# Mathematical Appendix to "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns"

Glenn Ellison
MIT

Edward Glaeser
Harvard University

William Kerr[1]
Harvard Business School

January 2009

[1]email: gellison@mit.edu, glaeser@fas.harvard.edu, wkerr@hbs.edu.

# 1 Measurement of Coagglomeration

In this mathematical appendix we discuss an index of coagglomeration introduced in Ellison and Glaeser (1997). We note that the index takes on a simpler form when used to measure pairwise coagglomeration. We further develop the economic motivation for the index as a measure of the importance of cross-industry spillovers and shared natural advantages.

## 1.1 Background

Consider a group of industries indexed by $i = 1, 2, \ldots, I$. Suppose that a geographic whole is divided into $M$ subareas and suppose that $s_{1i}, s_{2i}, \ldots, s_{Mi}$ are the shares of industry $i$'s employment contained in each of these areas. Let $x_1, x_2, \ldots, x_M$ be some other measure of the size of these areas, such as each area's share of population or aggregate employment. A simple measure of the *raw geographic concentration* of industry $i$ is

$$G_i = \sum_{m=1}^{M} (s_{mi} - x_m)^2.$$

Ellison and Glaeser (1997) note that it is problematic to make cross-industry or cross-country comparisons using this measure because it will be affected by the size distribution of plants in the industry and the fineness of the available geographic data. They propose an alternate measure of agglomeration we will refer to as the EG index:

$$\gamma_i \equiv \frac{G_i/(1 - \sum_m x_m^2) - H_i}{1 - H_i},$$

where $H_i$ is the plant-level Herfindahl index of industry $i$.[1] They show that the EG index "controls" for differences in the plant size distribution and the fineness of the geographic breakdown, in the sense of being an unbiased estimator of a parameter reflecting the importance of natural advantages and spillovers in a simple model of location choice.

Ellison and Glaeser (1997) also propose a measure of the coagglomeration of a group of $I$ industries. Let $w_i$ be industry $i$'s share of total employment in the $I$ industries. Let $s_1, \ldots, s_M$ be the shares of the total employment in the group of $I$ industries in each of the geographic subareas. (Note that $s_m = \sum_{i=1}^{I} w_i s_{mi}$.) Write $G$ for the raw geographic

[1] This is defined by $H_i = \sum_{k=1}^{N_i} z_{ki}^2$, where $k = 1, 2, \ldots, N_i$ indexes the plants in industry $i$ and $z_{ki}$ is the employment of plant $k$ as a share of the total employment in industry $i$.

concentration for the $I$-industry group: $G = \sum_{m=1}^{M}(s_m - x_m)^2$. Write $H$ for the plant-level Herfindahl of the $I$-industry group: $H = \sum_i w_i^2 H_i$. The EG index of coagglomeration is

$$(1) \quad \gamma^c \equiv \frac{G/(1-\sum_m x_m^2) - H - \sum_{i=1}^{I} \gamma_i w_i^2 (1-H_i)}{1 - \sum_{i=1}^{I} w_i^2}.$$

The index reflects excess concentration of the $I$-industry group relative to what would be expected if each industry were as agglomerated as it is, but the locations of the agglomerations were independent. The particular form is motivated by a proposition relating the expected value of the index to properties of the location-choice model.

**Proposition 0** *Ellison and Glaeser (1997)*

*In an $I$-industry probabilistic location choice model, suppose that the indicator variables $\{u_{km}\}$ for whether the $k^{th}$ plant locates in area $m$ satisfy $E(u_{km}) = x_m$ and*

$$Corr(u_{km}, u_{\ell m}) = \begin{cases} \gamma_i & \text{if plants } k \text{ and } \ell \text{ both belong to industry } i \\ \gamma_0 & \text{if plants } k \text{ and } \ell \text{ belong to different industries.} \end{cases}$$

*Then, $E(\gamma^c) = \gamma_0$.*

### 1.2 A simpler formula

The EG coagglomeration index is a measure of the average coagglomeration of industries in a group. A simpler equivalent formula can be given for the coagglomeration of two industries.

**Proposition 1** *An equivalent formula for the EG coagglomeration index when $I = 2$ is*

$$\gamma^c = \frac{\sum_{m=1}^{M}(s_{m1} - x_m)(s_{m2} - x_m)}{1 - \sum_{m=1}^{M} x_m^2}.$$

The formula makes clear that the EG coagglomeration index is closely related to the covariance of the state-industry employment shares in the two industries. The denominator rescales the simple covariance to eliminate a sensitivity to the fineness of the geographic breakdown. Note that plant-level Herfindahls do not enter into the formula: the lumpiness of plants causes an increase in the variance of the state-industry employment shares that could be mistaken for within-industry agglomeration, but does not by itself lead to a spurious increase in the cross-industry covariance. (Larger plant Herfindahls will, however, make $\gamma^c$ a noisier parameter estimate.)

## 1.3 Explicit models of location choice

Proposition 0 is in a sense quite general: it shows that the coagglomeration index is appropriate if location decisions are made in any manner that satisfies one property. This generality, however, is obtained at the expense of losing explicit connections to the economics of location decisions and how they are influenced by natural advantages, spillovers, etc. In this section we extend the single-industry model of Ellison and Glaeser (1997) to make these connections.

We will discuss spillovers and natural advantages separately using two models with many elements in common. There are two industries indexed by $i = 1, 2$, with $N_1$ plants in industry 1 and $N_2$ plants in industry 2. Plants are indexed by $k \in K_1 \cup K_2$, with $K_1$ being the set of plants in industry 1 and $K_2$ being the set of plants in industry 2. The plants choose among $M$ possible locations. Each plant has an exogenously fixed employment level $e_k$ that is independent of its location choice.

