working at the Center for Economic Studies at the Census Bureau and the Minnesota Census Research Data Center, and particularly thank Shawn Klimek, Arnie Reznek, Lynn Riggs, and Gina Erickson. The statistics reported in this paper that were derived from Census Bureau micro data were screened to ensure that they do not disclose confidential information. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis, the Federal Reserve Board, the Federal Reserve System, or the U.S. Bureau of the Census.

1 Introduction

The sizes of manufacturing plants exhibit wide variation, even within the most narrowly defined industry classifications used by statistical agencies. For example, in the wood furniture industry (NAICS industry code 337122), one can find plants with over a thousand employees and other plants with as few as one or two employees. The dominant theory of such *within-industry* plant-size differentials models plants as varying in terms of productivity. (See Lucas (1978), Jovanovic (1982), and Hopenhayn (1992).) In this theory, some plants are lucky and draw high productivity at start-up, while others are unlucky and draw low productivity. The size distribution is driven entirely by the productivity distribution.

The approach has been extremely influential. It underpins recent developments in the international trade literature. Melitz (2003) and Bernard, Eaton, Jensen, and Kortum (2003, hereafter BEJK) use the approach to explain plant-level trade facts. In Melitz, plants with high productivity draws have large domestic sales and also have the incentive to pay fixed costs to enter export markets. In this way, the Melitz model explains the fact—documented by Bernard and Jensen (1995)—that large plants within narrowly defined industries are more likely to be exporters than small plants. Relatedly, in BEJK, more productive plants have wider trade areas. Both the Melitz and the BEJK theories have a sharp implication about the effect of increased exposure to import competition on a domestic industry: the smaller plants in the industry—which are also the low productivity plants in the industry under these theories—are the first to exit.

In our view, the dominant approach to modeling plant-size differentials goes too far in attributing all variation in plant size within narrowly defined Census industries to differences in productivity. It is likely that plants that are dramatically different in size are performing different functions, even if the Census Bureau happens to classify them in the same industry. Moreover, these differences in function may be systematic and may very well be directly related to how increased import competition would affect the plants.

Take wood furniture as an example. The large plants in this industry with more than a thousand employees historically have been concentrated in North Carolina, centered around a place called High Point. These plants make the stock bedroom and dining room furniture pieces found at traditional furniture stores. Also included in the Census classification are small facilities making custom pieces to order, such as small shops employing skilled Amish craftsmen. Let us apply the standard theory of the size distribution to this industry. Entrepreneurs that enter and draw high productivity parameters would likely open up megaplants in High Point, North Carolina; those that get low draws might open Amish shops in other locations. The Melitz model and the BEJK model both predict that the large North Carolina plants will have large market areas, while the small plants will tend to ship locally. So far so good, because this result is consistent with the data, as we will show. But what happens when China enters the wood furniture market in a dramatic fashion, as has recently occurred? While all of the U.S. industry will be hurt, the Melitz and BEJK theories predict that the North Carolina industry will be relatively less affected because it is home to the large, productive plants. In fact, the opposite turns out to be true in the data.

To address this shortcoming, our theory takes into account that most industries have some segment that provides *specialty* goods, often custom-made goods, the provision of which is facilitated by face-to-face contact between buyers and sellers. The specialty segment is the province of small plants. Large plants instead produce in what we call the *primary* segment of an industry. Large production volumes in the primary segment facilitate standardization and the use of mass-production techniques. Here we follow the ideas of Piore and Sabel (1984) and a subsequent literature distinguishing between mass production in large plants and the craft production of specialty products in small plants. When China enters the wood furniture market, naturally it follows its comparative advantage and enters the standardized, or primary, segment of the market, making products similar to the stock furniture pieces produced in North Carolina. Thus, in our theory, the North Carolina industry is hurt the most by China's entry into the industry, as actually happened.

Our starting point is the Eaton and Kortum (2002) model of geography and trade as further developed in BEJK. We treat the model as applying at the level of a *product segment* within a Census-defined industry. The primary segment of an industry will have one set of BEJK parameters; specialty segments in the same industry will have other BEJK parameters. In particular, specialty products may entail customization, and we argue this is analogous to having high transportation cost parameters. An alternative formulation that ends up having the same reduced form treats specialty goods as niche goods with idiosyncratic sources of supply. We use this theory to explore two issues. First, how is the size distribution of plants connected to the geographic distribution of plants? (We call this the plant size/geographic concentration relationship.) Second, if there is a surge in imports, what is the relative effect of the trade shock across locations that vary by geographic concentration and mean plant size?

We estimate the model separately for individual industries, using Census data that include survey information from the Commodity Flow Survey on the origins and destinations of shipments. The shipment information is critical for our analysis because it enables us to recover parameters related to the transportation cost structure in the BEJK framework.

We obtain four main empirical results. First, we estimate that in most industries, more than half of the plants in an industry can be classified as being specialty segment plants. Second, a model with only a primary segment and no specialty segments fails to quantitatively match the plant size/geographic concentration relationship, whereas the full model including specialty segments fits this relationship well. For example, in High Point where the wood furniture industry concentrates, average plant size (as measured by sales revenue) is 6.6 times the national average. The estimated primary-only model predicts a factor of only 1.6, whereas the full model fits the data well with a factor of 6.9. The third empirical result concerns industries negatively affected by a surge of imports from China. In the data, when we look at leading locations with large plants like High Point in wood furniture, the more the industry is hit by imports, the greater the fall of the leading locations relative to the rest of the domestic industry. When we simulate the estimated model with only primary segments, the leading locations of industries hit by imports actually increase their share of the remaining domestic industry, contrary to the pattern in the data. When we simulate the estimated full model that includes specialty segments, the out-of-sample predictions fit the qualitative patterns of the data, and the magnitudes are roughly consistent as well. Fourth, we use the procedure to estimate changes in the split between the primary and specialty segments over time. In those industries that have been especially hard-hit by imports from China, the primary segment share of the remaining plant counts has declined significantly.

A broader contribution of this paper is that it pushes models originally designed to explain international trade to account for facts about industrial organization, geography, and intranational trade. Section 2 will lay out these facts. This paper emphasizes predictions about multiple source locations, a dimension that is largely overlooked in the micro-trade literature, since data are typically available from just one exporting country (although many importing countries). It improves on these models in a way that matters for predictions as basic as what types of plants get hurt by increased foreign competition.

There is an emerging new literature that allows for richer forms of heterogeneity across plants than the first generation of trade models with heterogeneous firms found in Melitz (2003) and BEJK. Hallak and Sivadasan (2009) allow plants to differ in the standard way regarding cost structure but also along a second dimension in terms of a plant's ability to provide quality. Baldwin and Harrigan (2011) allow firms to be differentiated by quality. Bernard, Redding, and Schott (2010) develop a multiple-product model of a firm with heterogeneity not only at the firm level but also at the level of product-specific attributes. Our paper is in the spirit of these papers in that it allows for richer heterogeneity. One difference is that we add heterogeneity in the extent to which goods are tradable. Holmes and Stevens

(2004) include a margin like this in a regional model linking plant size and geographic concentration. This paper is different from our earlier paper because it (1) uses BEJK to develop an entirely different modeling structure, (2) takes the model to the data and estimates its parameters, and (3) examines the effect of a trade shock.

Jensen and Kletzer (2005) develop an approach to inferring the tradability of a good by comparing the geography of the production of a good with the geography of the absorption of a good. The greater the discrepancy, the greater the apparent flow of goods, suggesting greater tradability. Our paper highlights that the geography of production of large plants is very different from the geography of production of small plants in the same narrow Census industry. Our use of this information to infer differences in the function of large and small plants is in the spirit of Jensen and Kletzer (2005).

In a recent paper, Autor, Dorn, and Hanson (2013) examine the effects of Chinese imports on labor market outcomes in the United States. Like our paper, their paper examines the effects of trade at the subnational level, exploiting differences in industry composition across regions. More broadly, there is a growing literature, such as Bloom, Draca, and Van Reenan (2011), on how industries are being affected by the emergence of China.

Finally, our paper is related to the macroeconomics literature on quantitative dynamic models incorporating plant heterogeneity. This literature makes heavy use of the standard theory in which size differences are driven entirely by productivity differences. Given a monotonic relationship between plant size and productivity, it is possible to invert the relationship and read off the distribution of plant productivities from the distribution of plant sizes. Hopenhayn and Rogerson (1993) is an early example, and Alessandria and Choi (2007) is a recent example. To the extent plants of different sizes are doing different things, this strategy overstates productivity differences. This paper is also related to Buera and Kaboski (2012), which emphasizes how structural change affects average plant size.

2 Basic Facts with Which a Theory Needs to Contend

In this section we list four basic facts with which a theory of the within-industry size distribution of plants needs to contend.

First, the degree of variation in size across manufacturing plants is remarkable, even when we drill down to the finest industry classifications available. Since 1997, the Census has classified plants according to the *North American Industry Classification System,* or *NAICS.* The finest classification level is the 6-digit NAICS level, e.g., NAICS 337122 for wood furniture mentioned in the introduction. There are 473 different 6-digit NAICS manufacturing industries. Suppose we start with the 360,000 manufacturing plants in the 1997 Census across all industries and sort by plant employment. The 10th percentile plant has only a single employee, while the 90th percentile plant has 97 employees, yielding a 90/10 ratio of 97 for manufacturing as a whole.[1] Drilling down to the industry level, we find that in three-quarters of all industries, the 10th percentile plant has two employees or fewer. In half of the industries, the 90/10 ratio is over 100, and in three-quarters, the 90/10 ratio is over 50. Suppose we decompose the variance in log employment across manufacturing plants into a *cross-industry* component and a *within-industry* component, by differencing out 6-digit NAICS means. The within-industry component accounts for 82 percent of the total variance in log employment. We conclude that to an overwhelming extent, variation in size across plants is a *within*-industry phenomenon.

Second, the size distribution tends to be skewed to the right, and a large fraction of plants in a typical industry are very small. While the average manufacturing plant size in the 1997 Census has 46 employees, 67 percent of all plants have fewer than 20 employees, which we will define as a *small plant* for this discussion. Small plants, while being in the majority, account for a small percentage of overall employment (8.6 percent). This skewness property holds within industries as well. In 323 out of 473 NAICS industries, at least 45 percent of the plants are small, and these industries account for 75 percent of total manufacturing employment. In 423 industries accounting for 90 percent of total employment, at least 26 percent of the plants in each industry are small. "Iron and Steel Mills" and "Cigarettes" are two industries that are famous for having huge plants—for the Cigarettes industry, mean size exceeded 1,300 employees. Yet even in these two industries, a large fraction of plants are small (23 percent in both cases).

Third, large plants tend to ship to more distant destinations compared with small plants in the same industry. In a well-known paper, Bernard and Jensen (1995) show that, controlling for industry, larger plants are more likely to export. An export is a particular kind of distant shipment. Holmes and Stevens (2012) show that an analogous pattern holds for domestic shipments: small plants (fewer than 20 employees) ship nearby at a significantly higher rate than large plants (more than 500 employees). Taking into account 6-digit NAICS industry controls, the probability that a shipment destination is within 100 miles of the originating plant is 38 percent for a small plant and only 20 percent for a large plant. (See Panels B and D of Table 7 in Holmes and Stevens 2012.)

Fourth, small plants tend to be dispersed, following the distribution of population, while

[1] The statistics reported in this paragraph are estimates based on the public tabulations of the 1997 Economic Census of cell counts by detailed employment range categories. The separate online Appendix explains our procedure.

large plants concentrate near other plants in the same industry. We call this the *plant size/geographic concentration relationship*. Holmes and Stevens (2002) document the pattern, and Table 1 illustrates it here. A common measure of the extent to which a particular industry concentrates in a particular location is the *location quotient*. It is defined as a location's share of total sales in an industry divided by a location's share of total population. If this ratio is significantly above one, the industry is highly concentrated at the location. To construct Table 1, for each dollar value of shipments, associate the location quotient of the 6-digit industry/location where the shipment originated. Then take the mean dollar-weighted value across shipments, conditioning on size of the plant. Locations are partitioned into population deciles in such a way that the maximum value the measure can take is 10. (This maximum occurs if locations with 10 percent of the population contain 100 percent of the industry.) The column labeled "Raw" in Table 1 contains the raw means. A plant in the largest size category is very likely to be in a location with a high concentration of its 6-digit industry, as the mean value is 6.84, more than two-thirds of the maximum possible value.[2] In contrast, the mean is close to one for the bottom size class, meaning the smallest plants tend to follow population. Industry effects are part of the story, but even after we include controls for 6-digit industries (the last column in Table 1), there remains a sizable relationship between plant size and concentration. In particular, the concentration measure increases by more than 50 percent from the smallest to largest size classes.

3 Theory

We develop a model of a product *segment* that follows BEJK. Each product segment is a particular version of the BEJK model, with its own model parameters. An *industry* will be defined as a set of product segments that are grouped together.

In BEJK, the geographic units are countries. Here we will work at the subnational level. We begin with narrowly defined geographic units that we refer to as *locations*. There are L locations, indexed by ℓ. We then group nearby locations together into *regions*.

3.1 Product Segments

Product segments are indexed by k. Each product segment consists of a set of differentiated goods that are aggregated to construct a segment-level composite commodity. Let $j \in [0, \nu^k]$

[2] Similar to Holmes and Stevens (2002), we can show that the pattern is unrelated to the issues raised in Ellison and Glaeser (1997).

index a particular differentiated good in segment k, with ν^k parameterizing the measure of different varieties. In BEJK, there is a single segment, and variety is normalized to one. Here, we have multiple segments and can normalize the variety parameter for one of the segments, i.e., $\nu^1 = 1$. Let q^k be a quantity of composite k. The utility for a consumption bundle of segment composites is Cobb-Douglas, with spending share β^k for segment k. We examine each segment in partial equilibrium, which, under Cobb-Douglas, means we can take spending x_ℓ^k on segment k at location ℓ as fixed. Now, to simplify notation, we leave implicit the index k of the particular segment being considered.

Differentiated goods within a particular segment are aggregated to the composite segment level in the usual constant elasticity of substitution (CES) way. Suppose the price of differentiated good j at location ℓ is given by $P_\ell(j)$. Let p_ℓ denote the overall price index of the product segment at ℓ. (Uppercase refers to a differentiated good, and lowercase refers to a composite.) With CES and elasticity of substitution σ, the price index equals

$$p_\ell = \left[\int_0^\nu P_\ell(j)^{1-\sigma} dj\right]^{\frac{1}{1-\sigma}}.$$

Let $X_\ell(j)$ be spending on product j at ℓ. With the CES structure, $X_\ell(j)$ has the following relationship to overall spending x_ℓ on the segment at ℓ:

$$X_\ell(j) = x_\ell \left(\frac{P_\ell(j)}{p_\ell}\right)^{1-\sigma}.$$

Following BEJK, there are potential producers at each location with varying levels of technical efficiency. Let $Z_{r,\ell}(j)$ index the efficiency of the rth most efficient producer of good j located at ℓ. This index represents the amount of good j that can be made by this producer, per unit of input. We impose the same Fréchet structure as in BEJK. In particular, let T_ℓ be the scaling parameter of the Fréchet distribution governing the distribution of productivity in the segment at location ℓ, and let θ be the curvature parameter.

