Agglomeration and sorting in U.S. manufacturing

Andrea Stella

2025-031

Agglomeration and sorting in U.S. manufacturing

Andrea Stella[*]

Federal Reserve Board

April 8, 2025

Abstract

Using data on U.S. manufacturing plants, I estimate a production function model that includes agglomeration intensity as a component of total factor productivity and allows agglomeration benefits to vary across establishments, which can lead to sorting. I find that agglomeration benefits decline with unobserved establishment-level raw productivity.

JEL classifications: D22, D24, E24, L11, R11, R32

Keywords: Agglomeration, Sorting, Census of Manufactures.

Most importantly, however, cities differ in productivity: large cities produce more output per capita than small cities do. This urban productivity premium may occur because of locational fundamentals, because of agglomeration economies, because more talented individuals sort into large cities, or because large cities select the most productive entrepreneurs and firms.

Behrens and Robert-Nicoud, *Handbook of Regional and Urban Economics*, 2015

1 Introduction

It is striking that roughly 80% of the U.S. population lives on just 5% of the country's land. Such extreme spatial concentration of people and industry suggests that cities confer significant advantages. A natural explanation is that agglomeration – the clustering of firms and workers – boosts productivity. Indeed, dense urban regions are observed to be more productive, on average, than less dense ones, but why this "urban productivity premium" exists has been a subject of debate. Economists have proposed several mechanisms: true agglomeration economies, whereby proximity yields external benefits such as labor pooling, knowledge sharing, and quicker, cheaper access to intermediate goods and services (Combes and Gobillon 2015); selection effects, where only high-productivity firms survive the tough competition of cities (Combes et al. 2012); and sorting, in which inherently more-productive firms and workers choose to locate in big cities (Combes et al. 2012). Disentangling these forces is crucial for understanding the benefits of city size. Are big cities productive because they make firms more efficient or because they attract the most efficient firms? This paper tackles that question with new evidence from U.S. manufacturing.

While agglomeration economies and selection effects have been extensively studied in the urban literature, the role of firm sorting has received far less empirical attention. If high-productivity firms systematically sort into dense regions, then the observed productivity advantage of cities might be a composition effect rather than a pure agglomeration effect. Conversely, if lower-productivity firms benefit more from being in clusters, it would mean that agglomeration acts to narrow productivity gaps, casting doubt on sorting as the primary explanation. This paper fills that gap in the literature by examining whether agglomeration benefits vary across establishments in a way that could induce sorting.

To answer this question, I develop an empirical framework that allows agglomeration's effect on productivity to differ for each establishment. Specifically, I propose a

model of establishment productivity that includes an interaction between local agglomeration intensity and the firm's unobserved productivity. In essence, the production function is specified such that agglomeration economies can depend on a plant's individual efficiency. This flexible, heterogeneous approach permits me to identify whether more-efficient firms experience larger or smaller productivity boosts from being in a dense manufacturing region. For example, if agglomeration benefits were higher for more-productive establishments, those firms would have a stronger incentive to locate in areas with greater industry density – a pattern I would interpret as evidence of positive sorting. On the other hand, if the benefits of agglomeration are greater for less-productive plants, it would imply an opposite sorting pattern (or perhaps no sorting), where high-productivity firms are not disproportionately drawn to clusters by productivity advantages.

Identifying sorting effects is challenging because firms' underlying productivity is not directly observable to the researcher. I tackle this challenge by extending the control function approach of Olley and Pakes (1996) to my setting. In practice, this means I use a production function estimation method that controls for unobserved productivity shocks while allowing those shocks to interact with agglomeration intensity. By doing so, I correct for the fact that more-productive firms might endogenously choose higher inputs and favorable locations. This approach allows me to recover each establishment's intrinsic productivity and to estimate the elasticity of output with respect to local agglomeration for firms at different points in the productivity distribution. In short, my empirical strategy disentangles genuine agglomeration economies from the sorting of firms by productivity, using panel data on manufacturing plants and modern econometric techniques to address bias.

Using confidential microdata on U.S. manufacturing establishments from the U.S. Census Bureau, I find clear evidence of heterogeneous agglomeration benefits. In my estimates, agglomeration boosts productivity more for low-productivity plants than for high-productivity plants. Quantitatively, when moving from a plant at the 25th percentile of the productivity distribution to one at the 75th percentile, the output elasticity with respect to local agglomeration declines by about 10 to 40 percentage points. In other words, a relatively less-efficient factory sees a much larger productivity gain from being located in a dense manufacturing cluster than a highly efficient factory does. I interpret this finding as evidence against strong positive sorting being the driver of the urban productivity premium. If anything, the pattern suggests that density levels the playing field: It helps weaker firms catch up more than it helps the superstar firms. This casts doubt on firm sorting as the primary explanation for the positive correlation

between density and productivity observed in U.S. manufacturing.

This paper contributes to and expands the small but growing literature on firm sorting and agglomeration. There are only a few empirical papers that directly examine sorting, and their findings have differed. Forslid and Okubo (2014), for example, develop a theoretical model of spatial sorting with heterogeneous firms and two regions, and they find a non-monotonic sorting pattern: Firms with very high productivity (and high capital intensity) and firms with very low productivity (low capital intensity) tend to relocate to the larger region. In contrast, Gaubert (2018) estimates a richer general equilibrium model using French firm-level data and finds evidence of a positive interaction between agglomeration and firm productivity, meaning that more-productive firms gain more from locating in big cities. My results for U.S. manufacturing suggest the opposite interaction – high-productivity firms appear to gain less – thereby providing a new perspective on this question. Methodologically, my approach, which is based on a partial equilibrium production function model, requires fewer structural assumptions than the general equilibrium models used in prior work. In particular, I am able to let the data speak on how inputs and productivity jointly determine output in clustered versus non-clustered settings rather than imposing calibrated parameters. Unlike Gaubert (2018), who computes total factor productivity (TFP) as a residual and calibrates certain elasticities before assessing agglomeration effects, I jointly estimate productivity and agglomeration elasticities within a unified framework. This approach avoids potential biases from misspecification and provides a direct, data-driven estimate of the extent of sorting.

