Supplementary Web Appendix Materials for "An Alternative Theory of the Plant Size Distribution, with Geography and Intra- and International Trade"

by

Thomas J. Holmes (University of Minnesota, Federal Reserve Bank of Minneapolis, and NBER)

and

John J. Stevens (Board of Governors of Federal Reserve System)

August 2012 (Revised October 26, 2013)

Note: 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.

This document contains:

1. The details on the first-stage estimation algorithm, for how we solve for the vector of cost efficiencies Γ.
2. The calculations used to derive our measure of internal distance (equation (8) in paper).
3. An analysis of the issue of wholesaling and local shipments in the use of CFS data.
4. Details about how we calculate the mean location quotients reported in Tables 1 and 2.
5. Details about how we calculate the 90/10 ratio of plant employment and the variance decomposition, as reported in Section 2.

1 First-Stage Estimation Algorithm

We provide a detail about the first-stage estimation algorithm. As explained in the text, conditioned on the parameters η of the distance discount, we solve for a vector of cost

efficiencies to exactly match the distribution of sales revenues across locations. From the equations derived earlier, total sales revenue at location i equals

$$y_i = \sum_{\ell=1}^{L} \frac{\gamma_i a_{\ell i}}{\sum_{k=1}^{L} \gamma_k a_{\ell k}} x_\ell,$$

where the $a_{\ell i}$ depend upon η, held fixed here. Given the y_i, x_ℓ, $a_{\ell i}$, we need to find a vector $\Gamma = (\gamma_1, \gamma_2, ..., \gamma_{177})$, with the normalization $\gamma_1 = 1$, that solves these equations.

Define ω_i by $\omega_i \equiv \frac{1}{\gamma_i}$ and $\omega = (1, \omega_2, \omega_3, ..., \omega_L)$. Define the mapping

$$f_i(\omega) = \frac{1}{y_i} \sum_{\ell=1}^{L} \frac{a_{\ell i}}{\sum_{k=1}^{L} \frac{1}{\omega_k} a_{\ell k}} x_\ell.$$

If we find a ω such that $f_i(\omega) = \omega_i$ for all $i \geq 2$, then the output equations all hold.

Note first that $f_i(\omega)$ is strictly increasing. Note second that $f_i(\omega)$ is bounded above by

$$\bar{f}_i = \frac{1}{y_i} \sum_{\ell=1}^{L} \frac{a_{\ell i}}{a_{\ell 1}} x_\ell,$$

which is the limit of $f_i(\omega)$ as ω goes to infinity. Suppose we have a point $\omega^\circ > f(\omega^\circ)$. Define the sequence $\{\omega_0, \omega_1, \omega_2 ...\}$ by $\omega_0 = \omega^\circ$, and $\omega_t = f(\omega_t)$. Since $f(\cdot)$ is monotonically increasing and is bounded from above, this sequence converges to a solution.

To run the algorithm, we need a starting point $\omega^\circ > f(\omega^\circ)$. In our estimation, the following procedure worked for finding a starting value. Without loss of generality, we label locations for a particular industry so that location 1 has the maximum sales revenue share, $y_1 = \max\{y_1, y_2, ..., y_{177}\}$. Then we set $\omega_i^\circ = \frac{1}{y_i}\lambda$ for some small λ.

If $f(\omega)$ is convex, then any solution $\omega = f(\omega)$ is unique. We have shown that $f(\omega)$ is convex for $L = 2$ or $L = 3$ locations.

2 Calculation of Internal Distance

In the text we report a formula for internal distance (equation (8)), that we restate here for convenience.

$$\hat{d}_{r,r} \equiv \sum_{\ell' \in \Lambda_r} \sum_{\ell^\circ \in \Lambda_r} \frac{\gamma_{\ell^\circ}}{\bar{\gamma}_r} \frac{x_{\ell'}}{\bar{x}_r} d_{\ell' \ell^\circ}. \tag{1}$$