Let d_{ℓ',ℓ° be the distance between locations ℓ° and ℓ'. Transportation cost takes the usual iceberg form: in order to deliver one unit at distance d, $\tau(d) \geq 1$ units must be shipped. Assume $\tau(0) = 1$, i.e., there is no transportation cost at the limit where the distance shipped is zero. Assume $\tau(d)$ is (weakly) increasing in d.

Eaton and Kortum (2002) show that for a particular differentiated good j, the probability that location ℓ° contains the lowest-cost producer for serving location ℓ' is

$$\phi_{\ell',\ell^{\circ}} = \frac{a_{\ell',\ell^{\circ}}\gamma_{\ell^{\circ}}}{\sum_{\ell=1}^{L} a_{\ell',\ell}\gamma_{\ell}} \tag{1}$$

for

$$\gamma_{\ell} \equiv T_{\ell} w_{\ell}^{-\theta}, \tag{2}$$

where w_{ℓ} is the cost of inputs at location ℓ, and

$$a_{\ell',\ell} \equiv \tau \left(d_{\ell',\ell}\right)^{-\theta}. \tag{3}$$

We refer to γ_{ℓ} as the *cost efficiency index* for location ℓ and $a_{\ell',\ell^{\circ}}$ as the *distance adjustment* between ℓ° and ℓ'. Let $\Gamma = (\gamma_1, \gamma_2, ..., \gamma_L)$ be the cost efficiency vector and A (with elements $a_{\ell',\ell^{\circ}}$) be the distance adjustment matrix. We can think of $a_{\ell',\ell^{\circ}}\gamma_{\ell^{\circ}}$ as an index of the competitiveness of source ℓ° at destination ℓ'. It takes location ℓ°'s overall cost efficiency and adjusts for distance to ℓ'. Location ℓ°'s probability $\phi_{\ell',\ell^{\circ}}$ of getting the sale at ℓ' equals its own competitiveness at ℓ' relative to the sum of all the other locations' indexes of competitiveness at ℓ'.

BEJK consider a rich structure with multiple potential producers at each location who each get their own productivity draws. Then firms engage in Bertrand competition for consumers at each location. The equilibrium may feature limit pricing, where the lowest-cost producer matches the price of the second-lowest-cost producer. Or the lowest cost may be so low relative to rivals' costs that the price is determined by the inverse elasticity rule for the optimal monopoly price. The very useful result of BEJK is that despite this complexity, the outcome has a simple structure. The distribution of prices to ℓ' is the same regardless of the source location ℓ°. This implies that sales revenue at destination ℓ' is allocated across sources ℓ° according to $\phi_{\ell',\ell^{\circ}}$. Thus, total revenue at source ℓ° from sales to destination ℓ' is given by

$$y_{\ell',\ell^{\circ}} = \phi_{\ell',\ell^{\circ}} x_{\ell'}. \tag{4}$$

Total revenue at source ℓ° across all destinations equals

$$y_{\ell^{\circ}} \equiv \sum_{\ell'=1}^{L} y_{\ell',\ell^{\circ}} = \sum_{\ell'=1}^{L} \phi_{\ell',\ell^{\circ}} x_{\ell'}. \tag{5}$$

Like BEJK, we associate production of a particular good j at ℓ as taking place within a *plant.* The number of plants producing in a given segment at a given location then depends on the number of goods produced at the location. Let ζ be a product-segment level parameter specifying the number of goods produced per plant. We can then derive the plant counts in a given segment at location ℓ as follows. The variety of goods in the segment produced locally at ℓ equals $\phi_{\ell,\ell}\nu$, overall segment variety ν times the local share $\phi_{\ell,\ell}$.[3] Plant counts then equal

$$n_\ell = \lambda\phi_{\ell,\ell} \tag{6}$$

for a scaling parameter λ equal to the ratio of overall variety ν to goods per plant ζ,

$$\lambda \equiv \frac{\nu}{\zeta}.$$

This parameter will appear in equation (16) below, which we use to estimate the number of plants in each segment.

3.2 Aggregating Locations to Regions

When we go to the data, the geographic units we use will vary in geographic coverage, and some geographic units will be larger than other units on account of the sometimes arbitrary ways in which different locations are grouped together into contiguous units. We extend the above framework to take this aggregation issue into account. Our conceptual approach is to treat the geographic units in the data as *regions* that are aggregations of *locations* as defined in the theory above. Here we show how to aggregate distances within a region to arrive at a region-level internal distance measure that we use in estimation. In constructing a measure of internal distance, we follow the lead of earlier work that takes internal distance as well as external distance into account. (See Anderson and van Wincoop (2003) and the references therein.)

Let regions be indexed by $r = 1, 2, ..., R$. Let Λ_r be the set of locations ℓ contained in region r. For example, one of the regions in the data is the New York City region. This region consists of the various locations (e.g., towns or neighborhoods) in the northern half of New Jersey, plus parts of Pennsylvania and Connecticut, as well as the area of New York State near the city.

Define the region-level variable $\bar{\phi}_{r',r^\circ}$ as the share of shipments from any source location in

[3] If a particular plant is the most efficient producer at any location, it must also be the most efficient producer at its own location. This follows from the standard triangle inequality on transportation costs, which necessarily holds here because transportation costs are monotonic functions of distance.

region r° to any destination location in region r'. The variable $\bar{\phi}_{r',r^\circ}$ is obtained by aggregating up the location-level shares ϕ_{ℓ',ℓ° as follows:

$$\bar{\phi}_{r',r^\circ} = \sum_{\ell'\in\Lambda_{r'}} \frac{x_{\ell'}}{\bar{x}_{r'}} \left(\sum_{\ell^\circ\in\Lambda_{r^\circ}} \phi_{\ell',\ell^\circ} \right), \tag{7}$$

where $\bar{x}_r$ is spending at region r aggregated across all locations ℓ in in the region, i.e., $\bar{x}_r \equiv \sum_{\ell\in\Lambda_r} x_\ell$. We provide an approximation for (7) constructed with region-level variables that is valid when internal distances within a region are small relative to external distances across regions. This is a sensible assumption in our empirical context.

The first step in our aggregation procedure is to define the following measure of internal distance within region r:

$$\hat{d}_{r,r} \equiv \sum_{\ell'\in\Lambda_r} \sum_{\ell^\circ\in\Lambda_r} \frac{\gamma_{\ell^\circ}}{\bar{\gamma}_r} \frac{x_{\ell'}}{\bar{x}_r} d_{\ell'\ell^\circ}, \tag{8}$$

where $\bar{\gamma}_r$ aggregates location-level productivity indices (2) to the region level,

$$\bar{\gamma}_r \equiv \sum_{\ell\in\Lambda_\ell} \gamma_\ell.$$

Suppose that the within-region spending share $x_\ell/\bar{x}_r$ of location ℓ and the within-region productivity share $\gamma_\ell/\bar{\gamma}_r$ both equal the within-region population share of location ℓ. In this case, the distance measure (8) can be interpreted as the expected distance between two randomly selected individuals living in the region. This is how we actually estimate $\hat{d}_{r,r}$ below. This is an intuitive and easy-to-calculate measure of distance associated with trade within a region.

Next we define $\hat{a}_{r,r} \equiv a(\hat{d}_{r,r})$ to be the within-region distance adjustment in a region-level model with internal distance $\hat{d}_{r,r}$. Also, let $\hat{a}_{r',r^\circ} = a(\bar{d}_{r',r^\circ})$ for $r^\circ \neq r'$ be the cross-region distance adjustments, where $\bar{d}_{r',r^\circ}$ is defined as the distance from region r' to region r°.[4] When we go to the data, we use the following region-level version of BEJK to model shipment flows from region r° to r', which takes the same form as the location-level equation (1):

$$\hat{\phi}_{r',r^\circ} \equiv \frac{\hat{a}_{r'r^\circ}\bar{\gamma}_{r^\circ}}{\sum_{r=1}^{R} \hat{a}_{r',r}\bar{\gamma}_r}. \tag{9}$$

[4] For expositional simplicity, we make the approximation that distances between locations from different regions depend on the pair of regions, not the particular locations within the regions.

If all distances within a region are zero, then this is equivalent to there being a single location in each region, and $\hat{\phi}_{r',r^\circ} = \bar{\phi}_{r',r^\circ}$ must hold. In the separate online Appendix, we use straightforward calculus arguments to show that if distances within a region are small, relative to distances across regions, then $\hat{\phi}_{r',r^\circ}$ is a first-order approximation to $\bar{\phi}_{r',r^\circ}$. We use $\hat{\phi}_{r',r^\circ}$ to estimate the model.

3.3 Aggregating Product Segments to Industries

An *industry* is an aggregation of a set of product segments by a statistical agency. For example, consider a list of product segments including "standard passenger cars and trucks," "limousines," "ambulances," "hearses," and so on. Under the 1987 Census Standard Industrial Classification System, these segments (and others such as "cars, armored," and "fire department vehicles") are grouped together into SIC 3711, "Motor Vehicles and Passenger Car Bodies."

Henceforth, we associate each segment with an index i of the industry containing the segment, in addition to its segment index k. Superscripts denote the segment index and subscripts the industry index. For example, β_i^k is the spending share of segment k of industry i. Let $K_i + 1$ be the total number of different segments included in industry i. The total spending share of industry i across all segments is then

$$\beta_i^{Ind} \equiv \sum_{k=1}^{K_i+1} \beta_i^k.$$

The structure developed so far is general. We add content by placing additional restrictions. First, we assume there is one *primary segment*, segment $k = 1$, that accounts for the majority of expenditure in the industry, and we label segments $k = 2$ through $K_i + 1$ as the *specialty segments*. (Thus, K_i equals the number of different specialty segments in industry i.) That is, we assume

$$\beta_i^P \geq \beta_i^S, \tag{10}$$

where

$$\begin{aligned} \beta_i^P &\equiv \beta_i^1 \\ \beta_i^S &\equiv \sum_{k=2}^{K_i+1} \beta_i^k. \end{aligned}$$

To motivate assumption (10), consider the automobile example. For most day-to-day transportation needs, consumers generally use standard passenger cars and trucks. But on rare occasions, consumers have need for a specialty vehicle: a limousine on wedding days, an ambulance during emergencies, a hearse at funerals. This pattern is consistent with assumption (10). The assumption is also consistent with the so-called Pareto principle, or the 80-20 rule, in which it has been observed in a variety of applications that things tend to be skewed so that "the vital few" have large market shares and "the trivial many" have small market shares (Juran (1954)). In particular, there is a rule of thumb where 80 percent of sales come from 20 percent of the products.

Besides spending shares, we explore two different ways in which the specialty segments may differ from the primary segment. In the first, we treat specialty segments as having *high transportation costs*. In the second, we treat specialty segments as *niche with idiosyncratic sources of supply.*

3.3.1 Specialty Segments with High Transportation Costs

In this approach, the BEJK model parameters of a given specialty segment are identical to the parameters of the primary segment in the same industry, with two exceptions. First, as already assumed above, the spending share is less, i.e., $\beta_i^k < \beta_i^1$. Second, transportation costs are higher for specialty goods. Formally, for each industry i,

$$\tau_i^k(d) > \tau_i^1(d)\text{, for } d > 0 \text{ and } k \geq 2. \tag{11}$$

Assumption (11) on relative transportation costs is key, and we motivate it as follows. The provision of specialty goods is often facilitated by face-to-face interactions between producers and consumers. These interactions are cheaper to accomplish when the producer and consumer are near each other. Almost by definition, specialty goods are intended to meet a special need. Direct interaction increases the probability of getting it right. Furthermore, local sourcing is relatively more valuable in making turnaround time faster for custom goods relative to standardized goods. To see why, note that standardized goods can be held in inventory for quick local delivery, even if the goods themselves have been imported from a distant location. In contrast, custom goods cannot be held in inventory.

We first show that average plant size (measured by sales) is smaller for specialty segments compared with the primary segment. Exposition of this result is simplified if we assume internal distance at the *location level* is zero, i.e., $d_{\ell,\ell} = 0$, which implies $a_{i,\ell,\ell}^k = 1$. (Note the setup allows for positive internal distances at the *region* level.) Then, using (1), and the fact that (11) implies $a^k_{i,\ell',\ell} < a^1_{i,\ell',\ell}$ for $k \geq 2$ and $\ell' \neq \ell$, we obtain the following:

$$\phi^k_{i,\ell,\ell} = \frac{\gamma_{i,\ell}}{\gamma_{i,\ell} + \sum_{\ell' \neq \ell} \gamma_{i,\ell'} a^k_{\ell',\ell}} > \frac{\gamma_{i,\ell}}{\gamma_{i,\ell} + \sum_{\ell' \neq \ell} \gamma_{i,\ell'} a^1_{\ell',\ell}} = \phi^1_{i,\ell,\ell}. \tag{12}$$

Inequality (12) says that the probability a good at ℓ is obtained from a local producer (i.e., at ℓ) is higher for a specialty segment than for the primary segment. Intuitively, higher transportation costs of specialty goods make local producers more likely to be competitive against imports. Inequality (12) implies that overall plant counts will be higher for specialty segments. Since overall spending is less (assumption (10)), average sales per plant must be lower for specialty plants.

Now let transport costs in specialty segments differ from the primary segment by a proportionate factor $\rho > 1$, i.e., $\tau^k_i(d) = \rho\tau^1_i\,(d)$, $d > 0$, and consider the limit as ρ gets large. Assuming $\gamma_{i,\ell} > 0$, we can see from (12) that the probability a specialty segment good in industry i is locally sourced goes to one for large ρ,

$$\lim_{\rho \to \infty} \phi^k_{i,\ell,\ell} = 1. \tag{13}$$

As explained earlier, when we go to the data, we will be aggregating locations into regions. Equation (13) implies that when transportation costs for specialty segments are high (large ρ), the number of specialty plants is approximately proportionate to the number of locations within the region (since each location is approximately self-sufficient in each specialty segment). For intuition, consider the example of the New York City region mentioned earlier, which is spread out across parts of four states. Retail establishments, such as furniture stores, are of course spread out throughout the region to maintain proximity to customers. The same basic economic force is at work with specialty manufacturing plants that offer a retail-like function, such as craft shops making custom furniture and kitchen cabinets. These can be found throughout various locations in the four states that make up the region.

3.3.2 Specialty Segments as Niche with Idiosyncratic Sources of Supply

Clearly, many specialty goods are traded across long distances and have transportation costs that are immaterial and outweighed by the importance of product quality. Take the example of hearses noted above. Relatively few regions in the United States produce their own hearses. A large percentage of them are produced in Ohio and then distributed throughout the country. Limousines are also a tradable good, for which Los Angeles is a significant source. Other niche segments in the motor vehicle industry can be found in other regions.