The findings have important implications for urban economic policy and industrial location decisions. Policymakers often promote local "cluster" development or offer incentives for firms to locate in their city, aiming to harness agglomeration spillovers to boost regional growth. My results suggest that such place-based policies should consider the types of firms they target. If smaller or less-productive manufacturers reap the largest productivity gains from clustering, then policies that support the formation of industrial hubs – for instance, providing shared infrastructure, facilitating supplier networks, or establishing industrial parks – could disproportionately raise the productivity of those firms and help lagging regions. On the other hand, if highly productive firms experience little additional boost from being in an already dense area, expensive tax breaks to attract a superstar firm might yield fewer spillover benefits than expected. Understanding the sorting dynamic also guides individual firms' location decisions: A moderately productive plant might see substantial gains from moving into a major industrial center, whereas a very productive plant may find that it can operate nearly as

efficiently in a smaller city or peripheral area without losing much in terms of agglomeration benefits. In summary, recognizing who benefits most from agglomeration can lead to more targeted and efficient urban economic policies, ensuring that efforts to stimulate local manufacturing yield the maximum possible improvement in productivity and competitive advantage.

The remainder of the paper is organized as follows. Section 2 describes the empirical model and identification strategy. Section 3 introduces the data used in the estimation. Finally, section 4 presents the empirical results and discusses their implications for agglomeration and sorting in manufacturing.

2 Empirical model

The production function of an establishment takes the following form in logs:

$$v_{it} \;=\; \beta_0 + \beta_L l_{it} + \beta_K k_{it} + \tau_{it} + \epsilon_{it}, \tag{1}$$

where i refers to the establishment, t is time, v_{it} is value added, l_{it} is labor, k_{it} is capital, τ_{it} is TFP, and ϵ_{it} is an i.i.d shock not observed by the establishment before making input decisions in time t and never observed by the econometrician. I assume that TFP is a function of the agglomeration intensity of the location where the establishment is based, a_{it}, and establishment-specific raw efficiency, ω_{it}. More specifically, I assume the following: $\tau_{it} = \beta_a a_{it} + \omega_{it} + \gamma a_{it}\omega_{it}$. ω_{it} represents the current and past productivity shocks that the establishment observes before making input decisions in t; the econometrician never observes ω_{it}.[1]

I measure agglomeration intensity as the number of employees of other establishments in the same sector that are within 10 kilometers from establishment i; more precisely, a_{it} is the logarithm of agglomeration intensity plus one.[2] There is evidence in the literature that agglomeration delivers economically and statistically significant benefits to productivity. I am interested in testing whether such agglomeration benefits vary with the establishment's raw efficiency.

---

[1]The assumption that agglomeration benefits are Hicks neutral is consistent with the literature; see Rosenthal and Strange (2004).

[2]Since taking the logarithm of a variable plus one has been shown to create bias in several estimation models, as a robustness check I estimated an alternative production function model that is a simple Cobb–Douglas when agglomeration intensity is zero and is the baseline model, equation (1), when agglomeration intensity is strictly higher than zero; in this alternative robustness model, a_{it} is equal to the logarithm of agglomeration intensity instead of the logarithm of agglomeration intensity plus one, as in the baseline model. The results are qualitatively and quantitatively very similar.

As shown by the algebra in (2), β_0 and β_a are not separately identified from the level of ω. I replace ω with $\bar{\omega} - c$, where $\bar{\omega} = \omega + c$, and I obtain an identical equation with $\bar{\omega}$, $\bar{\beta}_0$, and $\bar{\beta}_a$ replacing ω, β_0, and β_a, where $\bar{\beta}_0 = \beta_0 - c$ and $\bar{\beta}_a = \beta_a - c\gamma$.

$$
\begin{aligned}
v_{it} &= \beta_0 + \beta_L l_{it} + \beta_K k_{it} + \beta_a a_{it} + \omega_{it} + \gamma a_{it}\omega_{it} + \epsilon_{it} \qquad (2) \\
&= \beta_0 + \beta_L l_{it} + \beta_K k_{it} + \beta_a a_{it} + (\omega_{it} + c - c) + \gamma a_{it}(\omega_{it} + c - c) + \epsilon_{it} \\
&= \beta_0 - c + \beta_L l_{it} + \beta_K k_{it} + (\beta_a - c\gamma)a_{it} + (\omega_{it} + c) + \gamma a_{it}(\omega_{it} + c) + \epsilon_{it} \\
&= \bar{\beta}_0 + \beta_L l_{it} + \beta_K k_{it} + \bar{\beta}_a a_{it} + \bar{\omega} + \gamma a_{it}\bar{\omega} + \epsilon_{it}
\end{aligned}
$$

Since β_0, β_a, and the level of ω are not separately identified, I replace ω_{it} with $\tilde{\omega}_{it} = \omega_{it} + \beta_0$ and β_a with $\tilde{\beta}_a = (\beta_a - \gamma\beta_0)$. The production function I estimate is then

$$
v_{it} = \beta_L l_{it} + \beta_K k_{it} + \tilde{\beta}_a a_{it} + \tilde{\omega}_{it} + \gamma a_{it}\tilde{\omega}_{it} + \epsilon_{it}. \qquad (3)
$$

Before going into the details of the estimation, it is useful to provide the intuition behind the estimation approach and why it is needed. Estimating the production function with an agglomeration term – equation (3) – poses a classic identification challenge: More-productive firms may choose higher inputs and potentially different locations, which can bias naive estimates of agglomeration benefits. To address this, I employ a control function approach in the spirit of Olley and Pakes (1996) and Levinsohn and Petrin (2003). This semiparametric estimator uses a firm's decision on intermediate input purchases as a proxy for its unobserved productivity shocks. The intuition is that, given certain regularity conditions, a more-productive establishment will employ more intermediate inputs, so observed intermediate input use can reveal the establishment's current productivity level to the econometrician.