We claimt using this measure of internal region-level distance is valid as a first-order approximation, when distances within a region are small relative to distances across regions. To formalize this claim, we first parameterize the distance between any given pair of locations ℓ° and ℓ' as follows:

$$\begin{aligned} d_{\ell^\circ \ell'} &= m\delta_{r,\ell',\ell^\circ}, \text{ if } \ell^\circ \text{ and } \ell' \text{ are in the same region } r, \\ &= \bar{d}_{r',r^\circ}, \text{ if } \ell^\circ \text{ and } \ell' \text{ are in different regions } r^\circ \text{ and } r'. \end{aligned} \tag{2}$$

The parameter m scales up distances proportionately between two locations in the *same* region. Distances between pairs of locations from two *different* regions remain fixed, depending only on the pair of regions, not the particular locations within the regions. (That is, the distance from one location in the New York City region, to anywhere in the Los Angeles region, is the same as the distance from another location in New York City to anywhere in Los Angeles.) This approximation simplifies the exposition. When the scaling parameter m in (2) is small, internal distances within a region are small, compared to external, cross-region distances.

In the text, we first define $\bar{\phi}_{r',r^\circ}$ as the region-level aggregation of the location-level shipment shares,

$$\bar{\phi}_{r',r^\circ} = \sum_{\ell' \in \Lambda_{r'}} \frac{x_{\ell'}}{\bar{x}_{r'}} \left(\sum_{\ell^\circ \in \Lambda_{r^\circ}} \phi_{\ell',\ell^\circ} \right), \tag{3}$$

Next we define $\hat{a}_{r,r} \equiv a(\hat{d}_{r,r})$,and $\hat{a}_{r',r^\circ} = a(\bar{d}_{r',r^\circ})$, for $r^\circ \neq r'$, and use these to construct

$$\hat{\phi}_{r',r^\circ} \equiv \frac{\hat{a}_{r'r^\circ}\bar{\gamma}_{r^\circ}}{\sum_{r=1}^{R} \hat{a}_{r',r}\bar{\gamma}_r}, \tag{4}$$

which is the shipment share of a region-level model with internal distance $\hat{d}_{r,r}$ in region r. Our claim is that $\hat{\phi}_{r',r^\circ}$ approximates $\bar{\phi}_{r',r^\circ}$, near $m = 0$. Now write $\bar{\phi}_{r',r^\circ}(m)$ and $\hat{\phi}_{r',r^\circ}(m)$ as functions of m. It is immediate that

$$\bar{\phi}_{r',r^\circ}(0) = \hat{\phi}_{r',r^\circ}(0).$$

This follows because at $m = 0$, there is no internal distance within a region, and this is

equivalent to there being a single location in each region (so region-level and location-level coincide).

Next, we calculate the slope of both functions at $m = 0$, starting with the slope of $\hat{\phi}_{r,r}(m)$,

$$\frac{d\hat{\phi}_{r,r}(0)}{dm} = \left(\frac{\bar{\gamma}_r}{\sum_{r''=1}^{R} \bar{a}_{r''r}\bar{\gamma}_{r''}} - \frac{\bar{\gamma}_r\bar{\gamma}_r}{\left(\sum_{r''=1}^{R} \bar{a}_{r''r}\bar{\gamma}_{r''}\right)^2} \right) \frac{d\hat{a}_{rr}(0)}{dm}, \tag{5}$$

where

$$\frac{d\hat{a}_{rr}(0)}{dm} = a'(0)\frac{d\left(\hat{d}_{r,r}\right)}{dm} = a'(0)\sum_{\ell' \in \Lambda_r}\sum_{\ell^\circ \in \Lambda_r} \frac{\gamma_{\ell'}}{\bar{\gamma}_r}\frac{x_{\ell^\circ}}{\bar{x}_r}\delta_{\ell'\ell^\circ}.$$

To calculate the slope of $\bar{\phi}_{r',r^\circ}(m)$, we first need to calculate the slope of the location-level shipment share, for two locations ℓ' and ℓ° within the same region r. Recall ϕ_{ℓ',ℓ°,

$$\phi_{\ell',\ell^\circ} = \frac{a(m\delta_{r,\ell'\ell^\circ})\gamma_{\ell'}}{\sum_{\ell'' \in \Lambda_r} a(m\delta_{r,\ell''\ell^\circ})\gamma_{\ell'} + \sum_{r' \neq r} \bar{a}_{r,r'}\bar{\gamma}_{r'}}.$$