This subsection makes the point that while any *particular* niche segment might be highly geographically concentrated, when we combine different niche segments into the same industry, the aggregate totals will tend to be spread out and be less geographically concentrated than any one niche in isolation. Different niches will have different input requirements and different technologies, a property we refer to as *idiosyncratic sources of supply.* While one niche might be best located in one region, a second niche might be best located in a second region. Hence, things look more spread out when we combine the two niches together.

Recall our earlier discussion of the Pareto principle, or the rule of thumb in which 20 percent of the products ("the few") get 80 percent of the sales, and 80 percent of the products ("the many") get 20 percent of the sales. We associate the primary segment with "the few." With relatively few products, the primary segment will tend to be geographically concentrated. In contrast, as we combine the many different types of niche goods in specialty segments with idiosyncratic sources of supply, the aggregate totals of the combined specialty segments will tend to be spread out, at least relative to the primary segment.

If all regions are the same size, it is simple to write down a model capturing these ideas. Think of the region-level cost efficiency parameter $\bar{\gamma}^k_{i,r}$ for each specialty segment $k \geq 2$ in industry i and region r as a random variable, and assume it is drawn from some distribution that is i.i.d. across regions and segments for a given industry i. Suppose transportation costs are zero for each specialty segment. (We noted above that some specialty goods have transportation costs that are immaterial; for simplicity, just set them to zero.) Then from (9), for any particular specialty segment k in industry i, region r's share of sales (and plant counts) is

$$share^k_{i,r} = \frac{\bar{\gamma}^k_{i,r}}{\sum_{r'=1}^{R} \bar{\gamma}^k_{i,r'}}.$$

To calculate region r's share across all K_i specialty segments together in industry i, we take the average:

$$share^S_{i,r} = \frac{1}{K_i} \sum_{k=2}^{K_i+1} \left(\frac{\bar{\gamma}^k_{i,r}}{\sum_{r'=1}^{R} \bar{\gamma}^k_{i,r'}} \right).$$

Given the i.i.d. draws of the $\bar{\gamma}^k_{i,r}$, by standard law of large numbers arguments, the specialty share $share^S_{i,r}$ will be close to $\frac{1}{R}$ when the number of different segments K_i in industry i is large. That is, specialty segments as a group will tend to spread out equally across locations.

When regions have different sizes, some with larger populations than others, there is a complication. If specialty production on average spreads out equally across locations, it is inconsistent with why some regions are bigger than others in the first place. To address this issue, suppose the cost efficiency vector $\Gamma_i^k = (\bar{\gamma}_{i,1}^k, \bar{\gamma}_{i,2}^k, ...\bar{\gamma}_{i,R}^k)$ for segment k of industry i specifying the cost efficiency index at each of the R locations is drawn i.i.d. from a distribution satisfying

$$E\left(\frac{\bar{\gamma}_{i,r}^k}{\sum_{r'=1}^R \bar{\gamma}_{i,r'}^k}\right) = popshr_r \tag{14}$$

for each region r, where $popshr_r$ is region r's share of the total population. Appealing again to a law of large numbers argument, specialty segments will tend to spread out across regions in proportion to population, when the variety K_i of different specialty segments in industry i is large.

3.4 The Number of Specialty Plants and Population

In the first specialty segment model—high transportation costs—the number of specialty plants in a region is proportional to the number of locations in the region. In the second model—niche with idiosyncratic sources of supply—the number of specialty plants is proportional to population. Suppose the number of locations in a region is proportional to population. Then both specialty segment models have the same reduced-form relationship between the plant counts in region r and population share:

$$n_{i,r}^S \approx \lambda_i^S \times popshr_r. \tag{15}$$

Below we argue that a proportionate relationship across regions between location counts and population is a reasonable assumption in our application.

4 The Data and Some Descriptive Results

The first part of this section discusses data sources and industry and geographic definitions. The second part presents evidence from a selected set of industries for which we have a specialty proxy.

4.1 The Data

We analyze the confidential micro data from two programs of the U.S. Census Bureau. The first, the 1997 *Census of Manufactures* (CM), contains information about plant employment, sales revenue, location, and industry classification.

The second, the 1997 *Commodity Flow Survey* (CFS), is a survey of the shipments that leave manufacturing plants.[5] Respondents are required to take a sample of their shipments (e.g., every tenth shipment) and specify the destination, the product classification, the weight, and the value of each sampled shipment. On the basis of this probability-weighted survey, the Census tabulates estimates of figures such as the total ton-miles shipped of particular products. There are approximately 30,000 manufacturing plants with shipments in the 1997 survey.

While we have access to the raw confidential Census data, in some instances we report estimates based partially on publicly disclosed information rather than entirely on the confidential data. These are cases in which we want to report information about narrowly defined geographic areas, but strict procedures relating to the disclosure process for the micro-data-based results get in our way. In these cases, we make use of the detailed public information that is made available about each plant in the Census of Manufactures. Specifically, the Census publishes the cell counts in such a way that for each plant, we can identify its six-digit NAICS industry, its location, and its detailed employment size class (e.g., 1–4 employees, 5–9 employees, 10–19 employees, etc.). We use this and other information to derive estimates for narrowly defined geographic areas. The data appendix (Appendix A) provides details.

The data analog of a region in the model is an Economic Area (EA) as defined by the Bureau of Economic Analysis (BEA).[6] The BEA begins with counties as building blocks and aggregates to the level of an EA, with the aim of creating meaningful economic units. There are 3,110 counties in the contiguous United States (we exclude Alaska and Hawaii throughout the analysis), and these are combined into 177 EAs. As an example, the New York City economic area contains 36 counties. Henceforth, when we refer a region we mean an EA.

To estimate our region-level internal distance given by equation (8), we use Census population data disaggregated to the Census tract level to determine the distribution of population within a region. As there are approximately 65,000 Census tracts, this approach allows for a fine level of geographic resolution relative to the region level. As explained earlier, we use the expected distance between two randomly selected individuals within the region as our estimate of the internal distance (8) of a region. The distance between any pair of different regions is set equal to the distance between the population centroids of each region.

With our region definitions in place, we examine the relationship between counts of

[5] Hillberry and Hummels (2003, 2008) are the first economics papers to use this confidential micro data. See U.S. Bureau of the Census (1999) for public tabulations.

[6] We use the 2004 definition; see Johnson and Kort (2004).

locations in a region and region population. As our measure of counts of locations, we use square miles of land area within each region above a population density threshold. Much of the land area in the United States is in very rural areas, and we want to exclude these. Specifically, we define a Census tract as occupied if there are at least 50 people per square mile. Occupied land, according to this definition, accounts for 91 percent of the overall United States population, but only 17 percent of the total land area. If we regress occupied land area on population (in logs) across the 177 regions, the slope is .90 ($s.e. = .04$, $R^2 = .79$). We take this as evidence of a roughly proportional relationship, consistent with our use of (15).

4.2 Evidence for the Theory from a Proxy for Specialty Plants

This subsection presents descriptive evidence based on seven industries for which we have a proxy variable for specialty plants. In 1997, the Census changed industry classification from the SIC system to the NAICS system. Seven NAICS manufacturing industries were redefined to include plants that had previously been classified as retail under SIC. For example, under the SIC system, establishments that manufactured chocolate on the premises for direct sale to consumers were classified as retail. Think here of a fancy chocolate shop making premium chocolate by hand. These were moved into NAICS 311330, "Confectionery Manufacturing from Purchased Chocolate." This industry also includes candy bar factories with more than a thousand employees. This situation—in which retail candy operations are lumped into the same industry as mass-production factories making standardized goods—epitomizes what we are trying to capture in our model. Analogous to chocolate, facilities making custom furniture and custom curtains in storefront settings were moved from retail under SIC to manufacturing under NAICS.[7]

Table 2 shows the seven NAICS manufacturing industries affected this way. We will refer to these industries as the *1997 Reclassification Industries.* All are consumer goods industries, producing some kind of candy, textiles, or furniture. For the 1997 data, we have both the SIC and NAICS variables for each plant. We refer to the plants that are in retail under SIC as "R" plants and the remaining plants as "M." We expect the R plants to fit well within our notion of a specialty plant. Note, however, that the M plants potentially also include specialty plants as well, and we present evidence consistent with this below. Therefore, we view the variable as being a useful proxy, rather than as providing an exact partition into

[7] The logic underlying these reclassifications was an attempt under the NAICS system to use a "production-oriented economic concept" (Office of Management and Budget, 1994) as the basis of industry classification.

primary and specialty plants.

We now return to the key facts listed in Section 2 and examine the reclassification industries in the context of these key facts and the theory. The first basic fact in Section 2 is that there is wide variation in plant size even when we go to the 6-digit NAICS classification level. In the theory, we show that specialty plants are smaller than primary plants in the same industry, which is consistent with the first basic fact. Table 2 presents evidence on this implication. R plants, our proxy for specialty, are on the order of one-tenth the size of the M plants in the same 6-digit NAICS.

The second basic fact in Section 2 is that the size distribution skews right, with a high percentage of very small plants. In the theory, specialty plants follow the population distribution of regions. As there are many regions, there will be many specialty plants, and as specialty plants are small, the theory is consistent with the second basic fact. Table 2 presents direct evidence on this implication. In the candy and textile industries, the R plants are approximately half of the total number of plants. In the three furniture industries, the R plants are about a quarter of the total.

The third basic fact in Section 2 is that small plants tend to ship locally. In the theory, specialty plants tend to ship locally compared with primary plants in the version of the model in which specialty plants have higher transportation costs. Here we offer two pieces of direct evidence on this implication. The first is simply that the R plants were classified under retail under the original SIC system. Of course the term "retail" is associated with selling to local consumers. Thus, the R classification itself has information content regarding shipping distance. Second, for each of the seven industries, the export share of R plants is zero to two decimal places. In contrast, the M plants in each industry have positive export shares on average.

The fourth basic fact in Section 2 is that small plants are diffuse, while large plants are geographically concentrated. In the theory, for both specialty-segment models we derived the reduced-form relationship (15) in which the level of specialty activity at a region varies proportionally with region population. Again, as specialty plants tend to be small, the theory is consistent with the basic fact. Table 2 provides direct evidence that specialty plants follow the distribution of population, as the mean location quotient defined in Section 2 is close to one in all the R segments. It is significantly greater than one in all the M segments, meaning that these plants tend to be geographically concentrated. For example, in the wood furniture industry, the mean is 1.17 for R and 4.42 for M.

Up to this point, we have used Table 2 to contrast the R and M segments within a 6-digit NAICS industry. There is also an interesting contrast in Table 2 across the wood furniture and kitchen cabinet NAICS industries. Both industries make cabinets, but kitchen cabinets are more likely to be custom made, because they are built into the wall and have to fit a particular spot in a kitchen. We redo our above discussion, only this time compare the M segment of kitchen cabinets, which we now treat as a specialty segment, with the M segment of wood furniture, the primary segment. For the specialty segment versus the primary, average plant size is smaller (15 versus 41 employees), plant count is higher (5,908 versus 3,035), shipments are more local (export share of .01 versus .03), and the plants tend to be geographically diffuse (a mean location quotient of 2.14 versus 4.42). Our paper focuses on the distinction between special and primary segments within narrowly defined six-digit NAICS industries. Naturally, the same ideas extend to broader groupings of industries. If we were to define the three-digit NAICS 337, "Furniture and Related Products Manufacturing" as one industry, the entire six-digit NAICS kitchen cabinet industry can potentially be thought of as a specialty segment within the broader furniture industry.

As noted above, even within the M segment, we expect there to be some specialty plants. We use unique information available for the furniture industry to provide support for this claim. In the data on product-level shipments, a distinction can be made between custom-made and stock furniture pieces. For furniture plants in the sample, let the variable *custom share* be the share of sales receipts made up of custom goods. (Appendix A.3 presents details.) Table 3 shows how custom share varies by R or M classification, by separating out kitchen cabinets from household furniture, and by plant size within the M segment. Before getting to the main point, we note two other clear facts in the table. First, R plants have significantly higher custom rates than M plants, consistent with our hypothesis that they are undertaking a specialty function. Second, kitchen cabinets plants have remarkably high custom rates, consistent with our previous discussion. Turning to the main fact, note that within the M segment and holding industry fixed, smaller plants have significantly higher custom rates. This result is consistent with our claim that small plants within the M segment itself also tend to undertake a specialty function.

We conclude this section by revisiting Table 1, where we traced out the plant size/geographic concentration relationship. In the bottom panel, we tabulate the relationship for the seven reclassification industries. Our result about the mean location quotient is even sharper than before. There is a jump from 2.46 in the bottom size category to 5.83 at the top (with NAICS fixed effects), compared with a jump of 3.68 to 5.69 in the whole sample. The seven industries considered here are a selected sample for which we can be highly confident that the issues we have raised are important. However, the key takeaway point of Table 1 is that the pattern found for these seven selected industries is just a slightly sharper version of a general pattern found throughout the manufacturing sector. As we continue the empirical analysis, we will carry along the distinction between these seven selected industries and the remaining industries. And we will continue to see that the underlying patterns are similar.

5 Estimation of the Model

In this section we estimate the model and in the next section analyze the estimates. We have several motivations for estimating the model rather than limiting ourselves to examining how the qualitative patterns in the data relate to the implications of the theory. First, our paper highlights the role of specialty plants. Through our approach, we are able to produce an estimate of the share of specialty plants in overall industry plant counts and thereby shed light on the quantitative relevance of the theory. We will point to the features of the data that pin down this estimate. Second, as we will see, an industry model with only a primary segment is consistent at a qualitative level with the fourth basic fact in Section 2, the plant size/geographic concentration relationship. Whether it succeeds at a quantitative level depends on model parameters like transportation frictions, and these can be obtained through estimation of the model. Third, while implications of the theory regarding the effects of imports are amenable to qualitative data analysis, estimation of the model allows for a richer analysis. In particular, there are additional considerations such as regression-to-the mean phenomena and population change. With an explicit quantitative model, these additional considerations can be brought in. Fourth, the theory implies that industries being hit by imports will experience a decline in primary-segment count share. As the primary count share is not directly observed in the data, we need to estimate the model to examine this implication.

In the theory, while specialty plants may account for a relatively large share of plant counts, they account for only a relatively small share of the value of industry shipments (condition (10)). This discrepancy between the specialty share in counts and share in expenditure motivates our choice of a two-stage estimation procedure. In the first stage, we use information on the value of industry shipments to estimate parameters of the primary segment. In the second stage, we use the information on industry plant counts, and the first-stage results, to estimate the plant-count model parameters for the primary and specialty segments.

5.1 First Stage: Distance Adjustment in the Primary Segment

Let $\Gamma_i^P = (\gamma_{i,1}^P, \gamma_{i,2}^P, ..., \gamma_{i,R}^P)$ be the vector containing the reduced-form (2) cost efficiency of each region r in the primary sector of industry i.[8] Normalize so the $\gamma_{i,r}^P$ sum to one across regions r. Let A_i^P be the $R \times R$ matrix of the reduced-form distance adjustments in the primary segment of industry i, with elements $a_{i,r'r^\circ}^P$ given in (3). In this subsection we explain our strategy for producing estimates of Γ_i^P and A_i^P.