I extend the traditional control function method to allow unobserved productivity to interact with an observed agglomeration measure. Practically, this means that total productivity – including unobserved productivity, local agglomeration intensity, and their interaction – is assumed to be a function of intermediate input choices. By doing so, I account for the possibility that firms may invest differently in intermediate inputs, depending on the external economies they experience in their location.

Under the reasonable assumptions I discuss next, I can recover the unobserved productivity term as a function of observable inputs and agglomeration intensity. This yields an estimated production function that includes the following: (i) conventional input elasticities for labor and capital, (ii) an agglomeration elasticity term, and (iii) an interaction term capturing how agglomeration's effect varies with firm productivity (the

parameter γ in my model). I verify that the conditions for identification discussed in the literature hold in my setting, lending credibility to my estimation of the interaction effect.

I will now go into the specifics of the estimation procedure, starting with the assumptions I make.

Assumption 1 *Information Set: The establishment's information set at t – that is, I_{it} – includes current and past raw efficiency $\{\omega_{i\tau}\}_{\tau=0}^{t}$ but does not include future raw efficiency $\{\omega_{i\tau}\}_{\tau=t+1}^{\infty}$. The transitory shocks ϵ_{it} satisfy $E[\epsilon_{it} \mid I_{it}] = 0$.*

Assumption 2 *First-Order Markov: Raw efficiency evolves according to a first-order Markov process. This distribution is known to firms and is stochastically increasing in raw efficiency: $p(\omega_{it+1}|I_{it}) = p(\omega_{it+1}|\omega_{it})$.*

Assumption 3 *Timing of Input Choices: Firms accumulate capital according to $k_{it} = \kappa(k_{it-1}|i_{it-1})$, where investment i_{it-1} is chosen in period $t-1$. Labor input is flexible and chosen in period t.*

Assumption 4 *Timing of Location Choices: Location choices by establishments are made when the establishment is set up. Establishments do not move. Establishments enter a location as start-ups and exit when they shut down. Location choices need to be made a period in advance and are therefore based on the previous-period information set I_{it-1}.*

Assumptions 1 through 3 are standard in the literature. Assumption 4 on location choices is driven by the data I use; in the Census' Longitudinal Business Database (LBD), the establishment identity is defined by its geographical location, so establishments do not change location.[3] Gaubert (2018) only relies on data from one year for the structural estimation – specifically, 2000; consequently, she also does not exploit location changes resulting from firms endogenously relocating to identify the sorting parameter.

Assumption 5 *Firms' intermediate input demand is given by $m_{it} = f(k_{it}, l_{it}, \tau_{it})$.*

Assumption 6 *Strict Monotonicity: $f(k_{it}, l_{it}, \tau_{it})$ is strictly increasing in τ_{it}.*

I assume that establishments decide what level of intermediate input to use based on their capital and productivity levels and that the intermediate input demand is strictly increasing in productivity. It should be noted that the non-separability between raw efficiency ω and agglomeration a implies that the intermediate input choice in this model

---

[3]As I identify the location of an establishment with its ZIP Code, it is important to note that ZIP Code boundaries can change over time. Therefore, while changes in location within the data are possible, they should not be driven by the establishment's endogenous choices.

is not necessarily strictly increasing in ω, as assumed in the previous literature. However, it seems reasonable to assume that it is strictly increasing in overall productivity τ; in other words, I am assuming, like Levinsohn and Petrin (2003), that more-productive firms, conditional on capital, use more intermediate inputs. The conditions for Assumption 6 are the same as in Levinsohn and Petrin (2003), with τ instead of ω. Following Assumption 6, I can invert f and express productivity τ as a function of capital and intermediate input use:

$$v_{it} = \beta_0 + \beta_l l_{it} + \beta_k k_{it} + f^{-1}(k_{it}, l_{it}, m_{it}) + \epsilon_{it} \tag{4}$$

$$= \Phi(k_{it}, l_{it}, m_{it}) + \epsilon_{it} \tag{5}$$

where

$$\Phi(k_{it}, l_{it}, m_{it}) = \beta_l l_{it} + \beta_k k_{it} + \tilde{\beta}_a a_{it} + \tilde{\omega}_{it} + \gamma a_{it} \tilde{\omega}_{it}. \tag{6}$$

I can estimate $\Phi(k_{it}, l_{it}, m_{it})$ in (5) by replacing $\Phi(k_{it}, l_{it}, m_{it})$ with a third-order polynomial in k_{it}, l_{it}, and m_{it}. To avoid the identification issues raised by Ackerberg et al. (2015) and Gandhi et al. (2020), I take advantage of the firm's first-order condition with respect to labor to identify β_l. It can be shown that

$$\beta_l = \frac{W_{it} L_{it}}{P_{it} Q_{it}} \frac{e^{\epsilon_{it}}}{E_{it} e^{\epsilon_{it}}}, \tag{7}$$

where $\frac{W_{it} L_{it}}{P_{it} Q_{it}}$ is the labor share. Estimating (5) produces an estimate of ϵ_{it}, which I use to estimate $\hat{\beta}_l$ using (7). Raw efficiency can then be derived from (6) as

$$\tilde{\omega}_{it} = \frac{\Phi(k_{it}, l_{it}, m_{it}) - \hat{\beta}_l l_{it} - \beta_k k_{it} - \tilde{\beta}_a a_{it}}{1 + \gamma a_{it}}. \tag{8}$$