The slope is

$$\frac{d\phi_{\ell',\ell^\circ}}{dm} = \frac{a'(0)\delta_{r,\ell^\circ\ell'}\gamma_{\ell'}}{(\sum_{r'} \bar{a}_{r,r'}\bar{\gamma}_{r'})} - \frac{\bar{a}_{r,r}\gamma_{\ell'}}{(\sum_{r'} \bar{a}_{r,r'}\bar{\gamma}_{r'})^2}a'(0)\sum_{\ell' \in \Lambda_r}\delta_{\ell''\ell^\circ}\gamma_{\ell''}$$

Thus

$$\begin{aligned}
\frac{d\bar{\phi}_{r,r}}{dm} &= \sum_{\ell^\circ \in \Lambda_r}\frac{x_{\ell^\circ}}{\bar{x}_r}\sum_{\ell' \in \Lambda_r}\left[\frac{a'(0)\delta_{r,\ell^\circ\ell'}\gamma_{\ell'}}{(\sum_{r'} \bar{a}_{r,r'}\bar{\gamma}_{r'})} - \frac{\gamma_{\ell'}}{(\sum_{r'} \bar{a}_{r,r'}\bar{\gamma}_{r'})^2}a'(0)\sum_{\ell' \in \Lambda_r}\delta_{\ell''\ell^\circ}\gamma_{\ell''}\right] \\
&= a'(0)\sum_{\ell^\circ \in \Lambda_r}\frac{x_{\ell^\circ}}{\bar{x}_r}\sum_{\ell' \in \Lambda_r}\left[\frac{\delta_{r,\ell^\circ\ell'}\gamma_{\ell'}}{(\sum_{r'} \bar{a}_{r,r'}\bar{\gamma}_{r'})} - \frac{\gamma_{\ell'}}{(\sum_{r'} \bar{a}_{r,r'}\bar{\gamma}_{r'})^2}\sum_{\ell'' \in \Lambda_r}\delta_{\ell''\ell^\circ}\gamma_{\ell''}\right] \\
&= a'(0)\sum_{\ell^\circ \in \Lambda_r}\frac{x_{\ell^\circ}}{\bar{x}_r}\sum_{\ell' \in \Lambda_r}\frac{\delta_{r,\ell^\circ\ell'}\gamma_{\ell'}}{(\sum_{r'} \bar{a}_{r,r'}\bar{\gamma}_{r'})} - a'(0)\sum_{\ell^\circ \in \Lambda_r}\frac{x_{\ell^\circ}}{\bar{x}_r}\sum_{\ell' \in \Lambda_r}\frac{\gamma_{\ell'}}{(\sum_{r'} \bar{a}_{r,r'}\bar{\gamma}_{r'})^2}\sum_{\ell'' \in \Lambda_r}\delta_{\ell''\ell^\circ}\gamma_{\ell''} \\
&= a'(0)\sum_{\ell^\circ \in \Lambda_r}\sum_{\ell' \in \Lambda_r}\frac{x_{\ell^\circ}}{\bar{x}_r}\frac{\delta_{r,\ell^\circ\ell'}\gamma_{\ell'}}{(\sum_{r'} \bar{a}_{r,r'}\bar{\gamma}_{r'})} - a'(0)\sum_{\ell^\circ \in \Lambda_r}\frac{x_{\ell^\circ}}{\bar{x}_r}\frac{\bar{\gamma}_r}{(\sum_{r'} \bar{a}_{r,r'}\bar{\gamma}_{r'})^2}\sum_{\ell'' \in \Lambda_r}\delta_{\ell''\ell^\circ}\gamma_{\ell''}
\end{aligned} \tag{6}$$

This equals (4), completing the proof.

2.1 Calculation of Internal Distance Within a Census Track

Use use Census 2000 population by tract to calculate the measure of internal distance. The Census reports centroids for Census tracts and we use these to calculate $d_{\ell'\ell^\circ}$ within the same region, for $\ell' \neq \ell^\circ$. For $\ell' = \ell^\circ$, we proceed as follows. We attempt to estimate the expected distance between two randomly selected individuals. If individuals are uniformly distributed across the tract, and if the tract is a square, then the expected distance approximately equals

$$E[d_{\ell'\ell^\circ}] \approx sqrt(land_{\ell^\circ}) \times .521,$$

where $land_{\ell^\circ}$ is the area of tract ℓ°.