We begin by parameterizing how the distance adjustments $a_{i,r'r^\circ}^P$ vary with distance $d_{r'r^\circ}$. We consider two cases. The first is the standard log-log specification typically used in the trade literature:

$$\ln a_{i,r'r^\circ}^P = -\eta_i^{\text{loglog}} \ln d_{r'r^\circ},$$

which yields a constant distance elasticity η_i^{loglog}. The second is a semi-log specification,

$$\ln a_{i,r'r^\circ}^P = -\eta_i^{\text{semi,1}} d_{r'r^\circ} - \eta_i^{\text{semi,2}} \left(d_{r'r^\circ}\right)^2,$$

in which the distance elasticity can vary.[9] If the coefficient on the squared term is zero, $\eta_i^{\text{semi,2}} = 0$, then the coefficient $\eta_i^{\text{semi,1}}$ can be interpreted as a constant decay rate per unit distance for industry i.

There is enormous variation in population across regions in the data. The most populous region, the New York City region, is 255 times larger in population than the smallest, the region centered in Aberdeen, South Dakota. We use population to approximate differences in demand across regions, i.e., a region's spending share is approximated by its population share. Thus, for a given industry, demand in the New York City region will be 255 times larger than the demand in Aberdeen. We expect this to be a good first approximation for final good industries and for certain intermediate good industries where downstream production follows the distribution of population (such as intermediate goods for the construction industry). In contrast, for intermediate goods to the manufacturing industry, population is less useful as a proxy for local demand (since upstream manufacturing industries don't necessarily follow the distribution of population). Nevertheless, for completeness, we apply our procedure to the entire set of industries. Out of the 473 different industries, there are missing data or disclosure issues for 7 industries, leaving us with estimates for $466 = 473 - 7$ industries. We

[8] Henceforth, the geographic units will always be at the region level, and we will no longer use an overbar to distinguish region-level from location-level variables.

[9] In both specifications, we follow Anderson and van Wincoop (2003) and the references therein by allowing for an internal distance within each region, which we directly calculate as explained earlier. We note there is an alternative approach in the literature that leaves the internal distances in each region as free parameters that can be estimated with region-level dummy variables.

have also considered a restricted sample of 175 industries for which we argue that population is a particularly useful proxy for local demand. (See Appendix A.2.) Our results with the narrow sample are very similar to the results with the full sample.

As already noted, we expect the spending shares on specialty segments to be small. In this section, we estimate the model for the primary segment under the approximation that the spending share of specialty goods is close to zero. In this case, we can take the distribution of the value of shipments at the industry level across regions as an approximation to the distribution of the value of primary segment shipments across regions. The Census of Manufactures (CM) covers the universe of all plants in the United States. Subtracting out the exports of each plant, we can aggregate the plant-level domestic sales revenue data to approximate $y_{i,r}^{P}$, the share of primary segment domestic sales revenue in industry i originating from plants in region r.

Other than export information, there is no destination information in the CM. However, the CFS provides survey information on shipments and their destinations. A concern we have with the CFS data is that local shipments may be overrepresented in the data. These seem too high, in the sense that more is shipped locally than can be absorbed by local demand. We expect that sometimes shipments intended for faraway destinations get there by way of a local warehouse. In cases like these, the destination found in the CFS may be the local warehouse rather than the ultimate destination. The separate online Appendix discusses this issue further and provides some evidence on the importance of wholesaling for the manufacturing industries in our sample.

The form of our data leads us to the following strategy for estimating the cost efficiency vector Γ_i^P and the distance adjustment parameter η_i^P (where $\eta_i^P = \eta_i^{\text{loglog}}$ in the log-log case and $\eta_i^P = (\eta_i^{\text{semi},1}, \eta_i^{\text{semi},2})$ in the semi-log case). We pick (η_i^P, Γ_i^P) to perfectly match the domestic sales revenue distribution y_r^P across originating regions, as we directly observe the universe of sales at each region. Because of our concern about excessive local shipments in the CFS, we throw out all local shipments in the CFS that are less than 100 miles and fit the conditional distribution of the longer shipments. Formally, set $\overline{dist} = 100$ and let $B(r, \overline{dist})$ be the set of all destinations at least $\overline{dist}$ from an originating region r. The conditional probability that a primary segment industry i shipment originating in region r° goes to a particular destination $r' \in B(r^{\circ}, \overline{dist})$ equals

$$p_{i,r'r^{\circ}}^{P} = \frac{y_{i,r'r^{\circ}}^{P}}{\sum_{r \in B(r^{\circ}, \overline{dist})} y_{i,rr^{\circ}}^{P}},$$

which can be calculated in the model through equation (4). For each value of η_i^P, we solve for the vector Γ_i^P such that the predicted total sales of the industry at a given region equals total sales in the CM data. The supplementary online Appendix outlines our algorithm for finding a solution $\Gamma_i^P(\eta_i^P)$ to the 177 nonlinear equations for the 177 regions. (The approach is analogous to the inversion in Berry (1994).) We can then write the conditional probability above as a function of η_i^P. We pick η_i^P to maximize the conditional likelihood of the destinations observed in the shipment sample.[10]

We begin by discussing our estimates for the constant elasticity case (log-log). Industries at the extremes provide a useful illustration of the results. The industries with the four highest elasticities are "Ready-mix concrete," "Ice," "Asphalt," and "Concrete block," where the elasticities are, respectively, 4.2, 3.0, 2.9, and 2.8. These industries clearly have extremely high transportation costs, and shipments tend to be very local in these industries.[11] At the other extreme, there are 29 industries in which the distance elasticity is estimated to equal zero, the lower bound.[12] These include industries like "Semiconductors," "Analytical laboratory instruments," and "Aircraft," where transportation costs are obviously low relative to value. Table 4 reports the elasticity estimates for the seven reclassification industries, which are intermediate between the extremes just discussed, ranging from 0.31 for "Chocolate Candy" to 1.19 for "Upholstered Household Furniture." Table 5 reports summary statistics for the 466 out of 473 total NAICS 6-digit industries for which it was possible to estimate the model.[13] (Detailed estimates by industry are provided in the supplementary online Appendix.) The mean elasticity is .61, and the 25th, 50th, and 75th percent quartiles are .27, .53, and .82. We also distinguish between diffuse and nondiffuse demand industries. The mean elasticity for diffuse demand is slightly higher than for nondiffuse demand, .67 versus .58. We note that the standard errors for the elasticity estimate are relatively small, with a mean value of .03. (These are reported in the supplementary online Appendix.)

We have also estimated the model using the data from the 1992 Census of Manufactures, and the results are reassuringly very similar. The mean value for 1992 is .67 compared with the mean of .61 in 1997. For comparable industries between the two years, the elasticity

[10] The sample of plants selected for the CFS is stratified. We use the establishment sampling weights to reweight the cell count realizations and follow a pseudo-maximum likelihood approach. In writing down the likelihood, we condition on the origination of a given shipment.

[11] For these industries, we actually include shipments within 100 miles in the estimation because shipments beyond 100 miles are relatively rare. The supplementary online Appendix provides the details about which industries were treated this way.

[12] In the estimation, we impose $\eta_i^{\text{loglog}} \geq 0$.

[13] The $7 = 473 - 466$ industries without estimates are industries for which there is either no shipment information in the CFS or the sample size is too small to meet disclosure requirements.

estimates in the two years are very close.[14] For example, for the four highest elasticity industries listed above, with 1997 estimates of 4.2, 3.0, 2.9, and 2.8, the corresponding 1992 estimates are 4.0, 3.2, 2.5, and 2.9.

We turn now to the semi-log case. To facilitate comparison with the constant elasticity case, Tables 4 and 5 report the implied elasticities at 100 and 500 miles, rather than the coefficients themselves. There is a consistent pattern that the elasticity at 500 miles is almost five times as large as the elasticity at 100 miles. The estimate for the constant elasticity case lies between these two values. For the semi-log case, the percentage effect on the distance adjustment, of a *unit* increase in distance, tends to be roughly constant as a function of distance. This accounts for why the elasticity is roughly proportional to distance. In 427 cases the semi-log specification has a higher likelihood value than the log-log version, and the comparison is reversed in the remaining 39 cases. We can see in Table 4 that for the reclassification industries, the semi-log fits better in each case. For each industry, we use the specification with the highest likelihood. Appendix A.4 provides a discussion of goodness of fit. In particular, we show that the fitted values of the model do a good job of accounting for cross-industry variation in the distance of shipments.

5.2 Second Stage: Primary and Specialty Plant Counts

This section explains how we estimate model parameters determining plant counts in each industry. We begin by outlining how the procedure works in the limiting case where the specialty segment share of industry revenue is zero. Next we explain how we allow for a positive specialty-segment revenue share.

Consider first the limiting case where specialty-segment revenue share is zero in a particular industry. From the first stage, we have an estimate of the cost efficiency vector Γ_i^P and the distance adjustment η_i^P parameters for the primary segment of industry i. We can plug Γ_i^P and the implied distance adjustment A_i^P into equation (1) to obtain the probability $\phi_{i,rr}^P$ that region r sells to itself in the primary segment of industry i. Then with the plant count scaling coefficient λ_i^P for the primary segment, we obtain primary segment plant counts at r from (6), equal to $n_{i,r}^P = \lambda_i^P \phi_{i,rr}^P(\Gamma_i^P, \eta_i^P)$. Next, we use specification (15) for specialty plant counts in which $n_{i,r}^S = \lambda_i^S x_r$ approximately holds, where x_r is population share. Total plant counts at region r equal primary plant counts plus specialty plant counts,

[14] Industry definition changed from SIC for 1992 to NAICS for 1997. There are 316 6-digit NAICS industries that have an exact 4-digit SIC equivalent in 1992. Weighting by the number of shipment observations used, the correlation in the elasticity estimates for 1992 and 1997 is .98.

$$n_{i,r} = \lambda_i^P \phi_{i,rr}^P(\Gamma_i^P, \eta_i^P) + \lambda_i^S x_r. \tag{16}$$

To take (16) to the data, we introduce an error term. Suppose the observed total number of plants in the given industry at region r equals the above expression plus an error term $\xi_i + \varepsilon_{i,r}$,

$$\tilde{n}_{i,r} = \lambda_i^P \phi_{i,rr}^P(\Gamma_i^P, \eta_i^P) + \lambda_i^S x_r + \xi_i + \varepsilon_{i,r}, \tag{17}$$

where the error term has variance proportional to region r's population x_r. We use weighted least squares to construct estimates of the slopes λ_i^P and λ_i^S and the constant ξ_i for each industry. (Given the results of the first stage, $\phi_{i,rr}^P(\Gamma_i^P, A_i^P)$ are data for industry i at this point.)

We modify the above procedure to allow for positive specialty-segment revenues. Take as given a value Rev_i^S of specialty revenue per plant for industry i. Use Rev_i^S, along with the estimate λ_i^S from above, to construct an estimate of specialty-segment revenues at region r,

$$\hat{y}_{i,r}^S = Rev_i^S \lambda_i^S x_r,$$

and from this construct an estimate of primary segment revenue,

$$\hat{y}_{i,r}^P = \max\{y_{i,r} - \hat{y}_{i,r}^S, 0\}.$$

Next go back and solve for a new primary segment cost efficiency vector $\Gamma_i^{P\prime}$ that exactly fits the new estimate of the primary-segment sales distribution $\hat{y}_{i,r}^P$ across regions r. Using the new value of $\Gamma_i^{P\prime}$, run the weighted least squares regression above to produce new estimates $\hat{\lambda}_i^{P\prime}$ and $\hat{\lambda}_i^{S\prime}$ of the slopes. Iterate until convergence on estimates $\hat{\lambda}_i^P$ and $\hat{\lambda}_i^S$, the plant count coefficients for the primary and specialty segments. It remains for us to specify the choice of average specialty-plant sales revenue Rev_i^S. For each industry i, we set Rev_i^S equal to the average sales size of plants in the one to four employees category. We have experimented with alternative values for Rev_i^S, and it makes little difference for the estimates of $\hat{\lambda}_i^P$ and $\hat{\lambda}_i^S$.[15] While this iterative procedure takes account of positive specialty-segment revenues within the second stage, it holds fixed the first-stage estimated distance adjustment parameter vector η_i^P. In principle, after differencing out specialty revenues, we could have gone back to stage one to reestimate η_i^P and iterate this way. We did not do this because (1) it was desirable

[15] Doubling Rev_i^S relative to the baseline, or setting it to zero, makes virtually no difference in the results.

to complete the first stage with the secure confidential data and then do the second stage outside the Census with the publicly available data, and (2) the amount of revenues that would be differenced out is small, and so it would not be of much consequence.

Table 6 presents the results. The individual estimates are reported for the seven reclassification industries, and summary statistics are reported for broader sets of industries. We first note that allowing for the constant term ξ_i makes little difference; when we reestimate (17) without an intercept, we get similar results. Next note that the coefficient estimate $\hat{\lambda}_i^S$ for specialty goods tends to be quite large. Given the scaling that the population shares x_r sum to one, $\hat{\lambda}_i^S$ has an interpretation as an estimate of the total count of specialty plants in the industry.

There is no mechanical reason why the estimated coefficient $\hat{\lambda}_i^S$ in the regression (17) is necessarily positive. In fact, we can see in Table 6 that the minimum of $\hat{\lambda}_i^S$ over all 466 industries equals -1.4. Based on the interpretation of the scaling of $\hat{\lambda}_i^S$ just given, an estimate of -1.4 is negligible in absolute value. As the standard error of the estimate for this particular industry equals 2.2, it is not statistically significant from zero in any case. All together, there are six industries with estimates of $\hat{\lambda}_i^S$ that are just slightly less than zero, and none are statistically significant. For these six industries, our finding is that a standard model with a single primary segment fits the data. These industries include "Cellulosic organic fiber," "Primary aluminum production," and "Primary smelting of copper," which have, respectively, 6, 21, and 16 plants overall. It is not surprising that the primary-only model works well for industries of this nature and with so few plants.

Such industries are the rare exceptions to a general rule. Specialty plant counts are estimated to be quite high in the vast majority of industries. The last column of Table 6 reports statistics for the share of plants that are specialty.[16] In the 10th percentile industry, 35.4 percent of the plant counts are estimated to be specialty. The median is 68.0 percent, and the mean is 63.9 percent. The key property of the data that is driving this estimate is the fact that small plants tend to follow the distribution of population, while large plants are geographically concentrated. We will see in the next section that an industry model with only a primary segment is consistent with this qualitative pattern but cannot match the magnitude. In particular, given the estimated distance adjustments, the only way to rationalize the existence of such a large group of small plants following the distribution of population is for these plants to be in the specialty segment.

We can use Table 6 to compare the results for the seven reclassification industries with

[16] To construct this estimate, we set $\xi_i = 0$ and take the estimates of $\hat{\lambda}_i^P$ and $\hat{\lambda}_i^S$ to estimate fitted values of plant counts at each location and then aggregate up.

the results for the broader set of industries. For the reclassification industries, mean specialty count share is 75.6, about 12 percentage points higher than the overall mean of 63.9 percent. It is not surprising that the estimates for the reclassification industries are higher than overall, because it is a selected set of industries based on the existence of a particular proxy for specialty plants. More surprising is how small the difference is between the reclassification set and the rest. A large share of specialty plants within an industry is a pervasive feature throughout the manufacturing sector and is not limited to a narrow set of select industries.