Assumptions 2 and 3 imply that I can decompose raw efficiency into its conditional expectation at time $t-1$ and an innovation term η:

$$\tilde{\omega}_{it} = E(\tilde{\omega}_{it} | I_{t-1}) + \eta_{it}. \tag{9}$$

$E(\tilde{\omega}_{it} | I_{t-1})$ in (9) can be estimated non-parametrically with a third-order polynomial to retrieve the innovation term ϵ_{it}. ϵ_{it} and η_{it} are orthogonal to the information set in period $t-1$, I_{t-1}: $E[(\eta_{it} + \epsilon_{it}) | I_{t-1}] = 0$. This can be used to derive moment conditions and

estimate the model parameters with GMM. I use as instruments variables determined in $t-1$: capital and agglomeration from period t, labor and materials from period $t-1$, and interactions between them. The moment conditions are then as follows:

$$
E \left[(\hat{\eta}_{it}(\beta_k, \tilde{\beta}_a, \gamma) + \hat{\epsilon}_{it}(\beta_k, \tilde{\beta}_a, \gamma)) \otimes \begin{pmatrix} 1 \\ k_{it} \\ a_{it} \\ m_{it-1} \\ l_{it-1} \\ k_{it}a_{it} \\ k_{it}l_{it-1} \\ k_{it}m_{it-1} \\ a_{it}l_{it-1} \\ a_{it}m_{it-1} \\ l_{it-1}m_{it-1} \\ k_{it}a_{it}m_{it-1} \\ k_{it}a_{it}l_{it-1} \\ k_{it}m_{it-1}l_{it-1} \\ a_{it}m_{it-1}l_{it-1} \\ k_{it}a_{it}m_{it-1}l_{it-1} \end{pmatrix} \right] = 0. \tag{10}
$$

In Appendix A, I perform a brief Monte Carlo study, using a similar setup to Acker-berg et al. (2015), and find that my algorithm performs well.

3 Data description

The main data sources are the U.S. Census Bureau's Annual Survey of Manufacturers (ASM) and Census of Manufactures (CM). The model is estimated separately on seven four-digit sectors: 3111 (Animal Food Manufacturing), 3113 (Sugar and Confectionery Product Manufacturing), 3116 (Animal Slaughtering and Processing), 3118 (Bakeries and Tortilla Manufacturing), 3141 (Textile Furnishings Mills), 3152 (Cut and Sew Apparel Manufacturing), and 3162 (Footwear Manufacturing).[4] The sample period is 1991 to

---

[4]Since the estimation is very time consuming, I ran the model only on four-digit sectors within the NAICS 31 sector, with the exception of sectors ending in 9, which are residual groups that combine different types of establishments. I was not able to achieve identification for the sectors not included in Table 2, as some bootstrap simulations, or in a few cases the point estimates themselves, were economically

2019. I include in my sample only establishments that were included in the ASM with certainty during at least one survey year. Moreover, I drop observations with labor share below the 2.5th percentile and above the 97.5th percentile. After the cleaning, the resulting samples cover most of the value added in the original ASM samples.

I follow Foster et al. (2016) and Cunningham et al. (2023) in the computation of value added and the production function inputs: Nominal output is measured as total shipments and deflated using an industry-level measure from the NBER–CES Manufacturing Industry Database, capital is computed with a perpetual inventory method, labor is measured as total hours, materials are measured separately for physical materials and energy and each is deflated by an industry-level deflator, and, finally, value added is computed as output minus materials. I measure agglomeration intensity as the number of employees in the same four-digit sector within 10 kilometers. I merged the ASM–CM sample with the LBD to compute agglomeration intensity, as the LBD provides the location of all establishments in the United States during my sample period. The distances among establishments are computed as Euclidean distances among the ZIP Code centroids.

Table 1: PRELIMINARY EVIDENCE

	My dataset		2012 CM	
Log density	0.025		0.024	
	(0.002)		(0.002)	
Log total employment		0.016		0.025
		(0.002)		(0.003)
N. Obs.	28,500	28,500	2,900	2,900
R^2	0.005	0.002	0.036	0.032

Notes: Standard errors are reported in parentheses. N. Obs. is the rounded number of observations.

To get a first taste of the data, I ran regressions relating the density of an area to its productivity. Such regressions are popular in the literature; see Combes and Gobillon (2015). More precisely, I ran a regression of the mean log TFP of the establishments in a ZIP Code on either the log density of the ZIP Code or its log total manufacturing employment.[5] I ran the regressions on the dataset described above (including all seven four-digit sectors) and on the 2012 CM sample. As shown in Table 1, the agglomeration elasticity is estimated to be around 2.5%, on the lower end of the range of estimates in the literature. It must be noted that the R^2 statistics are low compared to similar models

---

nonsensical and the confidence intervals too large to make a precise inference possible. For sectors 3116 and 3162, I had to re-run a few bootstrap simulations to achieve identification.

[5]TFP is computed as in Foster et al. (2016) and Cunningham et al. (2023).

in other studies.

4 Empirical results and discussion

4.1 Main findings: heterogeneity in agglomeration benefits

Table 2 presents parameter estimates for two models: a simple Cobb–Douglas production function with only capital and labor (Column "CB"), and the full model (1) (Column "Full"). The table displays for each four-digit sector the coefficients on capital, β_k, labor, β_l, and the interaction between unobserved raw productivity and agglomeration intensity, γ.