### 1.3.1 Spillovers

We conceptualize spillovers as mechanisms that make plant $k$'s profits a function of the other plants' location decisions. A general model of this form would be to assume that firm $k$'s profits when locating in area $m$ are of the form $\pi_{km} = f(m, \ell_{-k}, \epsilon_{km})$, where $\ell_{-k}$ is the vector of all plants' location decisions and $\epsilon_{km}$ is a random shock. A difficulty with discussing the degree of geographic concentration in such a model is that the location choice process becomes a game that can have multiple equilibria. For example, if plants $k$ and $k'$ receive substantial benefits from co-locating, then there may be equilibria in which the two plants co-locate in any of several areas that are fairly good for each plant, and also an equilibrium in which the plants forego the spillover benefits and locate in the areas that are most advantageous for each plant separately. (This will only be an equilibrium if plant $k$'s most-preferred location is sufficiently unattractive to plant $k'$ and vice-versa.) The different equilibria will typically lead to different levels of measured concentration.

Ellison and Glaeser (1997) note that the impact of equilibrium multiplicity is substantially reduced if one considers random "all-or-nothing" spillovers. To extend their analysis, define a *partition* $\omega$ of $K_1 \cup K_2$ to be a correspondence $\omega : K_1 \cup K_2 \overrightarrow{\to} K_1 \cup K_2$ such that

$k \in \omega(k)$ for all $k$ and $k' \in \omega(k) \Rightarrow \omega(k) = \omega(k')$. Suppose that plants' location decisions are the outcome of game in which the plants choose locations in some (possibly random) exogenously specified order and plant $k$'s profits from locating in area $m$ are given by

$$\log(\pi_{km}) = \log(x_m) + \sum_{k' \in \omega(k)} I(\ell_{k'} \neq m)(-\infty) + \epsilon_{km}.$$

The first term on the right-side of this expression, $x_m$, is the measure of the size of area $m$ we used when constructing the concentration index. Its inclusion allows the model to match real-world data in which many more plants locate in California than in Wyoming.[2] The second term reflects the impact of spillovers: the interpretation is that a spillover exists between plants $k$ and $k'$ if $k' \in \omega(k)$ and that when spillovers exist they are sufficiently strong so as to outweigh all other factors in the location decision process. The third term in the profit function, $\epsilon_{km}$, is a Weibull distributed random shock that is independent across plants and locations.

**Proposition 2** *Consider the model of location choices with spillovers described above:*

*(a) The Perfect Bayesian equilibrium outcome is essentially unique. In equilibrium, each plant $k$ chooses the location $m$ that maximizes $\log(x_m)+\epsilon_{km}$ if no plant with $k' \in \omega(k)$ has previously chosen a location, and the location of previously located plants with $k' \in \omega(k)$ if some such plants have previously chosen a location.*

*(b) If $0 \leq \gamma_0^s \leq \gamma_1^s, \gamma_2^s$ or $0 \leq \gamma_1^s, \gamma_2^s$ and $0 \leq \gamma_0 \leq \min(1/N_1, 1/N_2)$, then there exist distributions over the set of possible partitions for which*

$$Prob\{k' \in \omega(k)\} = \begin{cases} \gamma_i^s & \text{if plants } k \text{ and } k' \text{ both belong to industry } i \\ \gamma_0^s & \text{if plants } k \text{ and } k' \text{ belong to different industries,} \end{cases}$$

*(c) If the distribution satisfies the condition in part (b), then in any PBE of the model the agglomeration and coagglomeration indexes satisfy*

$$\begin{aligned} E(\gamma_i) &= \gamma_i^s \\ E(\gamma^c) &= \gamma_0^s. \end{aligned}$$

[2] Ellison and Glaeser (1997) note that their model has an equivalent formulation in which each potential "location" is equally profitable on average and the reason why there are many more plants in California is that California is an aggregate of a larger number of "locations".

*Remarks:*

1. Note that Proposition 2 shows a degree of robustness to equilibrium selection: it shows that the agglomeration index has the same expected value in any PBE of the sequential move games obtained by ordering the plants in different ways.

2. Proposition 2 also shows some robustness to the distribution of spillover benefits. Our agglomeration and coagglomeration indexes have the same expected value for any distribution over partitions satisfying the condition in part (b). The proof of the proposition describes a couple different ways to generate distributions satisfying the condition. One is very simple technically and has a four point support. Another generates coagglomeration patterns that look more reasonable by first creating clusters within each industry and then joining clusters across industries.

### 1.3.2 Shared natural advantage

Another mechanism that can lead to the coagglomeration of plants in two industries is the presence in some areas of "shared natural advantages" that provide benefits to firms in both industries. The natural advantages can be exogenous factors, as when a coastal location makes a state attractive both to shipbuilding plants and to oil refineries. They can also be endogenous factor advantages of the types described in each of Marshall's theories, e.g. airplane manufacturers and automobile parts manufacturers may be coagglomerated because both benefit from locating in areas with skilled machinists.

To model natural-advantage influenced location choice, we suppose that profits for a plant $k$ that belongs to industry $i(k)$ and locates in area $m$ are given by

$$(2) \quad \log(\pi_{mk}) = \log(\eta_m + \xi_{mi(k)}) + \epsilon_{mk},$$

where the $\eta_m$, $\xi_{mi}$, and $\epsilon_{mk}$ are mutually independent random variables. The $\eta_m$ can be thought of as reflecting shared natural advantages of each area $m$ that make it attractive or unattractive to plants in both industries.[3] The $\xi_{mi}$ reflect additional factors that are idiosyncratic to industry $i$. As in the previous model, we also assume that there are plant-idiosyncratic factors, $\epsilon_{mk}$.

[3] These could include state policies as discussed in Thomas Holmes (1998).

**Proposition 3** *Suppose that profits are as in equation 2 and that each plant k chooses the location m that maximizes* $\pi_{mk}$.