For the log-log specification, we need the expected log distance between two randomly selected individuals. The calculation is analogous to above, except we use

$$\begin{aligned} E[\ln d_{\ell^\circ\ell'}] &\approx \ln d_{\ell^\circ\ell'}, \text{ if } \ell^\circ \neq \ell', \\ &\approx \ln(sqrt(land_{\ell^\circ})) - .350, \text{ if } \ell^\circ = \ell'. \end{aligned}$$

3 Local Shipments and Wholesaling

When estimating the model in the first stage, we condition on sales being greater than 100 miles. Here we discuss the issue of "excess local shipments" in a little more detail. As this is the way we did things initially, we restrict attention to the subset of industries with diffuse demand. We have no reason to believe the results would be different for the complete set of industries.[1]

Table S1 reports the distribution of shipment shares by distance category, both in the data and in the estimated models, conditioned on shipments above 100 miles. In the goodness of fit discussion in the paper, presented in Table A1 of the appendix, the information is reported for all diffuse demand industries together. Here in Table S1 we also include a breakdown of industries by differences in transportation costs. In particular, we calculate the distance

[1] We report results based on the model estimates from the 2010 working paper version. In this earlier version, we used the semi-log specification for all industries and set the internal distance within an economic area to zero. The difference between the model estimates in the final version and the earlier estimates are negligible because the internal distances within economic areas are small and because we use the semi-log instead of the log-log specification in most cases in the final version of the paper.

adjustment $a_i(100)$ at 100 miles for the particular industry i and group industries by $a_i(100)$ categories.[2] We see in the conditional distribution a tendency for the model to overstate distance shipped, compared with the data. However, for the most part, the discrepancy is moderate.

Table S2 uses the same estimated model (the model optimized to fit the conditional distribution) and reports how things look in the unconditional distribution. A large discrepancy in the under 100 miles category is readily apparent. In particular, consider the highest $a_i(100)$ industries, the ones that are most tradable. In the fitted model, local shipments less than 100 miles are predicted to be only 5 percent of sales, but in fact make up 19 percent of sales, a difference of almost a factor of four. In the limiting case where there is no discount adjustment (i.e., $a_i = 1$ at all distances), a plant's shipments to a particular city should be proportionate to the city's population. That is, the share of shipments going less than 100 miles should equal the share of the U.S. population within 100 miles. Many of the high a_i industries, like clothing, tend to be outside the major population centers, and that is why predicted shipments within 100 miles is so low for these industries.

In working with the CFS data, Hillberry and Hummels (2008) observe the prevalence of local shipments and, as an explanation, emphasize the role that intermediate goods might play. For some immediate goods, it may be efficient to ship to nearby downstream manufacturing plants for further processing. However, here we are focusing on diffuse demand industries. Here we have eliminated —through our use of the input-output tables to define diffuse-demand industries—those manufacturing industries that tend to ship downstream to other manufacturing plants. So the point about intermediate goods has less relevance for us than it would be without that selection.

We believe the wholesaling sector plays some role in the story. Shipments leaving plants may make stops at nearby warehouses before arriving to their ultimate destination, which may be thousands of miles away. In such a case, the shipment distance is recorded at less than 100 miles, rather than the true distance.

Table S3 provides evidence on the relevance of the wholesale sector, for our sample of manufacturing industries. The CFS is a sample of shipments leaving both manufacturing plants and wholesale plants. To link manufacturing shipments with wholesale shipments, we exploit the product-level information available in the CFS. Each shipment in the 1997

[2]Note that the conditional results in Table S1 are calculated from the unconditioned results in Table S2, i.e., we take the averages across industries in S2 and then calculate the conditional shares (as opposed to calculating the conditional share for each industry first and then taking averages).