Table 2: ESTIMATES

	Panel A							
	3111		3113		3116		3118	
	CB	Full	CB	Full	CB	Full	CB	Full
β_k	0.661	0.648	0.682	0.787	0.591	0.705	0.651	0.675
	(0.029)	(0.207)	(0.022)	(0.072)	(0.016)	(0.037)	(0.010)	(0.037)
β_l	0.203	0.203	0.248	0.248	0.299	0.299	0.303	0.303
	(0.004)	(0.004)	(0.006)	(0.006)	(0.004)	(0.004)	(0.004)	(0.004)
γ		-0.146		-0.473		-0.476		-0.667
		(0.121)		(0.183)		(0.358)		(0.128)
N. Obs.		11,500		4,800		17,000		16,000
N. Estab.		1,400		550		1,700		2,400
	Panel B							
	3141		3152		3162			
	CB	Full	CB	Full	CB	Full		
β_k	0.535	0.490	0.433	0.534	0.542	0.987		
	(0.026)	(0.055)	(0.016)	(0.049)	(0.049)	(0.220)		
β_l	0.345	0.345	0.434	0.434	0.415	0.415		
	(0.007)	(0.007)	(0.005)	(0.005)	(0.012)	(0.012)		
γ		-0.470		-0.618		-0.409		
		(0.124)		(0.124)		(0.205)		
N. Obs.		4,800		14,000		1,800		
N. Estab.		700		3,700		250		

Notes: The table shows the parameter estimates for 7 four-digit industries. CB stands for Cobb–Douglas, and Full refers to model (1). Bootstrapped standard errors are reported in parentheses. N. Obs. is the rounded number of observations. N. Estab. is the rounded number of unique establishments.

As shown in Table 2, the estimates of capital elasticity are a touch higher in the

full model than in the Cobb–Douglas model for most sectors. However, the confidence intervals are wide enough that the difference is not statistically relevant for most sectors. The elasticity of labor is, by construction, identical under the two models and is around 30%. The estimated interaction term (γ) between local agglomeration intensity and firm-level productivity is negative in virtually all specifications. This means that higher-productivity plants experience a smaller marginal gain from agglomeration than lower-productivity plants.[6]

To analyze the agglomeration side of the production function, I combine the estimates of $\tilde{\beta}_a$ and γ with the distribution of raw productivity $\tilde{\omega}$. This analysis is presented in Table 3, where I calculate the average output elasticity with respect to agglomeration, Avg. = $E(\beta_a + \gamma\omega)$, and its interquartile range, IQR = $\gamma(\omega_{75} - \omega_{25})$. Notably, the terms involving β_0 cancel out when computing both Avg. and IQR, allowing me to express them as functions of β_a and ω instead of their normalized counterparts $\tilde{\beta}_a$ and $\tilde{\omega}$.

The average agglomeration benefits ("Avg." in Table 3) appear to be quite small, as they are close to zero and statistically insignificant for most sectors. This contrasts with findings from Table 1 and the existing literature, which reports agglomeration externalities ranging from 1% to 12%; see Rosenthal and Strange (2004) and Combes and Gobillon (2015) on urbanization economies and Greenstone et al. (2010) for a similar production function setting. Even though the average agglomeration benefits appear modest, the IQR estimates in Table 3 reveal a broad variation, indicating that many firms derive substantial advantages from agglomeration. These findings serve as a caution against estimating agglomeration benefits without carefully identifying productivity and considering the interactions between productivity and agglomeration intensity.

Table 3: AGGLOMERATION ELASTICITY

	3111	3113	3116	3118	3141	3152	3162
Avg.	-0.652	0.003	-0.002	-0.012	0.000	0.000	0.013
	(0.201)	(0.027)	(0.024)	(0.089)	(0.026)	(0.036)	(0.107)
IQR	-0.419	-0.155	-0.112	-0.086	-0.184	-0.130	-0.322
	(0.107)	(0.060)	(0.024)	(0.039)	(0.022)	(0.050)	(0.196)

Notes: Avg. is equal to $E(\beta_a + \gamma\omega)$. IQR is equal to $\gamma(\omega_{75} - \omega_{25})$. Bootstrapped standard errors are reported in parentheses.

The interquartile range of the elasticity with respect to agglomeration intensity ("IQR" in Table 3) is always negative and statistically significant. The data strongly supports a

---

[6]Since interpreting $\tilde{\beta}_a$ is challenging because of the identification issue discussed in Section 2, I did not include it in Table 2.

negative interaction between establishment-level raw productivity and agglomeration benefits. Quantitatively, moving from a relatively low-productivity establishment (at the 25th percentile of the productivity distribution) to a high-productivity establishment (the 75th percentile) is associated with a decline in the agglomeration elasticity of output on the order of 10 to 40 percentage points. In other words, the establishments benefiting the most from a dense industrial environment tend to be the least productive ones, whereas the most-productive firms enjoy more modest gains from local agglomeration.

This finding is illustrated in Table 3 by the consistently negative estimates I obtain for the IQR difference in agglomeration elasticity across productivity levels. The negative and statistically significant IQR effect confirms that the production function is not log-supermodular in density and productivity, contrary to what one would expect if only the most productive firms reaped the largest external gains, as in Gaubert (2018). From an economic standpoint, this result is somewhat surprising: It suggests that smaller or less-efficient manufacturers rely more heavily on external economies of scale (e.g., shared suppliers and local knowledge) to improve their performance, whereas the industry leaders do so to a lesser extent. A possible explanation is that highly productive firms may already be operating near the technological frontier or have better internal processes, leaving less scope for external factors to further raise their productivity. In contrast, firms with lower innate productivity have more to gain from the networking, learning, and input-sharing opportunities that agglomeration provides. As a result, the productivity boost from the local industrial environment exhibits diminishing returns as a firm's own efficiency rises.

It is important to note that my evidence casts doubt on firm sorting by productivity as the primary explanation for the well-documented positive correlation between density and productivity. If the most productive firms were endogenously sorting into dense regions because they gain the most from those environments, I would expect to find a positive interaction term (higher agglomeration elasticity for higher-productivity firms). I find the opposite. Thus, while more-productive firms are indeed found in denser areas, on average, my results imply this pattern is not because those firms derive outsized productivity benefits from agglomeration. Other mechanisms must be at play to explain why high-productivity firms tend to be in cities.

4.2 Robustness checks

I now consider the robustness of these findings. I subject the core result – that agglomeration's benefits decline with firm productivity – to a variety of robustness checks.