*(a) Suppose* $0 < \gamma_1^{na} \leq \gamma_2^{na}$, *and that* $0 \leq \gamma_0^{na} \leq \frac{1-\gamma_2^{na}}{1-\gamma_1^{na}}\gamma_1^{na}$. *Write* $\delta_{mi}$ *for* $\eta_m + \xi_{mi}$. *Then, there exist distributional choices for the* $\eta_m$ *and* $\xi_{mi}$ *for which*

$$E(\delta_{mi} / \sum_{m'=1}^{M} \delta_{m'i}) = x_m,$$

$$Var(\delta_{mi} / \sum_{m'=1}^{M} \delta_{m'i}) = \gamma_i^{na} x_m (1 - x_m),$$

$$Cov(\delta_{m1} / \sum_{m'=1}^{M} \delta_{m'1}, \delta_{m2} / \sum_{m'=1}^{M} \delta_{m'2}) = \gamma_0^{na} x_m (1 - x_m).$$

*(b) If the distributions of the* $\eta_m$ *and the* $\xi_{mi}$ *are such that the conditions in part (a) are satisfied and the* $\epsilon_{mk}$ *are independent Weibull random variables, then the agglomeration and coagglomeration indexes satisfy*

$$\begin{aligned} E(\gamma_i) &= \gamma_i^{na} \\ E(\gamma^c) &= \gamma_0^{na}. \end{aligned}$$

*Remarks:*

1. As is described in more detail in the proof of Proposition 3, one specification of the shared- and industry-idiosyncratic natural advantages that can be made to satisfy the conditions in part (a) involves choosing the $\eta_m$ and $\xi_{mi}$ to be $\chi^2$ random variables with appropriately chosen degrees of freedom. In this specification the $\delta_{mi}$ are $\chi^2$ random variables with $2x_m(1 - \gamma_i^{na})/\gamma_i^{na}$ degrees of freedom. The lowest level of coagglomeration, $E(\gamma^c) = 0$, obtains when there are no shared natural advantages: if we assume that the $\eta_m$ are identically zero, then the $\delta_{mi}$ are independent across industries and state-industry employments will be independent across industries. The greatest degree of coagglomeration, $E(\gamma^c) = \frac{1-\gamma_2^{na}}{1-\gamma_1^{na}}\gamma_1^{na}$, obtains when we make the shared natural advantages as important as possible: if the $\xi_{m2}$ are identically zero, then all of the natural advantages affecting industry 2 are shared natural advantages.[4]

[4] In this case, the $\eta_m$ are distributed $\chi^2$ with $2x_m(1 - \gamma_2^{na})/\gamma_2^{na}$ degrees of freedom and the $\xi_{m1}$ are $\chi^2$ with $2x_m(\frac{1-\gamma_1^{na}}{\gamma_1^{na}} - \frac{1-\gamma_2^{na}}{\gamma_2^{na}})$ degrees of freedom.

2. Ellison and Glaeser (1997) also provide a result characterizing the expected value of the agglomeration index when both spillovers and natural advantages are present. This result does not have a clean generalization to the multi-industry case. The difficulty is that both agglomeration and coagglomeration are no longer independent of the equilibrium selection. For example, if a spillover exists between plants in separate industries, there will be more agglomeration in each industry if the plant from the more agglomerated industry chooses the joint location than if the plant from the less agglomerated industry does so.

## 2 Proofs of Propositions

Proof of Proposition 1: Note that

$$\begin{aligned} G &= \sum_{m=1}^{M}(w_1 s_{m1} + w_2 s_{m2} - x_m)^2 = \sum_{m=1}^{M}(w_1(s_{m1} - x_m) + w_2(s_{m2} - x_m))^2 \\ &= w_1^2 \sum_{m=1}^{M}(s_{m1} - x_m)^2 + w_2^2 \sum_{m=1}^{M}(s_{m2} - x_m)^2 + 2w_1 w_2 \sum_{m=1}^{M}(s_{m1} - x_m)(s_{m2} - x_m) \\ &= w_1^2 G_1 + w_2^2 G_2 + 2w_1 w_2 \sum_{m=1}^{M}(s_{m1} - x_m)(s_{m2} - x_m), \end{aligned}$$

and that $H = w_1^2 H_1 + w_2^2 H_2$. Hence,

$$\begin{aligned} (1 - \sum_{i=1}^{2} w_i^2)\gamma^c &= G/(1 - \sum_m x_m^2) - H - \sum_{i=1}^{2} \gamma_i w_i^2 (1 - H_i) \\ &= G/(1 - \sum_m x_m^2) - (w_1^2 H_1 + w_2^2 H_2) - \sum_{i=1}^{2} \left( \frac{G_i/(1 - \sum_m x_m^2) - H_i}{1 - H_i} \right) w_i^2 (1 - H_i) \\ &= \frac{w_1^2 G_1 + w_2^2 G_2 + 2w_1 w_2 \sum_{m=1}^{M}(s_{m1} - x_m)(s_{m2} - x_m) - \sum_{i=1}^{2} w_i^2 G_i}{1 - \sum_{m=1}^{M} x_m^2} \\ &= \frac{2w_1 w_2 \sum_{m=1}^{M}(s_{m1} - x_m)(s_{m2} - x_m)}{1 - \sum_{m=1}^{M} x_m^2}. \end{aligned}$$

The final formula results from noting that $2w_1 w_2 = 1 - \sum_{i=1}^{2} w_i^2$ when $w_1 + w_2 = 1$.

Proof of Proposition 2: Part (a) of the theorem follows immediately from backward induction. The final plant to move must choose in this way. Given that the final plant will locate in this way, the next-to-last plant maximizes its payoff by choosing the location that maximizes $\log(x_m) + \epsilon_{km}$ if it has no spillover with a previously located plant, because it

will receive full spillover benefits from the final plant (if such spillovers exist) regardless of its location choice. The qualification "essentially unique" in the proposition reflects that the maximizing choice is not unique when the maximizer of $\log(x_m) + \epsilon_{km}$ is not unique. This occurs with probability zero.

Part (b) states that we can choose a distribution over partitions that satisfies

$$\text{Prob}\{k' \in \omega(k)\} = \begin{cases} \gamma_i^s & \text{if plants } k \text{ and } k' \text{ both belong to industry } i \\ \gamma_0^s & \text{if plants } k \text{ and } k' \text{ belong to different industries,} \end{cases}$$

if either of two hypotheses holds.