Next we explore the connection between the estimated share of *specialty* plants in an industry and the share of *small* plants, where we define a small plant to have fewer than 20 employees. While we expect primary plants to be on the larger side, it is possible for a given primary plant to have fewer than 20 employees. And while we expect specialty plants to be on the smaller side, a given specialty plant might have more than 20 employees. Therefore, the specialty plant share and the small plant share are conceptually different objects (in addition to the fact that the first is estimated with the model and the second is read off the data). Still, we expect there to be a connection, and in fact, this is the case. For example, in the aluminum refining industry, in which it was noted above that the estimated specialty share is 0 percent, the small plant share is 0 percent. In the wood furniture industry, in which the estimated specialty share is 82.2 percent, the small plant share is 80.6 percent. If we regress the estimated specialty plant share on the small plant share for the 466 industries, the slope is .52 ($s.e = .04$, $R^2 = .23$).

6 Analysis of the Estimates

We use the estimated model to analyze two issues. The first is the plant size/geographic concentration relationship across regions. The second is the effect of the recent surge in imports from China on the distribution of plants across regions.

6.1 The Plant Size/Geographic Concentration Relationship

Recall the fourth fact from Section 2: small plants tend to be dispersed, following the distribution of population, while large plants concentrate near other plants in the same industry. Our full model—in which specialty segments are grouped with a primary segment in the same industry—accounts for this fact. Under either of the two ways we model them, specialty plants, which tend to be small, follow the distribution of population. In contrast, primary segment plants, which tend to be large, in general will not follow the distribution of population but instead will concentrate following Ricardian comparative advantage. Thus, there is a clear logic for how introducing specialty segments into an industry can account for the plant size/ geographic concentration relationship.

Do we need to introduce specialty segments? Might a model with only a primary segment be consistent with this relationship? Recall the parameter T_r^P governing region r-level productivity in the primary segment. Everything else the same, the larger T_r^P, the more the particular primary segment will concentrate at r. But how does this affect average plant size at r in the segment? There are two effects that work in opposite directions. On the one hand, plants that would have existed even before the increase are now more productive and therefore will be competitive at more distant regions. That is, existing plants will tend to become larger, pulling up average plant size. On the other hand, some plants at r that previously did not exist, because they were knocked out by lower-cost imports from other regions, are now competitive and will exist. The entry of these marginal plants will tend to pull down average plant size. In Appendix B we show that if transportation costs are zero, these two offsetting forces exactly counterbalance, and the net effect on average plant size of higher concentration is zero. We also show that in a two-location model with equal populations, if distance adjustments are positive, the net relationship between high concentration and average plant size is strictly positive.

In the estimated model from the previous section, we have estimates of distance adjustments, enabling us to quantitatively assess the success of a model with only primary segments in fitting the plant size/geographic concentration relationship. And we can compare this model with the success of the full model that includes specialty segments. We focus on regions with high concentrations of particular industries.

In particular, define a *high-concentration industry/region* to be one where the sales revenue location quotient is above 2 and where the region has at least 5 percent of the industry's revenues. Across the 7 reclassification industries, there are 23 different high-concentration industry/regions. These are listed in Table 7, sorted for each industry by descending sales revenue quotient. The table also reports a breakdown of the sales revenue quotient into a count and size quotient, $Q_r^{rev} = Q_r^{count} \times Q_r^{size}$. The count quotient is analogous to the revenue quotient, except it has a region's share of the national plant counts in the numerator instead of its share of revenue. The size quotient is the ratio of average plant size (in revenues) at the region to the national average plant size. It is clear from inspection of the data that the size margin plays an important role in how an industry expands at a region. Consider the wood furniture industry in the High Point region, where the revenue quotient is 27.7.[17]

[17] We use the abbreviated term "High Point" to refer to the BEA economic area containing the center

The breakdown is 27.7 = 4.2 × 6.6. Thus, average plant size in the region is 6.6 times the national average. A high contribution from the size margin holds for virtually all of the 23 individual industry/regions listed in Table 7. Across the 23 observations, the mean size quotient is 5.4, compared with a mean count quotient of 4.3.

The last two columns contain fitted values of the size quotient for the constrained model with only the primary segment, the *primary-only model*, and the *full model* that includes specialty segments.[18] As the discussion above of the theoretical results in Appendix B anticipated, the primary-only model fits the plant size/geographic concentration relationship at a *qualitative* level. In particular, it holds throughout all the estimated primary-only models in Table 7, with only three exceptions. However, it fails *quantitatively,* as the predicted size differences are much smaller than in the data. When we turn to the full model and allow for specialty segments, the predicted size differences are much closer to what they are in the data.

Next we consider summary statistics for the entire set of 466 industries, including a breakdown by diffuse and nondiffuse demand industries. We can see that the pattern we have just established for the reclassification industries continues to hold for the broad set of industries. The mean size quotient in high-concentration industry/regions in the broad set equals 4.4, indicating that in the data, the size margin plays a significant role throughout manufacturing in how regions specialize in an industry. However, in the primary-only model, the mean fitted value of the size quotient is only 1.2, i.e., the contribution of the size margin is relatively small, inconsistent with the size margin's large contribution in the data. In contrast, in the full model with specialty segments, the mean is 3.0, indicating a large role for the size margin, though still short of the mean value of 4.4 found in the data. There is skewness in the distribution of the size quotients, so it is of interest to also look at the median. The median fitted value of the size quotient in the full model equals 2.4, which is identical to the median of 2.4 in the data. The full model fits the plant size/geographic concentration relationship well; the constrained model, with a median size quotient of only 1.1, does not.

6.2 Effects of the China Surge

This section puts the model to work in examining the effects of imports. We take the estimated model for 1997 and simulate the effects of the surge in imports from China that

of the wood furniture industry in North Carolina. It consists of 22 counties, and its full name is the Greensboro–Winston-Salem–High Point, NC Economic Area.

[18] In the constrained model, plant counts are proportional to $\phi^{P}_{i,\ell'\ell^{\circ}}(\eta^{P}_{i},\Gamma^{P}_{i})$.

took place between 1997 and 2007. We examine the predicted geographic distribution of individual industries in 2007, and compare the predictions with what actually happened.

As of 1997, imports from China were already important for some industries. To focus on the effect of the *change* in imports over the period 1997 to 2007, we define *Existing China Imports* to equal 1997 imports from China, rescaled by the ratio between 2007 and 1997 of the value of domestic shipments plus Chinese imports. *New Imports* equal total imports from China in 2007, less existing imports. Define *New China Share* to equal the value of new imports in 2007, as a fraction of the sum of domestic shipments plus new imports. This approach effectively treats existing imports as being in a separate segment, in such a way that new imports and domestic production compete with each other, but not with existing imports.[19] We define the *New All-Country Share* by replacing "imports from China," in what we just described with "imports from all countries."

Table 8 partitions the 6-digit NAICS industries into 6 categories based on the new China share. In particular, there are 23 industries in which the share is greater than 50 percent; "House slippers," and "Infant cut-and-sew apparel" both have new China shares equal to 97 percent, an astonishing figure. There are another 23 industries in which the share is between 25 percent and 50 percent. Table 8 also reports the mean values of the new all-country share, revealing a very close connection to the new China share. For industries with a new China share in the range of 50 percent to 97 percent, the mean is 71 percent, while the mean new all-country value is 72 percent. For industries with a new China share in the range of 25 percent to 50 percent, the two means are 35 and 38. The correlation across all industries between the China and all-country figure is .90. The final column in Table 8 reports the mean industry employment growth over the period 1997 to 2007. In the two highest import categories, the mean employment changes were -75 percent and -51 percent, a remarkable decline over only a ten-year period. In the wood furniture industry example, the new China and new all-country shares are 32 percent and 38 percent, and employment fell from 128,000 to 63,000 over the period, a 51 percent decline.

6.2.1 What Happened in the Data

Before presenting predictions of various models, we examine what actually happened in the data. Panel A of Table 9 provides detailed cell counts for our example industry, wood furniture, for High Point and for the country as a whole. Note first the dominant position of High Point (shown in the last two columns) in 1997 in its share of large plants (the bottom

[19] This simplifies the 1997 baseline estimation, as it rationalizes the exclusion of existing imports.

two rows). For plants with 1,000 or more employees, High Point contained 5 out of the national total of 12 plants, and 9 out of the 36 in the next largest size category. In contrast, High Point contained only 51 out of 3,091 plants in the smallest size category. By 2007, a striking decline of the industry at High Point had taken place. The region no longer had *any* plants in the 1,000 and above employee size category, and the 500-999 category declined from 9 to 3 plants. Across all size categories, the plant count at High Point declined from 101 to 53, a very large percentage decline compared with the relatively small overall U.S. drop from 3,835 to 3,568 plants.

Panel B provides analogous plant counts for the broader set of 46 industries that came under the most pressure from China, the industries with new China shares above 25 percent. For each of these industries, we define the *primary region* to be the analog of what High Point is for wood furniture.[20] The key facts just noted about the wood furniture industry hold for the broader set of industries with a substantial new China share.

In Table 10, we look at the full set of high-concentration industry/regions considered earlier in Table 7, only now we break up the industries by new China share.[21] The earlier table reported that plants in high-concentration industry/regions on average are 4.4 times as large as the mean plants in their industries. In Table 10 we see this pattern holds across the different new China share groups. In the last column of Table 10, we calculate the percentage change in the count share at each high-concentration industry/region and take means within new China share groupings. (This is the same as the percentage change in the count quotient because the normalization cancels out.) For the highest group, in which the new China share was between 50 percent and 97 percent, the count share declined on average by 37 percent. In the next group, with a new China share between 25 percent and 50 percent, the count share declined on average by 20 percent. The average decline in these two top groups most affected by imports was significantly greater than the average decline of 6 percent to 10 percent in the remaining groups less affected by imports.[22] See Holmes (2011) for further discussion of the decline of large manufacturing plants in industries hit hard by imports.

[20] For each industry, it is the region with the highest revenue location quotient, among regions with at least 5 percent of sales (i.e., among high-concentration locations, as defined earlier).

[21] When we calculate the count quotient for 2007, we put the 2007 share of plant counts in the numerator and the 1997 share of population in the denominator. That is, we hold fixed the normalization.

[22] A test of the null hypothesis that the average of the highest group (50 to 97 percent new China share) is the same as the average within the combined group of 25 percent share and below yields a p-value of less than .0001. The p-value for group 2 (25 percent to 50 percent) being the same as 25 percent and below is .007.

6.2.2 Model Predictions

We model the surge in Chinese imports as arising from an exogenous shift in China's comparative advantage to sell primary segment goods in the United States. For the model of specialty segments as high transportation cost goods, the results we report follow directly from the logic of the theory. High transportation costs that impede domestic shipments of specialty goods obviously impede international shipments as well. For the second model of the specialty segments as niche with idiosyncratic supply, our simulations require an additional assumption that China has emerged as a source of primary segment varieties but not yet specialty segment varieties. Thinking of specialty goods as high-quality goods, it is natural to expect that China would enter on the bottom end with primary segment goods. The wood furniture coming from China is obviously more similar to the products of the large High Point factories than those of Amish craft shops. The way we model imports from China is very different from how we would model imports from the most advanced economies, which might very well fit our idea of specialty segment goods, e.g., machine tools made by the German Mittelstand or fashion goods from Milan.

Before moving to the quantitative analysis, we begin by discussing the qualitative implications of the theory. It is intuitive how the full model with specialty segments can account for the broad pattern in the data. Specialty plants are dispersed and small, and they are less affected by imports than primary plants, which are geographically concentrated and large. Thus, the analogs of High Point in wood furniture decline relative to the rest of the economy. Next consider a two-location, equal population version of the primary-only model, and let one location have a larger industry share than the other. In Appendix B we show that if a third source of supply emerges (China), the location with the larger initial share (and largest plants) will increase its share of the remaining domestic industry, opposite to the pattern in the data.

Moving to the quantitative analysis, we begin with 1997 estimated industry models as baselines and then take into account three considerations to arrive at model predictions for 2007. The first is the degree to which industries were affected by Chinese imports. The second is stochastic transition of region cost efficiency. The third is the change in population between 1997 and 2007. Here we explain how we implement the first two considerations (the third is immediate).

To discuss the first issue, imports, we address how to model the flow of imports internally within the United States. A complicating factor is that imports from China enter into the United States at a variety of different ports. Rather than introduce the complexity of an optimal logistics framework of port selection, we employ the following approach. We posit that at each region r with a port, and for each industry i, China has some cost-efficiency parameter in the primary segment equal to $\gamma_{i,r}^{P,China}$. We will think of this as a *new efficiency* that is creating new imports in industry i. We assume that Chinese goods at the port in r face the same internal transportation cost within the United States as would be faced by a domestic firm located at r.[23] In case of a region r with no port, we just set $\gamma_{i,r}^{P,China} = 0$. With this structure, the share of sales at destination r that are Chinese imports is given by

$$newshare_{i,r}^{P,China} = \frac{\sum_{r'=1}^{L} a_{i,r,r'}^{P} \gamma_{i,r'}^{P,China}}{\sum_{r'=1}^{L} a_{i,r,r'}^{P} \gamma_{i,r'}^{P,US} + \sum_{r'=1}^{L} a_{i,r,r'}^{P} \gamma_{i,r'}^{P,China}}, \tag{18}$$

which is analogous to (1), noting that we now add the superscript "US" to the cost efficiency term for domestic production. We calculate the national share $newshare_{i}^{P,China}$ by taking the weighted average of (18) across all destination regions r. We start by plugging in the estimated domestic cost-efficiency parameters $\gamma_{i,r'}^{P,US}$ for 1997 into (18). Next let

$$\gamma_{i,r}^{P,China} = \hat{\gamma}_{i}^{newChina} \times customs\_share_{r},$$

where $customs\_share_{r}$ is the share in the 2007 data of manufacturing imports from China going through customs at region r. We solve for the scaler $\hat{\gamma}_{i}^{newChina}$ so that the implied value of $newshare_{i}^{P,China}$ equals the China new-import share for industry i, constructed from the data as described above. With this approach, when the domestic transportation frictions are small, the predicted distribution of Chinese imports by customs district will approximate the distribution in the data.

We emphasize that in solving for $\hat{\gamma}_{i}^{newChina}$ to fit 2007 new China imports for each industry i, we are solving for an *industry-level* variable that does not use any data on the distribution of domestic production in 2007. Therefore, when we plug this parameter into the model to predict the distribution of plant counts in 2007 across regions, this is an *out-of-sample* prediction.