These checks verify that my findings are not an artifact of a particular measurement of "agglomeration" or driven by specific modeling assumptions. First, I alter the geographic scope used to define agglomeration intensity. My baseline measure considers employment in the same four-digit NAICS industry within a 10-kilometer radius. I re-estimate the models using a narrower radius of 5 kilometers and a much wider radius of 50 kilometers. Intuitively, a 5-kilometer radius captures very localized clusters, while a 50-kilometer radius extends to a broader regional scale, where spillovers might dissipate.

Second, I change the industry definition of the agglomeration measure: Instead of counting only same-four-digit NAICS neighbors, I use a coarser three-digit NAICS classification (grouping more-related industries together) to define nearby activity. This addresses whether my results are sensitive to using a tight industry definition (localization economies) versus a slightly broader industry grouping (which might incorporate some cross-industry externalities).

Table 4: ROBUSTNESS ON IQR ESTIMATES

	3111	3113	3116	3118	3141	3152	3162
Five Kilometers	-*	-	-	-**	-**	+	-
Fifty Kilometers	-	+	+	-*	-**	+	-*
Three-digit agglom.	-	n.c.	-*	-*	-**	-**	-

Notes: ** denotes statistical significance at the 1% level, and * denotes statistical significance at the 5% level. "n.c." stands for not converged.

Table 4 reports the results of these robustness exercises. Because of data confidentiality constraints, I report only the sign and significance of some estimates rather than the exact magnitudes. Across these alternative specifications, the pattern remains consistent. The estimated interaction effect (expressed again as the difference in agglomeration elasticity between higher- and lower-productivity plants) stays negative in almost all cases. In the majority of industries and alternative definitions, the less-productive establishments continue to show higher returns to local density than their more-productive counterparts. The only notable deviations occur under the 50-kilometer radius definition, where the interaction effect is positive, albeit not statistically significant, in three sectors. This attenuation is expected – a very wide radius likely dilutes the true local externalities by including relatively distant activity that offers limited direct interaction or spillovers. In contrast, using the broader three-digit industry aggregation yields the same sign and significance of the productivity–agglomeration interaction as in the baseline specification.

14

4.3   Implications for theory and policy

These empirical findings carry important implications for theories of urban agglomeration and the interpretation of past results. The evidence suggests that the production function for manufacturing establishments is not supermodular in firm productivity and local agglomeration intensity. In theoretical terms, I do not observe the complementarity that would make high-productivity firms benefit disproportionately from being in a dense cluster. Such complementarity is a key assumption in models where "the best firms end up in the best places" through endogenous sorting. By rejecting this assumption, my results indicate that one should be cautious in attributing the productivity advantages of cities to a matching of inherently more-efficient firms with richer external environments.

It is useful to contrast my findings with those of Gaubert (2018), one of the few other empirical studies on firm sorting. Gaubert (2018) develops a general equilibrium model of heterogeneous firms across cities and, using French data, finds evidence of a positive interaction between firm productivity and agglomeration benefits. In her framework, more-productive firms see larger gains from locating in big cities, which in turn reinforces their incentive to sort into those locations. By contrast, my results for U.S. manufacturing show a negative interaction – implying that, in my data, highly productive firms do not enjoy larger marginal external benefits. How can we reconcile these differences? One explanation may be methodological. Gaubert's approach entails calibrating certain parameters (e.g., factor elasticities), computing productivity as a residual, and then matching moments in a structural model. If productivity is mismeasured or if input adjustments are not fully accounted for, one might mistakenly infer a positive sorting effect. My estimation, by directly controlling for input choices and jointly estimating productivity, avoids relying on pre-calculated TFP measures. This could lead to a more accurate (and in this case, lower) estimate of the productivity–density interaction. Another reason could be country differences: The industrial structure and spatial configuration in the United States might differ from France in ways that affect agglomeration economies.

The finding that agglomeration benefits decline with firm productivity has practical implications for regional economic policy, especially place-based development programs. As shown by Bartik (2020), governments at the federal, state, and local levels increasingly invest in place-based policies – spending on the order of $60 billion per year – to spur job creation and productivity growth in specific areas. Examples include enterprise zones, relocation incentives, cluster development grants, and infrastructure investments targeted at lagging regions. A common rationale behind these policies is the expectation of

agglomeration spillovers: By attracting firms (ideally high-performing ones) to a region, policymakers hope to create self-reinforcing cycles of growth through clustering effects.

My results offer a more nuanced perspective that can improve the design and evaluation of such policies by underscoring the importance of considering firm heterogeneity in policy implementation. If policymakers aim to maximize local productivity spillovers, they should recognize that attracting the most-productive firms may not yield the largest external benefits per firm. Instead, policies that support a mix of firms, especially those that can gain the most from density, may be more effective in fostering overall economic growth. A superstar manufacturing plant, while certainly beneficial for many reasons, might not experience a large productivity jump just from being in a given location if it is already very efficient. In contrast, a moderately productive plant could substantially improve its performance when placed in a dense industrial cluster by leveraging knowledge transfers, skilled labor pools, or supplier networks that it did not have access to before. This suggests that policies that are purely focused on luring marquee firms (for example, offering large tax breaks to big multinationals to locate in a depressed area) might not fully capitalize on agglomeration economies – their presence might not diffuse as much additional efficiency to themselves or others as one would hope. Meanwhile, supporting a critical mass of small and medium enterprises or slightly less-productive firms to co-locate and interact could potentially generate more substantial productivity improvements collectively because these firms are more responsive to the agglomeration stimulus.

References

Ackerberg, Daniel A, Kevin Caves, and Garth Frazer, "Identification properties of recent production function estimators," *Econometrica*, 2015, *83* (6), 2411–2451.

Bartik, Timothy J, "Using place-based jobs policies to help distressed communities," *Journal of Economic Perspectives*, 2020, *34* (3), 99–127.