The first hypothesis is that $0 \leq \gamma_0^s \leq \gamma_1^s, \gamma_2^s$. In this case, a four-point distribution suffices. Let $\omega_0$ be the fully disjoint partition: $\omega_0(k) = \{k\}$ for all $k$. Let $\omega_i$ be the partition in which all plants in industry $i$ are in a single cluster with the remaining plants disjoint: $\omega_i(k) = K_i$ if $k \in K_i$ and $\omega_i(k) = \{k\}$ otherwise. Let $\omega_{12}$ be the partition with all plants in a single cluster: $\omega_{12}(k) = K_1 \cup K_2$ for all $k$. The distribution that place probability $\gamma_0^s$ on $\omega_{12}$, probability $\gamma_i^s - \gamma_0^s$ on $\omega_i$, and the remaining probability on $\omega_0$ has the desired property.

The second hypothesis is that $0 \leq \gamma_1^s, \gamma_2^s$ and $0 \leq \gamma_0 \leq \min(1/N_1, 1/N_2)$. In this case, it is simplest to describe the construction of a distribution on the set of partitions on $K_1 \cup K_2$ as a two-step process. Let $p_1$ be a probability distribution over partitions of $K_1$ that satisfies $p_1(\{\omega | k' \in \omega(k)\}) = \gamma_1^s$ for all $k, k' \in K_1$. This can be done easily by putting probability $\gamma_1^s$ on the partition with all plants in a single cluster and the remaining probability on a disjoint partition, and can also be done in many other ways if $\gamma_1^s$ is not too large. Similarly, let $p_2$ be a distribution over partitions of $K_2$ that satisfies $p_2(\{\omega | k' \in \omega(k)\}) = \gamma_2^s$ for all $k, k' \in K_2$. To choose a partition of $K_1 \cup K_2$, first draw partitions $\omega_1$ of $K_1$ and $\omega_2$ of $K_2$ according to $p_1$ and $p_2$. Let $C_i$ be the set of clusters in partition $i$: $C_i = \{S \subset K_i | \omega_1(k) = S \text{ for some } k \in K_1\}$. Assuming WLOG that $|C_1| < |C_2|$, let $f$ be a one-to-one function from $C_1$ to $C_2$ chosen uniformly from the set of all such functions. Then, define a partition $\omega$ on $K_1 \cup K_2$ by setting $\omega(k) = \omega_1(k)$ with probability $1 - |C_2|\gamma_0$ and $\omega(k) = \omega_1(k) \cup f(\omega_1(k))$ with probability $|C_2|\gamma_0$ for $k \in K_1$, and defining $\omega(k) = \omega_2(k)$ if $k \in K_2$ and $k$ has not previously been defined as belonging to some $\omega(k)$ with $k \in K_1$. (The randomization in this definition is perfectly correlated across $k$ and $k'$ if $k' \in \omega_1(k)$ and can have any correlation if $k$ and $k'$

are not in the same cluster.) It is straightforward that a partition created this way has the desired property.

Part (c) is a corollary of Proposition 0. Let $u_{km}$ be an indicator for plant $k$ locating in area $m$. A standard property of the logit model is that $\text{Prob}\{u_{km}=1\} = x_m/\sum_{m'} x_{m'} = x_m$. The locations of plants $k$ and $k'$ are identical if $k' \in \omega(k)$ and independent otherwise, so

$$E(u_{km}u_{k'm}|\omega) = \begin{cases} x_m & \text{if } k' \in \omega(k) \\ x_m^2 & \text{otherwise.} \end{cases}$$

The unconditional expectation is $E(u_{km}u_{k'm}) = x_m^2 + \text{Prob}\{k' \in \omega(k)\}(x_m - x_m^2)$. Using this we calculate

$$\begin{aligned} \text{Corr}(x_{km}, x_{k'm}) &= \frac{E(u_{km}u_{k'm}) - E(u_{km})E(u_{k'm})}{\sqrt{\text{Var}(u_{km})\text{Var}(u_{k'm})}} \\ &= \frac{x_m^2 + \text{Prob}\{k' \in \omega(k)\}(x_m - x_m^2) - x_m^2}{\sqrt{x_m(1-x_m)x_m(1-x_m)}} \\ &= \text{Prob}\{k' \in \omega(k)\} \end{aligned}$$

Hence, the hypothesis of Proposition 0 is satisfied whenever the condition on the distribution over partitions in part (b) of Proposition 2 holds.

PROOF OF PROPOSITION 3: Suppose that the $\eta_m$ and $\xi_{mi}$ are independent $\chi^2$ random variables with $\frac{1-\gamma_2^{na}}{\gamma_2^{na}} 2c_m x_m$ and $\frac{1-\gamma_i^{na}}{\gamma_i^{na}} 2x_m - \frac{1-\gamma_2^{na}}{\gamma_2^{na}} 2c_m x_m$ degrees of freedom, respectively, for some constants $c_m \in [0,1]$. The additive property of $\chi^2$ random variables implies that $\delta_{mi}$ is a $\chi^2$ random variable with $\frac{1-\gamma_i^{na}}{\gamma_i^{na}} 2x_m$ degrees of freedom. Note that $\delta_{mi}$ and $\delta_{m'i}$ are independent if $m \neq m'$. A standard result on Chi-square distributions implies that $\delta_{mi}/\sum_{m'=1}^{M} \delta_{m'i}$ has a Beta distribution with parameters $\frac{1-\gamma_i^{na}}{\gamma_i^{na}} x_m$ and $\frac{1-\gamma_i^{na}}{\gamma_i^{na}}(1-x_m)$.[5] A Beta distribution with parameters $\theta_1$ and $\theta_2$ has expectation $\theta_1/(\theta_1+\theta_2)$ and variance $\frac{\theta_1\theta_2}{(\theta_1+\theta_2)^2(\theta_1+\theta_2+1)}$. Using these formulas gives

$$E\left(\frac{\delta_{mi}}{\sum_{m'=1}^{M}\delta_{m'i}}\right) = \frac{\frac{1-\gamma_i^{na}}{\gamma_i^{na}} x_m}{\frac{1-\gamma_i^{na}}{\gamma_i^{na}}} = x_m$$

$$\text{Var}\left(\frac{\delta_{mi}}{\sum_{m'=1}^{M}\delta_{m'i}}\right) = \frac{\left(\frac{1-\gamma_i^{na}}{\gamma_i^{na}}\right)^2 x_m(1-x_m)}{\left(\frac{1-\gamma_i^{na}}{\gamma_i^{na}}\right)^2 \frac{1}{\gamma_i^{na}}} = \gamma_i^{na} x_m(1-x_m).$$

[5] See Chapter 25 of Johnson, Kotz and Balakrishnan (1995).