We turn now to explaining the second consideration, stochastic transition of cost efficiency. We can see in Table 10 that even in those high-concentration industry/regions not affected by new China imports (i.e., the last row), there tends to be some decline between 1997 and 2007 in count share. To understand this, recall that the 1703 industry/regions in the sample are a selection of cases in the right tail of the concentration distribution in 1997. We

[23] We obtain very similar results if we instead just assume that imported goods incur no additional transportation costs for shipment within the United States.

expect a *regression-to-the-mean* mechanism to play at least some quantitative role, as those at the very top, on average, tend to move down. To explicitly take into account this force, we employ the following procedure. We start with those 88 industries in the bottom category of Table 8 for which the new China share is zero. We make a grid of the cost efficiencies in 1997 and 2007 and estimate a stochastic transition process for the cost efficiencies over the grid. We then assume that the same transition matrix applies for the other industries and run 10,000 different simulations, taking averages over simulations for each region and industry. See Appendix C for more details.

We emphasize that when we plug in the estimated transition matrix to examine industries affected by Chinese imports, again this is an out-of-sample prediction. Information about the actual production distribution in 2007 in industries affected by Chinese imports is not being used in the construction.

We turn now to the results of our prediction exercise and begin by examining what happens in the standard model without specialty segments, which we call the *Primary-Only* model as above. The results in the "China Only" column of Table 11 are what we get when we allow China imports but shut down the stochastic process on cost efficiencies.[24] Mechanically, what we are doing is taking the 1997 estimate of the primary-only model as the baseline and introducing the $\hat{\gamma}_i^{newChina}$ parameter to match 2007 new imports for industry i. The bottom row of Table 11 contains the industries with zero new imports, so $\hat{\gamma}_i^{newChina} = 0$ for these industries. Of course, the predicted change is zero for these industries.

Next consider the industries affected the most by Chinese imports (the top row). The prediction of the primary-only model is that plant count shares at high-concentration regions increase *8 percent* on average from increased imports. Looking at the remaining rows in the "China Only" column, we can see that the greater the imports, the bigger the positive effect on high-concentration locations. High concentration is indicative of high efficiency, and firms in these locations tend to withstand the onslaught from China in the standard model. Earlier we noted that in a two-location version of the model, we prove formally that imports raise the domestic market share of high concentration locations. (See Appendix B for details.) Table 11 shows that the primary-only model estimated from the data has this same prediction, which is opposite to what happened in the data.

For quantitative analysis, we also need to take into account regression to the mean, as high-concentration locations can be expected to decline. The column labeled "RTM Only" (an abbreviation for "regression to the mean") starts with the 1997 baseline and adds in

[24] Population doesn't matter for the primary-only model, so accounting for population change makes no difference here.

the stochastic transition of cost efficiency and nothing else. The effect is a 15 percent to 17 percent decline across the different import categories. When we combine new imports and stochastic transition (column labeled "China+RTM"), the key prediction is that the change becomes less negative as we move up the column and increase new imports, from -17 percent at the bottom to -10 percent at the top. This is in sharp contrast to the actual pattern in the data in Table 10, where the change becomes substantially *more* negative as we move up the column and increase new imports.

We turn now to the results in Table 11 from the full model. These are the results that incorporate the specialty segment, which is what our paper is all about. When we simulate the effects of increased China import efficiency (the "China Only" column), there is a *decrease* in the high-concentration region share. Moreover, the decrease is *sharper*, the *larger* the change in China import efficiency (i.e., the more we move up the column). Adding in stochastic transition and population change, the predictions match the qualitative pattern in the data that the share changes are all negative and become sharply more negative for the highest categories (the two top rows). Moreover, the predicted magnitudes roughly approximate the actual outcomes, e.g., compare a predicted decline of 45 percent for the top import group with the actual decline of 37 percent.

The simulations discussed so far do not take into account the potential effect of imports on relative wages across regions. Autor, Dorn, and Hanson (2013) provide recent evidence on this effect of trade. While incorporating labor market equilibrium and wages is beyond the scope of this project, we have considered what happens when we plug the actual relative wage changes into our simulations, and it made little difference for our predicted distributions of industries across locations.[25]

We emphasize that our results follow from a compositional issue in which primary and specialty segments are being aggregated into the same Census industry. If it were somehow possible to separate out, a priori, primary segment plants in the data, then the primary-only model would apply when restricted to this subset. *Within* a segment, larger plants tend to be more productive than smaller plants and are more likely to survive an onslaught of imports. In general, such a priori classification is unavailable, but for illustration purposes it is useful to return once again to the wood furniture industry. For this industry, we have a proxy for specialty status, the R classification discussed in the construction of Table 2. Take the set of

[25] To take into account wage changes, we did the following. From equation (2), if technology parameters do not change across periods, then cost efficiency across two regions r and r' is proportional to how $\left(\frac{w_r}{w_{r'}}\right)^{-\theta}$ changes over time, where w_r and $w_{r'}$ are wages and θ is a model parameter. We use region-level data from the Quarterly Census of Employment and Wages to plug in for wages and set $\theta = 3.6$ as in BEJK. We redid the "China-only" versions of the primary-only and full models, and it made little difference in the results.

wood furniture plants in 1997 and regress whether or not a plant survives going into 2007, on a dummy variable for R classification as well as log 1997 employment. The coefficient on the R dummy is .20 ($s.e. = .04$), and the coefficient on log employment is .07($s.e. = .005$).[26] To the extent that the R classification accurately controls for specialty status, the coefficient .07 on log employment can be interpreted as the "within-segment" effect of size. Note that the magnitude of the R coefficient is quite large, equivalent to a three-fold increase in log employment (or a 20-fold increase in the level of employment).[27]

6.3 Measuring Changes in the Primary Segment Share

Taking the relative success of out-of-sample prediction results as validation of our general model, our final exercise brings in the 2007 outcome data to reestimate the model. We put our procedure to work taking a measurement of the division of industry between primary and specialty plants in 2007, using the 2007 plant distribution data, and compare this with our earlier estimates for 1997.[28] Following the theory, in those industries being hardest hit by imports, we expect the primary segment share to be falling relative to changes in the primary segment share in other industries.

Table 12 presents the results. Consider the top two import groups. In addition to the overall decline in plant counts, the primary segment share of what remains declines substantially over the period for both groups. In the first group, the primary segment share fell from 32 percent to 25 percent, in the second, the share fell from 27 percent to 16 percent. This is a very different outcome from what happens with the remaining industries. In particular, in the next three groups, the primary segment share is virtually unchanged over the period. In the bottom category with zero new China share, the primary segment share actually increases over time.

[26] The constant is .16 ($s.e. = .03$), and $R^2 = .044$, and $N = 3884$.

[27] If we take out the R dummy variable, the coefficient on size remains positive, but it is attenuated. Very small plants have high entry and exit for a variety of reasons. Our theory of the effect of imports applies at the segment/region level. A two-employee craft furniture shop in Chicago in 1997 might have exited by 2007 and been replaced by another two-employee craft shop in Chicago. In contrast, a thousand-employee factory in High Point that contracted or exited was not replaced by a similar factory in High Point.

[28] We hold fixed the η estimate from 1997 (the parameter vector governing distance adjustment) and otherwise run the second-stage estimation for 2007 the same way as we did for 1997. The 2007 CFS data were not available, so we could not use those data to create a 2007 specific estimate. We think it is sensible to treat the underlying transportation structure governing the η parameter as relatively constant over the period 1997–2007. This justifies our use of the 1997 estimate for η.

7 Conclusion

This paper develops a model in which industries are made up of mass-production factories making standardized goods and specialty plants making custom or niche goods. The paper uses a combination of confidential and public Census data to estimate parameters of the model and, in particular, produces estimates of plant counts in the specialty and primary segments by industry. The estimated model fits the observed plant size/geographic concentration relationship relatively well. The estimates reveal that, for those industries that have been heavily affected by a surge of imports from China, there has been a significant decline in the share of plants in the primary segment. These results are consistent with the hypothesis that products imported from China are close substitutes to the products of large plants (like those in High Point) and not so close substitutes to the products of the small plants that are diffuse throughout the country. The results are inconsistent with standard theories that assert that small plants are just like large plants, except for having a low productivity draw.

We highlighted one particular mechanism—the specialty/primary distinction—as an alternative to the standard productivity mechanism and then incorporated both mechanisms into the theory. There are potentially other systematic differences between large and small plants, in addition to the two considered here. For example, small plants might tend to be different technologically from large plants (e.g., perhaps they have a different way of combining inputs). While there are other mechanisms one might consider, the fact that this theory is able to account for a variety of facts about industrial organization, geography, and intranational trade motivates our choice to focus on it.

There is a labor literature about how firms have responded to the threat of foreign competition by *changing* the products they produce. Specifically, there are case studies of firms getting out of the business of trying to produce mass quantities of standardized products, focusing instead on custom manufacturing. See in particular the Freeman and Kleiner (2005) study on how personnel practices change in conjunction with product changes. There is also a trade literature on how a firm's product quality choice may be affected by trade, e.g., Verhoogen (2008). Undoubtedly, trade can have significant effects on what surviving firms do, and future work can take that into account. However, there are limits to what a given plant can do to reinvent itself. Consider a megaplant making standardized furniture in North Carolina and reinventing itself to make craft furniture. The initial comparative advantage that attracted the industry in the first place was the abundance of low-wage, low-skill workers in the area. Indeed, in an earlier period, labor-intensive manufacturing moved to North Carolina for the same reasons it is moving to China today. But having an abundance of low-skill labor does not necessarily imply a location is particularly rich in the skilled craftsmen that would be needed for craft work. Also, if it's advantageous to procure custom work locally, this limits how far North Carolina can expand into the craft business. In short, this discussion shows how taking into account geography and intranational trade can be crucial for understanding the effects of international trade.

Appendix A: Data

This section of the Appendix discusses various topics regarding the data. We have posted online a supplementary document, Holmes and Stevens (2013), that contains links to public-use data sets, supplementary tables, and various programs. In particular, we post the first-stage and second-stage model estimates for each individual industry.

A.1. Joint Use of Confidential and Public Data

The Census cannot disclose results that could potentially compromise the confidentiality of the Census data. In particular, statistics at a high level of geographic and industry detail are problematic for getting through the disclosure process.

It is useful to be able to simulate the estimated model outside a secure Census facility. This makes it possible to easily replicate findings of the paper. It also makes it possible to make tables of the predictions of the model at narrow geographic and industry detail. To accomplish this goal, we made joint use of confidential and publicly available data. First, we used the confidential micro CFS data to estimate the parameter vector η_i^P of the distance adjustment function $a_i^p(dist)$ for each industry i. The release of these aggregate industry variables posed no disclosure risk. Second, we used publicly available data to estimate the distribution of industry revenue across regions and then used these revenue estimates to estimate the productivity vector Γ_i^P for each industry.

To construct revenue estimates by region with public data, we begin with the 1997 Location of Manufacturing plants (LM) data series. The LM data are public data of cell counts by industry, county, and narrowly defined employment size categories (0–4, 5–9, 10–19, 20–49, 50–99, 100–249, 250–499, 500–999, 1,000–2,499, 2,500 and above). This is file E9731e2 from the 1997 Economic Census CD (U.S. Bureau of the Census (2001)). These cell counts are not considered a disclosure, and no information is held back. Thus, for each of the 362,829 plants in the Census of Manufactures, we have as public data the plant's county, industry, and detailed employment size category. We produce an estimate of average sales revenue conditioned on industry and employment size, and assign this to each plant as an estimate of its sales.[29] We then aggregate up the plant-level sales estimates to the BEA Economic Area level for each industry to obtain estimates of the region-level data needed to implement the second stage of our procedure. We found that in running the second-stage procedure, it made little difference when we used the estimated revenue data instead of the actual revenue data. For 2007, we use the analogous cell count data from County Business

[29] We used public data on sales by size category to estimate a regression model of log sales, with employment size dummies and industry dummies. (This is file E9731g4 from the above-mentioned 1997 Economic Census CD.) We pooled the data across industries because the sales revenue data for some cells are held back due to disclosure issues. We scaled the fitted values so the aggregate totals matched published aggregates.

Patterns (U.S. Bureau of the Census 2007), and we use the same mapping between sales estimate and employment size category that we use for 1997.

A.2. Defining a Narrow Set of Industries

As discussed in the text, we apply our procedure to all industries, but also consider a narrow sample of 175 industries in which we argue that population is likely to be a good proxy for demand. We use the U.S. Benchmark Input-Output Accounts for 2002 (Stewart, Stone, and Streitwieser (2007)) to classify industries. We first define a set of downstream demand (or "use") of goods that we categorize as "local demand." These are all uses in the following categories: structures, transportation, retail trade, information (publishing), finance, insurance, real estate, services, personal consumption expenditures, fixed private investment, and government services. Then for each commodity, we determine which of these deliver 75 percent or more of their sales to the local demand categories. We define these to be *Diffuse Demand* industries. We also include borderline cases of Surgical and Dental Equipment (NAICS 339112) and Dental Equipment (NAICS 339114). There are 175 6-digit NAICS industries fitting this description. The remaining 298 industries (out of a total count of 473 = 175 + 298 6-digit industries) are the *Nondiffuse Demand Industries.*

A.3. More Information about Data Sources

Section 2. To calculate the first and second basic facts, we use the public 1997 LM data described above.

Tables 1 and 2. See the supplementary online Appendix for the formula for how we calculate the location quotients.

Table 2. The cell counts that cross-classify the 1997 Census plants by NAICS and SIC use public data files E97B1 and E97B2 distributed at the Census website. These files report employment by SIC and NAICS, from which mean plant size by SIC/NAICS was derived. The confidential micro data were used to calculate export shares and mean plant sales location quotients in the table.

Table 3. We create a variable "custom share" using the product shipments file in the 1997 Census. For each narrowly defined product category, we inspect the text of the definition and define the shipment as custom or not depending on whether the word "custom" appears in the text field and how it appears. For example, product code 3371107121 is defined as "Wood vanities and other cabinetwork, custom." This is distinguished from another product in which "stock line" is used in place of "custom." In the calculation we exclude plants with imputed product-level data. For each plant with actual data, we take the dollar average across shipments to determine custom share. The statistics reported in Table 3 are the unweighted

means across plants with the available data.

Tables 4 and 5. These tables are estimated with confidential micro data at the Census. Population at the EA level was calculated using Census county-level estimates for 1992 and 1997 (and below for 2007) and aggregating up to the EA level.

Tables 6, 7, and 9–12 use publicly available data as discussed in A.1 above.

Table 8 uses import information posted by the U.S. International Trade Commission at its website. (For two furniture industries, NAICS 337122 and 337125, we used revised figures reported at the website of the International Trade Administration.) Imports are "Customs Value of U.S. Imports for Consumption." The variables are reported at the six-digit NAICS level, though in some cases the five-digit level is used. To simulate the geographic distribution of imports from China, we use published Census data on the geographic distribution of 2007 manufacturing imports from China across custom districts, file IMP_DETL in U.S. Bureau of the Census (2008). We assign custom districts to economic areas based on the location of the main office of the custom district.