Behrens, Kristian and Frédéric Robert-Nicoud, "Agglomeration theory with heterogeneous agents," *Handbook of regional and urban economics*, 2015, *5*, 171–245.

Biesebroeck, Johannes Van, "Robustness of productivity estimates," *The Journal of Industrial Economics*, 2007, *55* (3), 529–569.

Combes, Pierre-Philippe and Laurent Gobillon, "Chapter 5 - The Empirics of Agglomeration Economies," in Gilles Duranton, J. Vernon Henderson, and William C.

Strange, eds., *Handbook of Regional and Urban Economics*, Vol. 5 of *Handbook of Regional and Urban Economics*, Elsevier, January 2015, pp. 247–348.

___ , **Gilles Duranton, Laurent Gobillon, Diego Puga, and Sébastien Roux**, "The Productivity Advantages of Large Cities: Distinguishing Agglomeration From Firm Selection," *Econometrica*, November 2012, *80* (6), 2543–2594.

Cunningham, Cindy, Lucia Foster, Cheryl Grim, John Haltiwanger, Sabrina Wulff Pabilonia, Jay Stewart, and Zoltan Wolf, "Dispersion in Dispersion: Measuring Establishment-Level Differences in Productivity," *Review of Income and Wealth*, 2023, *69* (4), 999–1032.

Foster, Lucia, Cheryl Grim, and John Haltiwanger, "Reallocation in the great recession: cleansing or not?," *Journal of Labor Economics*, 2016, *34* (S1), S293–S331.

Gandhi, Amit, Salvador Navarro, and David A. Rivers, "On the identification of gross output production functions," *Journal of Political Economy*, 2020, *128* (8), 2973–3016.

Gaubert, Cecile, "Firm sorting and agglomeration," *American Economic Review*, 2018, *108* (11), 3117–3153.

Greenstone, Michael, Richard Hornbeck, and Enrico Moretti, "Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings," *Journal of Political Economy*, 2010, *118* (3), 536–598.

Levinsohn, James and Amil Petrin, "Estimating production functions using inputs to control for unobservables," *The Review of Economic Studies*, 2003, *70* (2), 317–341.

Olley, G Steven and Ariel Pakes, "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 1996, *64* (6), 1263–1297.

Rosenthal, Stuart S and William C Strange, "Evidence on the nature and sources of agglomeration economies," in "Handbook of regional and urban economics," Vol. 4 2004, pp. 2119–2171.

A   Online Appendix: Monte Carlo simulations

I now posit a data generating process to simulate data and test the estimation algorithm described in Section 2.

A.1   Data generating process

It is well outside the scope of this paper to develop a general equilibrium model of firm location choices with dynamic capital investment. Thus, I will replace agglomeration with an AR(1) process. Following Ackerberg et al. (2015), the production function is Leontief in materials:

$$Q_{it} = \min\{e^{\beta_0}K_{it}^{\beta_K}L_{it}^{\beta_L}e^{\tau_{it}}, \beta_M M_{it}\}e^{\epsilon_{it}} \tag{11}$$

where i refers to the establishment, t to time, L is labor, K is capital, M is materials, ϵ_{it} is an unobservable shock to production, and τ_{it} is productivity. I assume $\epsilon_{it} \sim N(0,\sigma_\epsilon^2)$ and constant returns to scale, $\beta_K + \beta_L = 1$. I assume that productivity is a function of both agglomeration, a_{it}, and idiosyncratic establishment-specific raw efficiency, ω_{it}. More specifically, I assume: $\tau_{it} = \beta_a a_{it} + \omega_{it} + \gamma a_{it}\omega_{it}$. The firm observes productivity at time t before making labor decisions. Capital is not flexible and evolves according to $K_{it+1} = (1 - \delta)K_{it} + I_t$. I also assume that firms' wages follow the process .

The establishment chooses labor maximizing revenues minus labor costs, where the price of output is set to 1:

$$\max_{L_{it}} E_t\{Q_{it} - L_{it}W_{it}\} \tag{12}$$

The first order condition is

$$E_t(\beta_0\beta_L K_{it}^{\beta_K}L_{it}^{\beta_L-1}e^{\tau_{it}}e^{\epsilon_{it}}) - W_{it} = 0 \tag{13}$$

which provides the optimal labor choice:

$$L = \beta_0^{\frac{1}{1-\beta_L}}\left(\frac{\beta_L}{W_{it}}\right)^{\frac{1}{1-\beta_L}}K_{it}^{\frac{\beta_K}{1-\beta_L}}e^{\frac{\tau_{it}}{1-\beta_L}}e^{0.5\frac{\sigma_\epsilon^2}{(1-\beta_L)^2}} \tag{14}$$

where I use that $E(e^{\alpha\epsilon}) = e^{0.5\alpha^2\sigma_\epsilon^2}$.

The investment choice is an intertemporal choice. The establishment maximizes the

discounted stream of profits. The discount factor is β and there are investment adjustment costs equal to $\frac{\phi}{2}I_{it}^2$.

$$\max_{I_{it}} E_t \sum_{t=0}^{\infty} \beta^t \{Q_{it} - \frac{\phi}{2}I_{it}^2\} \tag{15}$$

subject to the production function and the capital accumulation dynamics. Following Van Biesebroeck (2007), the first order conditions are as follows

$$-\phi I_{it} + E_t \beta [\frac{\partial Q_{it+1}}{\partial I_t} - \phi I_{it+1} \frac{\partial I_{t+1}}{\partial I_t}] = 0 \tag{16}$$

$$\phi I_{it} = \beta E_t [\beta_0^{\frac{1}{1-\beta_L}} \left(\frac{\beta_L}{W_{it+1}}\right)^{\frac{\beta_L}{1-\beta_L}} \beta_K K_{it+1}^{\frac{\beta_K}{1-\beta_L}-1} e^{\frac{\tau_{it+1}}{1-\beta_L}} e^{0.5 \frac{\beta_L \sigma_\epsilon^2}{(1-\beta_L)^2}} e^{\epsilon_{it+1}}] - \phi_i \beta E_t [-(1-\delta)I_t] \tag{17}$$

$$\phi I_{it} = \beta \beta_0^{\frac{1}{1-\beta_L}} \beta_K \beta_L^{\frac{\beta_L}{1-\beta_L}} E_t [e^{\frac{\tau_{it+1}}{1-\beta_L} - \frac{\beta_L}{1-\beta_L} ln W_{it+1} + 0.5 \frac{\beta_L \sigma_\epsilon^2}{(1-\beta_L)^2} + \epsilon_{it+1}}] + (1-\delta)\beta E_t [\phi_i I_{t+1}] \tag{18}$$