This shows that the distributions have two of the three desired properties given in part (a) of the Proposition.

To complete the proof of part (a) it suffices to show that the third property,

$$\text{Cov}(\delta_{m1}/\sum_{m'=1}^{M}\delta_{m'1}, \delta_{m2}/\sum_{m'=1}^{M}\delta_{m'2}) = \gamma_0^{na}x_m(1-x_m)$$

holds for some choice of $c_m \in [0,1]$. The covariance is a continuous function of $c_m$. When $c_m = 0$, the covariance is zero. Hence, by the intermediate value theorem we can complete the proof by showing that the covariance is equal to $\frac{1-\gamma_2^{na}}{1-\gamma_1^{na}}\gamma_1^{na}x_m(1-x_m)$ when $c_m = 1$.

When $c_m = 1$ the covariance can be written as

$$\begin{aligned}\text{Cov}\left(\frac{\delta_{m1}}{\sum_{m'=1}^{M}\delta_{m'1}}, \frac{\delta_{m2}}{\sum_{m'=1}^{M}\delta_{m'2}}\right) &= \text{Cov}\left(\frac{\eta_m}{\sum_{m'=1}^{M}\eta_{m'}}, \frac{\eta_m+\xi_{m2}}{\sum_{m'=1}^{M}\eta_{m'}+\xi_{m'2}}\right), \\ &= \text{Cov}\left(\frac{Y_0}{Y_0+Y_0'}, \frac{Y_0+Y_1}{Y_0+Y_0'+Y_1+Y_2}\right),\end{aligned}$$

where $Y_0 = \eta_m$, $Y_0' = \sum_{m'\neq m}\eta_{m'}$, $Y_1 = \xi_{m2}$, and $Y_2 = \sum_{m'\neq m}\xi_{m'2}$. Note that $Y_0$, $Y_0'$, $Y_1$ and $Y_2$ are mutually independent Chi-square random variables. By rewriting the last term on the right side of the above expression as $\frac{Y_0}{Y_0+Y_0'}\frac{Y_0+Y_0'}{Y_0+Y_0'+Y_1+Y_2} + \frac{Y_1}{Y_1+Y_2}\frac{Y_1+Y_2}{Y_0+Y_0'+Y_1+Y_2}$ we find that it is equal to

$$\begin{aligned}&\text{Cov}\left(\frac{Y_0}{Y_0+Y_0'}, \frac{Y_0}{Y_0+Y_0'}\frac{Y_0+Y_0'}{Y_0+Y_0'+Y_1+Y_2} + \frac{Y_1}{Y_1+Y_2}\frac{Y_1+Y_2}{Y_0+Y_0'+Y_1+Y_2}\right) \\ &\quad = \text{Cov}\left(\frac{Y_0}{Y_0+Y_0'}, \frac{Y_0}{Y_0+Y_0'}\frac{Y_0+Y_0'}{Y_0+Y_0'+Y_1+Y_2}\right) + \text{Cov}\left(\frac{Y_0}{Y_0+Y_0'}, \frac{Y_1}{Y_1+Y_2}\frac{Y_1+Y_2}{Y_0+Y_0'+Y_1+Y_2}\right)\end{aligned}$$

Another standard property of Chi-square (and Gamma) random variables is that $\frac{Y_0}{Y_0+Y_0'}$ and $Y_0+Y_0'$ are independent.[6] This immediately implies that the second covariance in the line above is zero. It also implies that $\frac{Y_0}{Y_0+Y_0'}$ and $\frac{Y_0+Y_0'}{Y_0+Y_0'+Y_1+Y_2}$ are independent. This implies

$$\text{Cov}\left(\frac{Y_0}{Y_0+Y_0'}, \frac{Y_0}{Y_0+Y_0'}\frac{Y_0+Y_0'}{Y_0+Y_0'+Y_1+Y_2}\right) = \text{Var}\left(\frac{Y_0}{Y_0+Y_0'}\right)E\left(\frac{Y_0+Y_0'}{Y_0+Y_0'+Y_1+Y_2}\right).$$

Plugging in the appropriate degrees of freedom into the formulas for the mean and variance of Beta-distributed random variables we find that this is equal to

$$\gamma_2^{na}x_m(1-x_m)\frac{\frac{1-\gamma_2^{na}}{\gamma_2^{na}}}{\frac{1-\gamma_1^{na}}{\gamma_1^{na}}} = x_m(1-x_m)\frac{1-\gamma_2^{na}}{1-\gamma_1^{na}}\gamma_1^{na}.$$

[6] See Chapter 17 of Johnson, Kotz and Balakrishnan (1994).

## References

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

Johnson, Norman L., Kotz, Samuel and Balakrishnan, N. (1994): *Continuous Univariate Distributions,* Vol. 1. New York: John Wiley & Sons.

Johnson, Norman L., Kotz, Samuel and Balakrishnan, N. (1995): *Continuous Univariate Distributions,* Vol. 2. New York: John Wiley & Sons.

# Data and Empirical Appendix to "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns"

Glenn Ellison
MIT

Edward Glaeser
Harvard University

William Kerr*
Harvard Business School

January 2009

*email: gellison@mit.edu, glaeser@fas.harvard.edu, wkerr@hbs.edu.

# 1 Overview

This appendix provides supporting materials for the results presented in Ellison, Glaeser and Kerr (2009). We combine materials first circulated in the text and data appendix of NBER Working Paper 13068 with new information related to subsequent revisions. We first outline the construction of the coagglomeration metrics from the US Census Data. We then describe the data sources and design of our explanatory variables and instrumental variables. Comprehensive empirical results are then presented. A separate mathematical appendix contains the theoretical model of location choice.