CFS is classified by SCTG product code.[3] Define the *ton-miles* of a shipment to be the shipment's weight times its distance. For each SCTG product, we estimate the share of ton-miles originating out of wholesale plants rather than manufacturing plants. Next, for each NAICS industry, we use the manufacturing plants in the CFS to estimate the revenue share across different SCTGs. Finally, we use the SCTG sales shares to weight the SCTG-level ton-mile shares, to produce a NAICS-level Ton-Mile Wholesale Share. On average across the sample industries, about a quarter of ton-miles go through the wholesale sector. When we repeat the exercise for the 1992 SIC sample industries, the results are virtually the same.[4] We conclude that for our sample industries, the wholesale sector plays a large enough quantitative role for it to potentially be a factor in accounting for why there are more local shipments than are predicted by the model.

4 Details of How the Location Quotients are Calulated in Tables 1 and 2

To explain the measure, we begin with the the definition of the location quotient Q_i^{rev} at location i. Lety_i is the total sales revenue of producers located at i and x_i be the population share. Letting y and x be the aggregate totals, the revenue location quotient Q_i^{rev} is a location's share of sales revenue (i.e., production) over its share of expenditure (i.e., consumption),

$$Q_i^{rev} \equiv \frac{y_i/y}{x_i/x}. \tag{7}$$

for a given industry.

Analogous to Holmes and Stevens (2004), for each industry we sort locations (economic areas) by the location quotient from lowest to highest and then aggregate locations into 10 approximately equal-sized population-decile classes. This aggregation helps smooth the data. Let Q_d^{rev} be the location quotient of decile d. By definition, $Q_d^{rev} \leq Q_{d+1}^{rev}$. If all sales are concentrated in the top decile, then $Q_{10}^{rev} = 10$, as 100 percent of the industry is concentrated among 10 percent of the population.

We are interested in comparing the geographic dispersion of different groups of plants

[3] This is the Standard Classification of Transported Goods code.

[4] The 1992 CFS uses the Standard Transportation Commodity Codes (STCC) for product codes rather than the SCTG product codes. Despite this difference, we get no change in results.

within the same industry. Let g index a particular group of plants. (For example, for what we do in Table 2, the index g signifies whether a plant is SIC/Retail or SIC/Man.) Suppose plants located in decile d of type g are indexed by k, and let $y_{d,g,k}$ be the sales of plant k of type g at decile d. Let $\bar{Q}^{rev}$ be the sales-weighted overall mean location quotient across plants from all locations of all types. Then

$$\bar{Q}^{rev} \equiv \frac{\sum_d \sum_g \sum_k y_{d,g,k} Q_d^{rev}}{\sum_d \sum_g \sum_k y_{d,g,k}} = \frac{\sum_d y_d \frac{\frac{y_d}{y}}{\frac{1}{10}}}{y} = 10 \left[\sum_{d=1}^{10} \left(\frac{y_d}{y} \right)^2 \right].$$

Hence, the mean location quotient $\bar{Q}^{rev}$ is exactly the standard Herfindahl index of concentration, times a factor of 10. If the entire industry is concentrated in the top decile, then $\bar{Q}^{rev} = 10$. If it is spread equally across the 10 deciles, then $\bar{Q}^{rev} = 1$. The main interest for this subsection is the *conditional mean location quotient* of plants of type g,

$$\bar{Q}_g^{rev} = \frac{\sum_d \sum_k y_{d,g,k} Q_d^{rev}}{\sum_d \sum_k y_{d,g,k}}.$$

Note that conditioning on type g enters only through the weights; plants of all types in decile d are used to define the Q_d^{rev} associated with a sale.[5] Conceptually, we are taking each dollar of sales in the data and associating it with the location quotient of its origin and taking means. This is the statistic that is reported in Tables 1 and 2.

5 Calculation of the 90/10 Ratio and the Variance Decomposition in Section 2

We use the public tabulations of employment size cell counts by industry to estimate the 90/10 ratio and variance decompositions reported in Table 1. The published tabulations report count for the following size classes:0–4, 5–9, 10–19, 20–49, 50–99, 100–249, 250–499, 500–999, 1,000–2,499, 2,500 and above.

To calculate the 90/10 ratio, we proceed as follows. For each cell, we allocate the plant counts uniformly across the various employment sizes contained within the range of the cell.

[5] In Holmes and Stevens (2002) we calculate an analogous measure that for each plant excludes the plant's own contribution to the location quotient and only uses the neighboring plants. This correction makes little difference in what we do here.