Forward substituting the optimal investment choice in equation (18) I get

$$I_{it} = \frac{\beta \beta_K \beta_0^{\frac{1}{1-\beta_L}}}{\phi_i} \beta_L^{\frac{\beta_L}{1-\beta_L}} E_t \sum_{s=0}^{\infty} ((1-\delta)\beta)^s [e^{\frac{\tau_{it+1+s}}{1-\beta_L} - \frac{\beta_L}{1-\beta_L} ln W_{it+1+s} + 0.5 \frac{\beta_L \sigma_\epsilon^2}{(1-\beta_L)^2} + \epsilon_{it+1+s}}] \tag{19}$$

Since τ_i does not have a simple dynamic progression, the expectation in (19) will have to be numerically computed.

I will consider two cases. In Case A, a_{it} and ω_{it} evolve according to unrelated autoregressive processes:

$$\omega_{it} = \rho_\omega \omega_{it-1} + \xi_{it}^\omega \tag{20}$$

$$a_{it} = \rho_a a_{it-1} + \xi_{it}^a \tag{21}$$

where $\xi_{it}^\omega \sim N(0, \sigma_{\xi_\omega}^2)$ and $\xi_{it}^a \sim N(0, \sigma_{\xi_a}^2)$. In Case B, a_{it} is influenced by the previous period shock to raw productivity:

$$\omega_{it} = \rho_\omega \omega_{it-1} + \xi_{it}^\omega \tag{22}$$

$$a_{it} = \rho_a a_{it-1} + \alpha_a \xi_{it-1}^\omega + \xi_{it}^a \tag{23}$$

where α_a measures the responsiveness of a_{it} to the productivity shock.

A.2 Simulation results

I simulate 20,000 firms over 10 periods for 500 times.[7] The parameters of the model are calibrated similarly to Ackerberg et al. (2015). $\beta_0 = 2$, $\beta_L = 0.6$, $\beta_K = 0.4$, $\beta_a = 0.3$, and $\gamma = 0.2$. The depreciation rate $\delta = 0.2$, the discount rate $\beta = 0.95$, the standard deviation of the optimization error in labor $\sigma_{\tilde{\zeta}_l} = 0.37$, the variation in the capital adjustment cost ϕ is such that $Std(\phi) = 0.6$, the standard deviation of ϵ is 0.1, the AR parameters for raw productivity omega are $\rho_\omega = 0.7$ and $\sigma_{\tilde{\zeta}_\omega} = 0.2142$. In both Case A and B, $\rho_a = 0.3$ and $\sigma_{\tilde{\zeta}_a} = 0.7$, and in Case B $\alpha_a = 0.5$.

Table 5: Monte Carlo simulation results

DGP	Statistic	β_L	β_K	γ
Case A	Mean	0.6000	0.4003	0.2005
	St. Dev	0.0002	0.0028	0.0269
Case B	Mean	0.6000	0.4000	0.1990
	St. Dev	0.0002	0.0032	0.0344
Truth		0.6	0.4	0.2

Notes: The table shows the mean and standard deviation of the parameter estimates obtained simulating 20,000 firms over 10 periods for 500 times.

Table 5 shows the mean and standard deviation of the parameter estimates in the 500 Monte Carlo simulations; β_a is not shown because it cannot be identified as explained in Section 2. In both Case A and Case B, the estimation procedure is able to correctly and precisely estimate the elasticities β_L and β_K, and the agglomeration parameter γ.

Table 6: γ with equation (25)

DGP	Statistic	γ
Case A	Mean	0.2005
	St. Dev	0.0270
Case B	Mean	0.1989
	St. Dev	0.0344

Notes: See Table 5.

---

[7]The firms are simulated for 100 periods and then the first 90 periods are removed as burn-in.

Since $\tau_{it} = \beta_a a_{it} + \omega_{it} + \gamma a_{it}\omega_{it}$, the distribution of TFP, τ, conditional on agglomeration, a, depends on the parameters β_A and γ. If β_A is different from zero, the conditional distribution will be shifted depending on the level of agglomeration, but its variance will not be influenced. If γ is different from zero, both the mean and variance of the conditional distribution of TFP will change depending on agglomeration. The way agglomeration impacts the conditional distribution of TFP can help understand the identification of the two agglomeration parameters.

For instance, taking the variance of TFP conditional on agglomeration I obtain

$$V(\tau_{it}|a) = (1 + \gamma a)^2 V(\omega_{it}|a) \tag{24}$$

Equation (24) can be used to obtain γ as

$$\gamma = \frac{Std(\tau_{it}|a) - Std(\omega_{it}|a)}{a \times Std(\omega_{it}|a)} \tag{25}$$

Table 6 shows the γ estimates obtained with equation (25) after estimating τ and ω with the moment condition in (10); the agglomeration parameter γ is identified by its role in determining the distribution of TFP conditional on agglomeration. This is not an alternative way to estimate the agglomeration parameters, as the GMM estimation to obtain τ and ω also delivers the agglomeration parameters, but I find it useful to understand how the agglomeration parameters are identified within the model.