# 2 US Coagglomeration Metrics

This section begins with an outline of the US Census Bureau data employed. We then construct two metrics of coagglomeration: the discrete index of Ellison and Glaeser (1997) and an approximation to the continuous metric of Duranton and Overman (2005).

## 2.1 US Census Bureau Data

Our estimates of industrial coagglomeration patterns are developed through confidential data housed by the US Census Bureau. The Census of Manufacturing is conducted every five years (those ending with 2 or 7) and surveys the universe of manufacturing plants operating in the US. With appropriate clearance, researchers can analyze the microdata of these Censuses, which is essential for estimating coagglomeration levels of detailed industries as public reports suppress values that risk disclosing the operating details of individual firms. Moreover, as the microdata for plants can be linked longitudinally across Censuses, we can compare the coagglomeration of existing establishment with that of new entrants. We focus on the six Censuses conducted from 1972 to 1997, providing approximately 300k establishment observations employing 17m workers in each census year.

We define manufacturing industries through the three-digit level of the 1987 Standard Industrial Classification (SIC3). 9870 pairwise industry combinations of the 140 SIC3 divisions are possible (including own-industry agglomeration). Tobacco (210s), Fur (237), and Search and Navigation Equipment (381) are excluded throughout the paper due to major industry reclassifications at the plant-level in the Census of Manufacturing that are difficult to interpret. In the empirical estimations of the main paper, the remainder of Apparel (230s), a portion of Printing and Publishing (277-279), and Secondary Non-Ferrous Metals (334) are also excluded due to either an inability to construct appropriate Marshallian explanatory matrices or outlier concerns in the explanatory data. Finally, we exclude same-industry pairs (i.e., agglomeration) for a total of 7381 unique pairwise industry combinations per Census of Manufacturing, representing 122

underlying industries.[1]

## 2.2 Ellison and Glaeser (1997)

Our first metric descends from Ellison and Glaeser (1997, hereafter EG). Following Proposition 1 of the mathematical appendix, the pairwise EG coagglomeration between industry pair $i$ and $j$ can be analyzed with the simple formula

$$\gamma_{ij}^{c} = \frac{\sum_{m=1}^{M}(s_{mi} - x_m)(s_{mj} - x_m)}{1 - \sum_{m=1}^{M} x_m^2},$$

where $m$ indexes geographic regions. $s_{1i}, s_{2i}, \ldots, s_{Mi}$ are the shares of industry $i$'s employment contained in each of these areas. $x_1, x_2, \ldots, x_M$ are some other measure of the size of these areas, such as each area's share of population or aggregate employment. We model $x_m$ in this project through the mean employment share in the region across manufacturing industries.[2] Our primary measure of the economic activity in an industry $j$ in a given geographic area $m$ is the total employment in all manufacturing establishments excluding auxiliary units. The $s_{mj}$ measure is then the share of the industry $j$'s employment in region $m$.

Throughout the paper, we simultaneously report EG coagglomeration metrics calculated at the state (including the District of Columbia), PMSA, and county levels. These variants only adjust the $M$ demarcations on which $s$ and $x$ are calculated. App. Table 1 documents the summary statistics for these metrics as employed in the coagglomeration estimations. App. Tables 2A-2C provide additional descriptive statistics for the EG metric.

App. Table 2A presents descriptive statistics of several measures of agglomeration and coagglomeration. The table is divided into three panels. The top panel presents indices calculated from state-level employment data. The first row shows that the EG industrial agglomeration index remains fairly stable between 1972 and 1982, and then falls by about 10% in the following decade. The next two rows summarize trends in the pairwise coagglomeration index. The mean pairwise coagglomeration is approximately zero. This is largely by definition: our benchmark measure of a state's "size" is its share of manufacturing employment so each industry's deviations from the benchmark will be approximately uncorrelated with the average of the deviations of all other industries. The standard deviation of the coagglomeration index is more interesting, showing a decline (tighter distribution) from 1972 to 1997.

The second panel presents corresponding figures computed using PMSA-level employments. The average decline in agglomeration from 1982 to 1992 is smaller at this geographic level and agglomeration appears to have increased from 1992 to 1997. The coagglomeration distribution

[1] The 2002 Census of Manufacturers recently became available. It employs the NAICS industry codes, however, that make it difficult to compare to earlier years. Ellison *et al.* (2006) discusses the calculation of coagglomeration measures under the NAICS framework. Dunne *et al.* (1989), McGuckin and Peck (1992), Davis *et al.* (1996), Autor *et al.* (2007), and Kerr and Nanda (2008) provide detailed accounts of the Census Bureau data.

[2] While the EG (1997) formula allows for the $x_m$ to vary across industries, the equivalency formula in Proposition 1 of the mathematical appendix requires that they be the same.

again shows a declining standard deviation. At the industry-pair level, the coagglomeration indices computed using the PMSA-level data have a 0.6 correlation with indices computed from state-level data (see App. Table 3C).

A nice feature of the Census of Manufacturing is that one can track plants over time and separate new plants from old plants. The third panel provides statistics on agglomeration and coagglomeration indices for the new "startups" in each industry.[3] The agglomeration and coagglomeration of these startups could be different from the overall pattern because they are less tied to past industrial centers than existing plants or the new establishments of existing firms (e.g., Dumais *et al.* 2002) and their location choices come after the inter-industry dependencies described below are formed. These measures are naturally more noisy than those calculated through total employment due to smaller number of plants involved and the distinct sets of plants being considered in each census year. The agglomeration data show an initial decline and a later increase, particularly in the final census year. This pattern is reflected in the standard deviation of the coagglomeration index, too. At the industry-pair level, the correlation between coagglomeration measures computed at the state level using all firms and those computed using new startups is 0.3.

Dumais *et al.* (2002) noted that the EG agglomeration index for an industry is highly correlated over time, even relative to the magnitude of state-industry employment changes. App. Table 2B indicates that coagglomeration indices are also highly correlated over time. For example, the number in the upper left cell indicates that the correlation between the 1972 and 1977 coagglomeration indices for an industry-pair is 0.953. The correlations are at least 0.9 for each five-year period. The correlation between 1972 and 1997 coagglomeration indices is still about 0.7.