A.4. Goodness of Fit of First-Stage Estimates

Here we examine the goodness of fit of the first-stage estimates. Recall that by construction, total shipments originating in each region across all destinations in the estimated model perfectly fit the data. For a notion of goodness of fit, we break up destinations by distance and compare the model and the data. In particular, we break shipments above 100 miles into three distance categories: (1) 100 to 500 miles, (2) 500 to 1,000 miles, and (3) over 1,000 miles. For industry i, let $share_{c,i}$ be the share of the shipments above 100 miles that are in distance category $c \in \{1, 2, 3\}$ in the data. Let $\widehat{share}_{c,i}$ be the fitted value in the estimated model. Table A1 presents descriptive statistics, where industries are separated by whether or not they are classified as Diffuse Demand or Nondiffuse Demand (see Appendix A.2). For diffuse-demand industries, in the data, on average across the industries, a share .45 of the 100-mile-plus shipments are in the 100- to 500-mile category. This figure compares with an average share of .38 in the estimated model. The model has a tendency to somewhat understate the shortest distance category and somewhat overstate the two longer distance categories. By construction, the destination of shipments in the model exactly follows the distribution of population. So regions far away from any producers will nevertheless be required in the model to receive their share of shipments. The bottom part of Table A1 shows that the fitted values of the model do a very good job in accounting for the cross-industry variation in the distance distribution. The slope of $share_{c,i}$ in a regression on $\widehat{share}_{c,i}$ across industries i is approximately one for each category $c \in \{1, 2, 3\}$. Next note that the fit in these regressions for the nondiffuse demand industries is not quite as good as it is for the diffuse demand industries. This is not surprising because the diffuse demand industries were selected out as industries for which the assumption that demand follows the distribution of population could be expected to work well.

Appendix B: Theoretical Results on the Plant Size/Geographic Concentration Relationship and Imports

This part of the appendix presents theoretical results referred to in Section 6 regarding the properties of the primary-only model. The first result is that with no transportation costs, average plant size is constant across locations.

Proposition A1. In the primary-only model with no transportation costs ($d_{\ell,\ell'} = 1$ all $\ell \neq \ell'$), average plant size (in sales volume) is identical at all locations.

Proof. With $d_{\ell',\ell} = 1$ all $\ell \neq \ell'$, the probability (1) that ℓ serves ℓ' reduces to

$$\phi_{\ell',\ell} = \frac{\gamma_\ell}{\sum_{\ell^\circ} \gamma_{\ell^\circ}},$$

which is independent of destination ℓ'. From BEJK this also equals the sales share at each location. Therefore, average plant size (in sales revenue) at each location is

$$\bar{r}_\ell = \frac{\sum_{\ell'} \beta \phi_{\ell'\ell} x_{\ell'}}{n_\ell} = \frac{\sum_{\ell'} \beta \phi_{\ell'\ell} x_{\ell'}}{\lambda \phi_{\ell,\ell}} = \frac{\beta \sum_{\ell'} x_{\ell'}}{\lambda}, \tag{19}$$

which is constant across locations. ■

We next examine a two-location version of the primary-only model with equal population at each location. We show that if transportation costs are not zero, then the location with a greater share of the industry will also be the location with the higher average plant size. We also show that if a source of foreign imports emerges, the location that initially has the larger market share and larger average plant size will increase its share of what is left of domestic production.

Assume there are two equal population domestic locations. Assume there is a source of imports from a third location C. Assume symmetric distance adjustment between the two domestic locations, $a_{21} = a_{12} = \alpha < 1$. Let a_C be the distance adjustment of imports, assumed to be the same at each domestic location. The probabilities that a sale at location 1 originates at location 1 or location 2 equal

$$\phi_{11} = \frac{\gamma_1}{\gamma_1 + \alpha\gamma_2 + a_C\gamma_C}, \tag{20}$$

$$\phi_{12} = \frac{\alpha\gamma_2}{\gamma_1 + \alpha\gamma_2 + a_C\gamma_C}. \tag{21}$$

The equations ϕ_{21} and ϕ_{22} are analogous. Assume $\gamma_2 > \gamma_1$, so that location 2 has more than half of domestic sales, while location 1 has less than half.

Proposition A2. Under the above assumptions: (i) If imports are zero, average plant size is strictly higher at location 2 than at location 1. (ii) Location 2's share of domestic sales (the sum of location 1 and location 2 sales) strictly increases in the China competitiveness parameter γ_C, and the same is true for its share of domestic plant counts.

Proof. Using the formulas (20) and $\gamma_1 < \gamma_2$, it is immediate that $\phi_{22} > \phi_{11}$. Recalling that population is the same at the two locations, the ratio of sales between the two locations is

$$\begin{aligned} \frac{y_2}{y_1} &= \frac{\phi_{12} + \phi_{22}}{\phi_{11} + \phi_{21}} = \frac{\frac{\alpha\gamma_2}{\gamma_1+\alpha\gamma_2+a_C\gamma_C} + \frac{\gamma_2}{\alpha\gamma_1+\gamma_2+a_C\gamma_C}}{\frac{\gamma_1}{\gamma_1+\alpha\gamma_2+a_C\gamma_C} + \frac{\alpha\gamma_1}{\alpha\gamma_1+\gamma_2+a_C\gamma_C}} \\ &= \frac{\gamma_2}{\gamma_1} \frac{(1+\alpha^2)\gamma_1 + 2\alpha\gamma_2 + (1+\alpha) a_C\gamma_C}{(1+\alpha^2)\gamma_2 + 2\alpha\gamma_1 + (1+\alpha) a_C\gamma_C}, \end{aligned} \tag{22}$$

where the second line follows from straightforward manipulations. The ratio of plant counts is

$$\frac{n_2}{n_1} = \frac{\phi_{22}}{\phi_{11}} = \frac{\gamma_2}{\gamma_1} \frac{\gamma_1 + \alpha\gamma_2 + a_C\gamma_C}{\alpha\gamma_1 + \gamma_2 + a_C\gamma_C}.$$

To prove part (i), we must show that $\frac{y_2}{y_1} > \frac{n_2}{n_1}$. Setting imports to zero, $\gamma_C = 0$, this follows from straightforward calculations using the formulas above and $\alpha < 1$ and $\gamma_1 < \gamma_2$. To prove part (ii), we need to show that the ratios $\frac{y_2}{y_1}$ and $\frac{n_2}{n_1}$ increase in γ_C, for fixed γ_1 and γ_2. From inspection of (22), $\frac{y_2}{y_1}$ increases if

$$\left(1+\alpha^2\right)\gamma_1 + 2\alpha\gamma_2 < \left(1+\alpha^2\right)\gamma_2 + 2\alpha\gamma_1$$

or

$$0 < (\gamma_2 - \gamma_1)\left(1 - 2\alpha + \alpha^2\right) = (\gamma_2 - \gamma_1)(1-\alpha)^2,$$

which holds since $\alpha < 1$ and $\gamma_2 > \gamma_1$. Analogously, $\frac{n_2}{n_1}$ increases in γ_C. ■

Appendix C: Details of the Stochastic Transition of Cost Efficiency Used in Section 6.

We allow for an 11-point grid for cost efficiency γ, assigning all zero values to one point and positive values to 10 points, according to the following percentiles (conditioned on a positive value): 1, 10, 25, 50, 75, 90, 95, 99, 99.5. After first normalizing the γ in each industry so they sum to one across regions, we set the value of $\hat{\gamma}_g$ for grid point g equal to the exponential of mean log of the estimated normalized values of γ within the grid interval. We also estimate the model for 2007 and normalize the γ to sum to one and grid out the 2007 estimates as well. Through this process, for each industry, region, and time period 1997 to 2007, we can assign cost efficiency levels in the 11-point grid $g \in \{0, 1, ..., 10\}$.

We estimate the 11×11 transition matrix from g_{1997} to g_{2007} by plugging in the empirical values of the matrix for the 88 industries classified in Table 8 has having 0 China new import share. We weight industry/regions by employment levels of the industry/region (plus one). When we simulate the model, we assume that the other industries follow the same transition matrix.

For the exercises in Table 11 that employ the stochastic transition matrix, we generate 10,000 simulation draws for each industry/region and take averages.

To construct Table 11, we first solve for $\hat{\gamma}_i^{newChina}$ for a given industry i, assuming complete persistence of cost efficiency between 1997 and 2007. We use this to construct the "China Only” columns. When we add in stochastic transition, we hold fixed $\hat{\gamma}_i^{newChina}$ as calculated in the “China Only” case. Given that we have estimated a transition matrix such that the γ sum to one at each date, on average the imports from China will be approximately the same when we incorporate stochastic transition.

References

Alessandria, George, and Horag Choi. 2007. "Establishment Heterogeneity, Exporter Dynamics, and the Effects of Trade Liberalization." Working Paper no. 07-17 (July), Research Department, Federal Reserve Bank of Philadelphia.

Anderson, James E., and Eric van Wincoop. 2003. "Gravity with Gravitas: A Solution to the Border Puzzle." *American Economic Review* 93 (March): 170–92.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review* 103 (October): 212–68.

Baldwin, Richard, and James Harrigan. 2011. "Zeros, Quality, and Space: Trade Theory and Trade Evidence." *American Economic Journal: Microeconomics* 3 (May): 60–88.

Bernard, Andrew B., Jonathan Eaton, J. Bradford Jensen, and Samuel Kortum. 2003. "Plants and Productivity in International Trade." *American Economic Review* 93 (September): 1268–90.

Bernard, Andrew B., and J. Bradford Jensen. 1995. "Exporters, Jobs, and Wages in U.S. Manufacturing, 1976–1987." *Brookings Papers on Economic Activity: Microeconomics* 1995 (1995): 67–119.

Bernard, Andrew B., Stephen J. Redding, and Peter K. Schott. 2010. "Multiple-Product Firms and Product Switching." *American Economic Review* 100 (March): 70–97.

Berry, Steven T. 1994. "Estimating Discrete-Choice Models of Product Differentiation." *RAND Journal of Economics* 25 (Summer): 242–62.

Bloom, Nicholas, Mirko Draca, and John Van Reenen. 2011. "Trade Induced Technical Change? The Impact of Chinese Imports on Innovation, IT and Productivity." CEP Discussion Paper no. 1000 (January), Centre for Economic Performance, LSE.

Buera, Francisco J., and Joseph P. Kaboski. 2012. "Scale and the Origins of Structural Change." *Journal of Economic Theory* 147 (March): 684–712.

Eaton, Jonathan, and Samuel Kortum. 2002. "Technology, Geography, and Trade." *Econometrica* 70 (September): 1741–79.

Ellison, Glenn, and Edward L. Glaeser. 1997. “Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach.” *Journal of Political Economy* 105 (October): 889–927.

Freeman, Richard B., and Morris M. Kleiner. 2005. “The Last American Shoe Manufacturers: Decreasing Productivity and Increasing Profits in the Shift from Piece Rates to Continuous Flow Production.” *Industrial Relations* 44 (April): 307–30.

Hallak, Juan Carlos, and Jagadeesh Sivadasan. 2009. “Firms’ Exporting Behavior under Quality Constraints.” Working Paper no. 14928 (April), NBER, Cambridge, MA.

Hillberry, Russell, and David Hummels. 2003. “Intranational Home Bias: Some Explanations.” *Review of Economics and Statistics* 85 (November): 1089–92.

Hillberry, Russell, and David Hummels. 2008. “Trade Responses to Geographic Frictions: A Decomposition Using Micro-Data.” *European Economic Review* 52 (April): 527–50.

Holmes, Thomas J. 2011. “The Case of the Disappearing Large-Employer Manufacturing Plants: Not Much of a Mystery After All.” Economic Policy Paper no. 11-4, Federal Reserve Bank of Minneapolis.

Holmes, Thomas J., and John J. Stevens. 2002. “Geographic Concentration and Establishment Scale.” *Review of Economics and Statistics* 84 (November): 682–90.

Holmes, Thomas J., and John J. Stevens. 2004. “Geographic Concentration and Establishment Size: Analysis in an Alternative Economic Geography Model.” *Journal of Economic Geography* 4 (June): 227–50.

Holmes, Thomas J., and John J. Stevens. 2012. “Exports, Borders, Distance, and Plant Size.” *Journal of International Economics* 88 (September): 91–103.

Holmes, Thomas J., and John J. Stevens. 2013. “Supplementary Online Appendix for ‘An Alternative Theory of the Plant Size Distribution, with Geography and Intra- and International Trade’ with Data and Programs.” Available at http://www.econ.umn.edu/~holmes/data/plantsize.

Hopenhayn, Hugo A. 1992. “Entry, Exit, and Firm Dynamics in Long Run Equilibrium.” *Econometrica* 60 (September): 1127–50.

Hopenhayn, Hugo, and Richard Rogerson. 1993. “Job Turnover and Policy Evaluation: A General Equilibrium Analysis.” *Journal of Political Economy* 101 (October): 915–38.

Jensen, J. Bradford, and Lori G. Kletzer. 2005. "Tradable Services: Understanding the Scope and Impact of Services Offshoring." *Brookings Trade Forum* 2005 (2005): 75–116.

Johnson, Kenneth P., and John R. Kort. 2004. "2004 Redefinition of the BEA Economic Areas." *Survey of Current Business* 84 (November): 68–75.

Jovanovic, Boyan. 1982. "Selection and the Evolution of Industry." *Econometrica* 50 (May): 649–70.

Juran, Joseph M. 1954. "Universals in Management Planning and Control." *Management Review* 43 (November): 748–61.

Lucas, Robert E., Jr. 1978. "On the Size Distribution of Business Firms." *Bell Journal of Economics* 9 (Autumn): 508–23.

Melitz, Marc J. 2003. "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity." *Econometrica* 71 (November): 1695–1725.

Office of Management and Budget. 1994. "Economic Classification Policy Committee: Standard Industrial Classification Replacement." *Federal Register* 59 (July).

Piore, Michael J., and Charles F. Sabel. 1984. *The Second Industrial Divide: Possibilities for Prosperity.* New York: Basic Books.

Stewart, Ricky L., Jessica Brede Stone, and Mary L. Streitwieser. 2007. "U.S. Benchmark Input-Output Accounts, 2002." *Survey of Current Business* 87 (10): 19–48.

U.S. Bureau of the Census. 1999. *1997 Commodity Flow Survey*, EC97TCF-US.

U.S. Bureau of the Census. 2001. 1997 Economic Census. CD-ROM. Washington, DC: U.S. Department of Commerce.

U.S. Bureau of the Census. 2007. County Business Patterns (CBP). www.census.gov/econ/cbp/.

U.S. Bureau of the Census. 2008. U.S. Imports of Merchandise, Statistical Month - December 2007, Issued February 2008 (DVD-ROM).

Verhoogen, Eric A. 2008. "Trade, Quality Upgrading, and Wage Inequality in the Mexican Manufacturing Sector." *Quarterly Journal of Economics* 123 (May): 489–530.