Note that we allocate plant counts to the "employment equal 0" category since there are plant in the data with employment at this level. We define the 10 percentile employment to be the employment level emp^{10} such that at least 10 percent of all plants have employment less than or equal to emp^{10}, while less then 10 percent of plants have employment less than $emp^{10}-1$. Analogously we define emp^{90}. Note we truncate the "2,500 and above" category at 9,999 employees (this is only relevant for calculating the emp^{90} for a few industries). The 90/10 ratio equals emp^{90}/emp^{10}. In cases where $emp^{10}=0$, we replace $emp^{10}=1$ in the formula.

To calculate the variance decomposition of log employment, we the average employment in each cell, take logs, and use this as an estimate of log employment of each plant in the cell.

References

Hillberry, Russell, and David Hummels. 2008. "Trade Responses to Geographic Frictions: A Decomposition Using Micro-Data." *European Economic Review* 52(3), 527–50.

Holmes, Thomas J., and John J. Stevens. 2002. "Geographic Concentration and Establishment Scale." *Review of Economics and Statistics* 84(4), 682–90.

Holmes, Thomas J., and John J. Stevens. 2004. "Spatial Distribution of Economic Activities in North America." In *Handbook of Regional and Urban Economics*, vol. 4, *Cities and Geography*, ed. J. Vernon Henderson and Jacques-François Thisse, 2797–843. Amsterdam: North-Holland.

Supplementary Appendix Table S1
Mean Share of Shipments in Data and Model
Conditioned on Distance Shipped over 100 Miles
Averages across Sample Industries and by Industry Distance Adjustment

Industry Grouping	Number of Industries	Distance Shipped		
		100 to 500	500 to 1,000	Over 1,000
Data, All Sample Industries	172	0.43	0.30	0.26
Model, All Sample Industries	172	0.38	0.33	0.30
Data, $a(100)<.5$	15	0.89	0.09	0.03
Model, $a(100)<.5$	15	0.77	0.17	0.06
Data, $.5 \leq a(100) < .75$	31	0.57	0.27	0.16
Model, $.5 \leq a(100) < .75$	31	0.48	0.31	0.20
Data, $.75 \leq a(100) < .9$	73	0.41	0.32	0.27
Model, $.75 \leq < a(100) < .9$	73	0.35	0.34	0.31
Data, $.9 \leq a(100)$	53	0.34	0.32	0.34
Model, $.9 < a(100)$	53	0.28	0.34	0.38

Supplementary Appendix Table S2
Unconditioned upon Shipments above 100 Miles

Industry Grouping	Number of Industries	Distance Shipped			
		Under 100	100 to 500	500 to 1,000	Over 1,000
Data, All Sample Industries	172	0.27	0.32	0.22	0.19
Model, All Sample Industries	172	0.11	0.34	0.29	0.27
Data, $a(100)<.5$	15	0.65	0.31	0.03	0.01
Model, $a(100)<.5$	15	0.39	0.47	0.10	0.04
Data, $.5 \leq a(100) < .75$	31	0.33	0.39	0.18	0.11
Model, $.5 \leq a(100) < .75$	31	0.14	0.42	0.27	0.18
Data, $.75 \leq a(100) < .9$	73	0.22	0.32	0.25	0.21
Model, $.75 \leq < a(100) < .9$	73	0.08	0.33	0.32	0.28
Data, $.9 \leq a(100)$	53	0.19	0.28	0.26	0.28
Model, $.9 < a(100)$	53	0.05	0.26	0.33	0.36

Supplementary Appendix Table S3
Estimates of Ton-Mile Wholesale Share
Averages across Sample Industries and by Industry Distance Adjustment

Industry Grouping	1997 NAICS Industry Sample		1992 SIC Industry Sample	
	Number of Industries	Mean of Ton-Mile Wholesale Share (Percent out of 100)	Number of Industries	Mean of Ton-Mile Wholesale Share (Percent out of 100)
All Sample Industries	172	24.4	175	24.2
By value of *a*(100) for industry				
Below .50	15	15.6	14	11.4
From .50 to .75	31	15.0	35	18.2
From .75 to .90	73	27.1	88	27.3
Above .90	53	28.7	38	27.3