Table 2 of the main paper contains a list of the fifteen most coagglomerated industry pairs. Most involve textile and apparel industries, which are heavily concentrated in North Carolina, South Carolina, and Georgia. None of these coagglomerations are as strong as the within-industry agglomerations of the most agglomerated industries. For example, Ellison and Glaeser (1997) find that $\gamma = 0.63$ for the fur industry (SIC 237). Many industry-pairs have approximately zero coagglomeration. Negative values of the index arise when pairs of industries are agglomerated in different areas. The lowest value of -0.065 obtains for the coagglomeration of the Guided Missiles and Space Vehicles (376) and Railroad Equipment (374) industries. We imagine that most strong negative coagglomerations like this are mostly due to coincidence.

App. Table 2C summarizes the mean 1987 coagglomeration between SIC3 pairs within SIC2

[3]More precisely, we first compute the total employment in each state-industry attributable to plants that did not appear in the previous census and did not belong to a firm that existed in the previous census (in this or any other industry). We then compute the agglomeration and coagglomeration indices using these totals as the state-industry employments. Approximately 80% of new manufacturing plants are startups in this sense. These startups enter at smaller sizes and account for about 50% of entering establishment employment. See Kerr and Nanda (2008) for more detail regarding the differences in entry sizes and entry rates between firm births and the expansion establishments of existing firms.

pairwise bins. The matrix confirms that SIC3 pairs within the same SIC2 category are generally positively coagglomerated. Apart from the high coagglomeration of the subindustries of the textile industry (SIC 22), none of the means are very large. This further illustrates that there is a great deal of idiosyncratic variation in coagglomeration levels across industry pairs.

## 2.3 Duranton and Overman (2005)

Duranton and Overman (2005, hereafter DO) construct a continuous metric of agglomeration. DO criticize indices like EG that employ discrete spatial units. This discreteness in effect makes the distance from Detroit MI to Chicago IL equivalent to that of Detroit MI to Miami FL. DO instead propose analyzing agglomeration of industry $i$ through a continuous index

$$\hat{K}_i(d) = \frac{1}{n_i(n_i-1)h} \sum_{r=1}^{n_i-1} \sum_{s=r+1}^{n_i} f\left(\frac{d-d_{r,s}}{h}\right),$$

where $d_{r,s}$ is the Euclidean distance between plants $r$ and $s$ within the focal industry $i$. $f$ is a Gaussian kernel density function with bandwidth $h$. The summations are over every pairwise bilateral distance of plants within the industry analyzed (i.e., $n_i(n_i-1)/2$ distances).

This agglomeration index can be generalized into a coagglomeration density for industries $i$ and $j$ measured through firm counts or employments:

$$\begin{aligned} \hat{K}_{ij}^{Ct}(d) &= \frac{1}{n_i n_j h} \sum\nolimits_{r=1}^{n_i} \sum\nolimits_{s=1}^{n_j} f\left(\frac{d-d_{r,s}}{h}\right) \\ \hat{K}_{ij}^{Emp}(d) &= \frac{1}{h \sum_{r=1}^{n_i} \sum_{s=1}^{n_j} e(r)e(s)} \sum\nolimits_{r=1}^{n_i} \sum\nolimits_{s=1}^{n_j} e(r)e(s) f\left(\frac{d-d_{r,s}}{h}\right). \end{aligned}$$

$n_i$ and $n_j$ are the number of plants in industries $i$ and $j$, respectively. The summations are over every bilateral distance between plants of industry $i$ and industry $j$ (i.e., $n_i n_j$ distances). We consider below densities that are weighted by plant employments, $e(r)$ and $e(s)$, and those considering just plant counts, $e = 1$.

This observed coagglomeration density is then compared to an underlying distribution of manufacturing activity akin to the $x_m$ of EG. This baseline is calculated through a thousand draws of two hypothetical industries of equivalent size to $i$ and $j$ and repeating the density estimation. Industry pairs where the observed coagglomeration density shows substantial deviations from the random draws are said to exhibit global localization or dispersion. As density estimations sum to one, this comparison is performed over a defined range from zero distance to a threshold level. We adjust the threshold distances in our empirical analysis.

Our approximations to DO are limited by two significant constraints. The first is the need to use county-to-county distances to measure bilateral distances between plants. Second, the continuous DO methodology is much more computationally intensive vis-a-vis simpler discrete

indices.[4] We must simplify the DO approach when working with 7381 industry pairs. Nonetheless, our approximations do capture much of the continuous nature sought by DO and are useful for examining how sensitive the relative rankings of agglomeration forces are to metric design and distances considered.

### 2.3.1 Base Index Calculations

We use county location identifiers for each plant to calculate our lumpy approximation of the DO coagglomeration index. The county spatial unit is the most-detailed spatial unit available for all manufacturing plants in the Census of Manufacturing.[5] The 3141 counties yield over 9.8m potential pairwise distances. We calculate distances between county centroids using the Haversine flat-earth formula that abstracts from the earth's curvature. Plants in the same county are given a uniform distance of one mile. Multiple factors yield conflicting recommendations regarding this latter choice — for example, accounting for larger land areas versus congestion effects with higher urban density. For coagglomeration, however, most of the action lies much further out in the distance distribution.

While the Census Bureau data contain the universe of manufacturing establishments, computational restrictions require that we cap the number of plants observed for each SIC3 industry. US plant counts range from under 100 to over 35k at this three-digit level. Given that the coagglomeration metric is being calculated for 7381 pairwise combinations of industries, we chose 1k random plant observations without replacement for any industry that had over 1k plants. Smaller industries were completely sampled. This 1k-plant upper bound influences 58 of our 122 industries, and we have confirmed with several industry pairs that the resulting distance distributions are very close to the full distributions. This sampling greatly increases our computational speed without introducing significant sampling error to the metric.[6]

This approach yields 1m plant-to-plant distances (or fewer) for an industry pair that are assigned Euclidean distances using county centroids. We then collapse these 1m observations

[4] As a guide for future researchers, the DO calculations discussed below require three months of computing time with four fully-utilized 3+ GHz processors and 32-bit CPUs. Approximately 20 gigabytes of disk space is required. This burden would be clearly less for studies of within-industry agglomeration.