Table 1
Mean Location Quotient by Plant Size for Three Groups of Industries

Industry Grouping	Plant Size Category	Number of Establishments	Employment Share (Percent)	Mean Location Quotient	
				Raw	NAICS Fixed Effect
All Industries (473 Industries)					
	All	361,516	100.0	5.04	5.04
	1–19	241,339	8.6	1.80	3.68
	20–99	85,069	22.3	2.66	4.09
	100–499	30,324	36.6	4.19	4.80
	500+	4,784	32.5	6.84	5.69
Reclassification Industry Sample (7 Industries)	All	18,585	100.0	4.32	4.32
	1–19	15,687	17.5	1.36	2.46
	20–99	2,073	20.4	2.62	3.03
	100–499	690	33.7	4.18	3.97
	500+	135	28.4	6.13	5.83

Source: The mean location quotients and establishment counts are the authors' calculations with confidential data from the 1997 Economic Census. The employment shares are the authors' calculations with published tabulations from the 1997 Economic Census.

Table 2
Descriptive Statistics for the 1997 Reclassification Industries
by R and M Status

NAICS Industry Classification	Classification Based on SIC	Number of Plants	Mean Plant Employ.	Export Share	Mean Location Quotient
Chocolate Candy	R	440	8	.00	1.01
(NAICS 311330)	M	421	70	.03	4.87
Nonchocolate Candy	R	349	4	.00	1.01
(NAICS 311340)	M	276	88	.03	4.61
Curtains	R	1,085	4	.00	1.54
(NAICS 312121)	M	999	21	.03	3.22
Other Apparel	R	724	3	.00	1.71
(NAICS 315999)	M	966	25	.03	2.67
Kitchen Cabinets	R	2,055	5	.00	1.25
(NAICS 337110)	M	5,908	15	.01	2.14
Upholstered Household Furniture	R	576	5	.00	1.52
(NAICS 337121)	M	1,130	77	.03	7.20
Wood Household Furniture	R	815	6	.00	1.17
(NAICS 337122)	M	3,035	41	.03	4.42

Source: Authors' calculations with confidential Census data and public Census tabulations.
Note: An "R" plant is classified in retail under SIC and manufacturing under NAICS. An "M" plant is in manufacturing under both classifications.

Table 3
Mean Custom Share across Sample Plants in the Furniture Industry
by R and M Status and within M by Employment Size

Industry Grouping	Classification	Number of Sample Plants	Mean Custom Share
Kitchen Cabinets and Household Furniture (NAICS 337110, 337121, 337122)	R	102	.82
	M	2,944	.42
	Within M by Emp. Size		
	1–19	1,628	.59
	20–99	877	.30
	100 and above	437	.09
Kitchen Cabinets (NAICS 337110)	R	48	.87
	M	1,854	.64
	Within M by Emp. Size		
	1–19	1331	.70
	20–99	426	.56
	100 and above	97	.28
Household Furniture (NAICS 337121, 337122)	R	54	.78
	M	1,090	.05
	Within M by Emp. Size		
	1–19	297	.08
	20–99	451	.05
	100 and above	340	.03

Source: Authors' calculations with confidential Census data.

Table 4
First-Stage Results for the Seven Reclassification Industries

Reclassification Industries	Log-Log Case Constant Elasticity	Semi-Log Case Elasticity 100 miles	Semi-Log Case Elasticity 500 miles	Log-Log Case Log-Like	Semi-Log Case Log-Like	Number of Shipment Obs.
Chocolate Candy (311330)	0.31	0.06	0.37	-6947.0	-6842.8	1,633
Nonchocolate Candy (311340)	0.38	0.09	0.45	-10843.6	-10690.6	2,592
Curtains (314121)	0.69	0.23	0.90	-12917.9	-12864.0	3,068
Other Apparel (315999)	0.55	0.23	0.85	-17470.7	-17414.4	3,819
Kitchen Cabinets (337110)	1.16	0.41	1.62	-27046.6	-26767.7	6,655
Upholstered Household Furn. (337121)	1.19	0.36	1.47	-36714.7	-36439.6	8,837
Wood Household Furn. (337122)	0.67	0.22	0.84	-67889.5	-67746.4	15,624

Source: See supplementary online Appendix.

Table 5
Estimated Elasticities from First Stage
Means and Percentiles Across *N* = 466 Industries

Statistic	Log-Log Case Constant Elasticity	Semi-Log Case Elasticity 100 miles	Semi-Log Case Elasticity 500 miles
Mean	.61	.24	.99
25th Percentile	.27	.08	.36
50th Percentile	.53	.17	.68
75th Percentile	.82	.29	1.12

Source: See supplementary online Appendix.

Table 6
Second-Stage Estimates of the Plant Count Parameters and Related Model and Data Statistics

	Regression Results (s.e. in Parentheses)				Estimated Share of Plants that Are Specialty (Percent)
	Constant ξ	Slope Specialty λ^S	Slope Primary λ^P	R^2	
Reclassification Industries					
Chocolate Candy (311330)	.4	621.3	76.6	.69	80.8
	(.2)	(52.3)	(13.0)		
Nonchocolate Candy (311340)	.1	487.8	62.6	.79	80.4
	(.1)	(31.2)	(7.9)		
Curtains (314121)	-.9	2186.4	20.9	.92	97.1
	(.3)	(56.4)	(8.1)		
Other Apparel (315999)	-1.8	1287.0	287.0	.90	62.3
	(.4)	(110.8)	(20.4)		
Kitchen Cabinets (337110)	1.5	5963.3	210.2	.89	78.4
	(1.2)	(230.5)	(21.5)		
Upholstered Household Furn. (337121)	-1.3	975.5	270.3	.88	49.8
	(.5)	(94.7)	(8.6)		
Wood Household Furn. (337122)	-1.1	3362.7	270.4	.85	82.2
	(.8)	(138.9)	(24.7)		
Mean Reclass. Industries (*N* = 7)	-.4	2122.4	171.1	.85	75.9
All Industries (*N* = 466)					
Mean	-.3	550.9	91.1	.71	63.9
Minimum	-19.7	-1.4	.2	.21	0.0
10th Percentile	-.9	33.8	7.1	.50	35.4
25th Percentile	-.4	95.3	13.8	.61	54.0
50th Percentile	-.1	238.2	35.4	.72	68.0
75th Percentile	.0	515.2	90.3	.82	79.0
90th Percentile	.2	1172.5	221.4	.88	86.3
Maximum	6.2	18854.2	3160.9	.98	99.9

Source: See supplementary online Appendix.

Table 7
Sales, Count, and Size Quotients in Data, Size Quotients for Two Models
in High-Concentration Industry/Regions

			Data			Primary-	
						Only	Full
		Revenue				Model	Model
Industry	Region	Share	Q^{rev}	Q^{count}	Q^{size}	Q^{size}	Q^{size}
Chocolate Candy (NAICS 311330)	Harrisburg, PA	.07	9.2	1.4	6.4	1.3	4.2
	Nashville, TN	.06	6.6	.8	7.9	1.4	3.8
	Chicago, IL	.15	4.1	1.2	3.6	1.4	3.0
	Philadelphia, PA	.08	3.3	2.1	1.6	1.2	2.5
	San Francisco, CA	.08	2.3	1.5	1.6	0.4	1.3
Nonchocolate Candy (NAICS 311340)	Grand Rapids, MI	.07	11.2	1.2	9.2	1.4	4.5
	Chicago, IL	.24	6.8	1.5	4.6	1.4	3.8
	Atlanta, GA	.07	3.5	.8	4.5	1.2	2.5
Curtains (NAICS 314121)	San Antonio, TX	.07	10.3	.6	16.8	0.7	7.0
	Raleigh-Durham, NC	.09	10.1	1.1	8.9	1.5	8.4
	Charlotte, NC	.06	7.6	1.7	4.5	1.5	6.7
	Boston, MA	.07	2.5	1.4	1.8	1.1	2.4
Other Apparel (NAICS 315999)	New York, NY	.28	3.5	2.6	1.4	1.5	2.2
	Los Angeles, CA	.16	2.5	2.1	1.2	1.0	1.5
Kitchen Cabinets (NAICS 337110)	Harrisburg, PA	.05	7.4	1.7	4.4	2.1	4.4
	Dallas, TX	.05	2.4	1.0	2.4	1.1	1.8
Upholstered Household Furn. (NAICS 337121)	Tupelo, MS	.21	107.4	43.4	2.5	1.5	2.8
	Charlotte, NC	.19	23.2	11.3	2.1	1.8	3.2
	Knoxville, TN	.09	22.2	2.6	8.6	1.7	3.0
	High Point, NC	.12	19.2	11.0	1.8	1.8	3.2
Wood Household Furn. (NAICS 337122)	High Point, NC	.17	27.7	4.2	6.6	1.6	6.9
	Charlotte, NC	.13	15.5	2.9	5.4	1.6	5.8
	Toledo, OH	.05	13.5	.8	17.5	1.3	4.8
Summary Statistics							
Reclass. Industries (*N* = 7 Industries)	*N* = 23 Industry/Regions						
	Mean	.11	14.0	4.3	5.4	1.4	3.9
	Median	.08	7.6	1.5	4.5	1.4	3.2
All Industries (*N* = 466 Industries)	*N* = 1708 Industry/ Regions						
	Mean	.11	17.6	6.5	4.4	1.2	3.0
	Median	.09	9.4	3.2	2.4	1.1	2.4

Source: See supplementary online Appendix.

Table 8
Summary Statistics for 6-Digit NAICS Industries Classified by New China Share Category

	Count of Industries	Mean of New China Share 1997-2007	Mean of New All-Country Share 1997-2007	Mean Industry Employment Growth 1997-2007 (percent)
All Industries	465	8	12	-21
By New China Share Category (percent)				
50 to 97	23	71	72	-75
25 to 50	23	35	38	-51
10 to 25	50	16	21	-29
5 to 10	38	7	14	-33
>0 to 5	243	1	7	-12
None	88	0	1	-14

Note: There are 12 industries for which the new China share is negative. In all of these cases, the value is negligible. (The minimum is -.013 and the mean is -.003.) For these cases we truncate the new China share at zero. Analogously, we truncate the new all-country share at zero. There are 465 industries in this table rather than 466; NAICS 339111, "Laboratory apparatus & furniture mfg," had no data for 2007 because the industry was discontinued.

Table 9
Cell Counts by Plant Size in 1997 and 2007
for the United States as a Whole and at the Primary Region of an Industry
for Wood Furniture and for the 46 Industries with New China Share above 25 Percent

Panel A: Wood Furniture Industry

	Plant Size Cell Counts in United States		Plant Size Cell Counts in Primary Region (High Point, NC)	
Employment Size Class	1997	2007	1997	2007
All Plants	3,835	3,568	101	53
1 to 19	3,091	3,079	51	29
20 to 49	323	278	9	6
50 to 99	165	95	6	5
100 to 249	130	68	8	3
250 to 499	78	28	13	7
500 to 999	36	15	9	3
1,000 and above	12	5	5	0

Panel B: All 46 Industries with New China Share Above 25 Percent

	Plant Size Cell Counts in United States		Plant Size Cell Counts in Primary Region	
Employment Size Class	1997	2007	1997	2007
All Plants	24,192	16,534	1,666	732
1 to 19	16,258	12,674	1,043	507
20 to 49	3,300	1,857	261	97
50 to 99	1,865	891	121	50
100 to 249	1,601	712	106	38
250 to 499	654	248	57	25
500 to 999	344	94	39	8
1,000 and above	170	58	39	7

Source: The cell counts are based on public tabulations from the Census discussed in Appendix B. The High Point, NC region consists of the BEA Economic Area containing High Point, NC (22 counties). For each industry, the Primary Region is the Economic Area with the highest sales revenue location quotient among regions with at least 5 percent of U.S. sales.

Table 10
Actual Values in Data for High-Concentration Industry/Regions
Size and Count Quotients and Change in Count Quotients
Means Across New China Share Categories

	Number of High Concen. Industry/ Regions	Q^{size} 1997	Q^{count} 1997	Q^{count} 2007	Percent Change in Count Shares 1997-2007
All	1703	4.4	6.5	5.5	-10
By New China Share Category (percent)					
50 to 97	92	5.7	4.8	2.6	-37
25 to 50	101	4.5	3.8	2.8	-20
10 to 25	179	4.2	5.8	5.1	-5
5 to 10	130	5.0	5.0	4.4	-9
>0 to 5	890	4.1	7.7	6.7	-6
None	311	4.6	5.6	4.6	-10

Source: See supplementary online Appendix.

Table 11
Predicted Values of Various Models for High-Concentration Industry/Regions
Percent Change in Count Shares Between 1997 and 2007
Means Across New China Share Categories

By New China Share Category (percent)	Number of High Concen. Industry/ Regions	Primary-Only Model			Full Model			
		China Only	RTM Only	China +RTM	China Only	RTM Only	China +RTM	China +RTM + 2007 Population
50 to 97	92	8	-17	-10	-37	-13	-43	-45
25 to 50	101	4	-15	-12	-12	-11	-20	-23
10 to 25	179	2	-17	-15	-5	-12	-16	-19
5 to 10	130	1	-16	-16	-2	-12	-14	-18
>0 to 5	890	0	-17	-17	0	-14	-14	-18
None	311	0	-17	-17	0	-13	-13	-16

Source: See supplementary online Appendix. The *China Only* case incorporates only the effect of China's import efficiency change. The *RTM Only* case incorporates only regression to the mean, calculated with the estimated stochastic transition process on regional efficiencies. *China+RTM* combines both effects.

Table 12
Estimated Primary Segment Plant Count Share for 1997 and 2007
by New China Share Categories

New China Share Category (percent)	Total Establishment Count (1,000 plants)		Estimated Primary Segment Count Share (percent)		Change in Percent Primary
	1997	2007	1997	2007	
50 to 97	10	5	32	25	-6
25 to 50	15	11	27	16	-11
10 to 25	34	30	31	31	1
5 to 10	24	20	30	32	2
>0 to 5	156	153	34	33	-1
None	108	95	32	37	6

Source: See supplementary online Appendix.

Appendix Table A1
Goodness of Fit of First-Stage Estimates
Distance Distribution of Shipments Conditioned on Shipments Being at Least 100 miles
Semi-log Specification for 1997
Split up by Diffuse and Nondiffuse Industries (as defined in Appendix A.2)

Statistic Reported	Category 1 100 ≤ distance < 500		Category 2 500 ≤ distance < 1,000		Category 3 1,000 ≤ distance	
	Diffuse	Nondiffuse	Diffuse	Nondiffuse	Diffuse	Nondiffuse
Mean *share*$_{c,i}$ Across Industries (Data)	.45	.48	.30	.30	.25	.23
Mean *share*$_{c,i}$ Across Industries (Model)	.38	.37	.32	.33	.30	.29
Regression of *share*$_{c,i}$ (Data) on *share*$_{c,i}$ (Model)						
Intercept	.06 (.02)	.11 (.02)	-.02 (.02)	.00 (.02)	-.03 (.01)	-.05 (.01)
Slope	1.02 (.04)	.97 (.05)	.97 (.07)	.89 (.07)	.96 (.03)	.95 (.04)
R^2	.81	.58	.56	.36	.83	.70
Number of Industries in Regression	167	294	167	294	167	294

Source: Authors' calculations with confidential Census data.