[5] The Census Bureau data unfortunately lack plant-specific locations like the UK data employed by DO. The SSEL Business Register, also housed at the Census Bureau, provides tract-level identifiers. Tracts are a much-finer geographic unit than the county, and Arzaghi and Henderson (2008) employ this detail to analyze the agglomeration of advertising agencies within Manhattan. We are unable to exploit this tract-level data for our comprehensive manufacturing sample, however, due to extensive areas that are not assigned tract identifiers (accounting for over 30% of manufacturing observations in 1997). These unassigned areas are non-random and primarily in less-populated regions (i.e., more common in rural Michigan than Manhattan). These less-populated areas are clearly more important for manufacturing plants that have a greater presence on the outskirts of metropolitan areas and in smaller cities (e.g., Kolko 2000). Agglomeration forces (e.g., input-output linkages) for manufacturing are also more likely to operate on larger spatial dimensions than the networking described by Arzaghi and Henderson (2008). Nonetheless, examining tract-level coagglomeration in other sectors is a promising area for future research.

[6] The extreme industry counts are 35 (industry 321), 38 (261), and 50 (315) and 13,772 (344), 25,098 (359), and 33,513 (275). 64 industries have fewer than 1000 plants, 90 fewer than 2000 plants, 104 fewer than 3000 plants, and 117 fewer than 4000 plants. Some of the excluded tobacco-related industries have fewer than 35 plants.

to a distribution that uses mile units, ranging from [same county] to just over 6k miles.[7] We then apply the Gaussian kernel density function $f$ to smooth the distance series. For this smoothing, the data are reflected around zero and the bandwidth $h$ is chosen to minimize the mean integrated squared error. This smoothed density function is calculated for both plant counts and employments. This process is repeated for each of the 7381 pairwise combinations of manufacturing industries.

As a robustness check, we also calculated DO coagglomeration metrics using county-level firm counts and industry employments with associated distances between counties. This approach has the benefit of including the universe of observations, regardless of industry size, but a liability for calculating confidence intervals (and therefore aggregate localization) as the establishment-level information is sacrificed. We use the confidence intervals from the sub-sampled data with the distributions developed through the county-industry aggregates. As seen below, these county-industry alternatives are highly correlated with our bilateral plant-based metrics.

### 2.3.2 Confidence Intervals and Global Metrics

As in DO, these distance distributions are compared to counterfactuals of distances among randomly-drawn plants from manufacturing. This counterfactual baseline controls for the overall spatial distribution of manufacturing. Industry pairs are found to be substantially coagglomerated only to the extent they are more spatially clustered than US manufacturing as a whole. In preparing this test, it is important that each counterfactual have a similar number of observations (and therefore precision) to the focal distribution. It is also important, given the random nature of the plant draw, to replicate the counterfactual baseline multiple times. DO converge on 1k counterfactuals as the comparison for their agglomeration densities.

Computational burdens again require that we simplify how the confidence bands are calculated. We thus randomly drew without replacement 1k plants for two hypothetical industries; this sampling was repeated 1k times. We then calculated "sub-draws" of these 1k samples using 100-plant increments for both industry pairs. Effectively, we created 1k random draws of plants for 100 x 100 observations, 100 x 200, ... , 100 x 1000, 200 x 200, 200 x 300, ... , 1000 x 1000. The 1000 x 1000 case is the upper bound given the capped sampling procedure noted above. This process created 55 permutations of the 1k random draws.

For each of these 55 permutations, the 1k draws were prepared as in the observed data to determine a distance distribution. To compute confidence baselines akin to 5% and 95% confidence intervals, we identified the 10th and 990th extreme values by distance mile (i.e., 1% and 99% local intervals, further discussed below). Appending these values together, we created a confidence band series and applied the kernel density procedure. The 7381 pairwise industry

[7] The large upper bound is due to our inclusion of Alaska and Hawaii, but in general the density at this level is very small. We obtain equivalent results if plant-to-plant distances are capped at 3k miles, which is roughly the distance from Miami FL to Seattle WA. The correlation in metrics for the capped and uncapped distributions is 0.99.

combinations are matched to the constructed confidence bands that most closely reflect the underlying observation counts (e.g., 123 x 872 actual $\Longrightarrow$ 100 x 900 confidence bands). We designate the confidence intervals as $K_{ij}^{UC}(d)$ and $K_{ij}^{LC}(d)$.[8]

The DO global indices of localization $\Gamma_{ij}$ and dispersion $\Psi_{ij}$ for industry pair $i$ and $j$ are calculated using the formulas:

$$\begin{aligned}\Gamma_{ij}^{Emp} &\equiv \sum\nolimits_{d=0}^{threshold} \max\left[\hat{K}_{ij}^{Emp}(d) - K_{ij}^{UC}(d), 0\right] \\ \Psi_{ij}^{Emp} &= \sum\nolimits_{d=0}^{threshold} \max\left[K_{ij}^{LC}(d) - \hat{K}_{ij}^{Emp}(d), 0\right] \text{ if } \Gamma_{ij}^{Emp} = 0 \\ &\quad \text{and 0 otherwise.}\end{aligned}$$

The calculation of the DO global index of localization depends significantly on the distance threshold analyzed (*threshold*). As density functions sum to one over the support, localization in one distance range will correspond to dispersion in other distances.

When studying the UK, DO employ a distance threshold for identifying localization of 180 kilometers, which is the median distance among UK manufacturing plants. The US's large land mass and scattered regional industrial centers do not provide as clear of a threshold, and the appropriate threshold is not dictated by theory. There is also a question of edge effects as discussed in Harrison and Kominers (2008). We therefore contrast four levels: 1000, 500, 250, and 100 miles. 1000 miles is approximately the median distance between US plants and is on the order of magnitude for the distance between Detroit MI and Dallas TX. The shorter thresholds approximate the 25th, 10th, and 3rd percentiles of plant bilateral distances, respectively. They are on the order of magnitude of Detroit MI to Washington DC, Cincinnati OH, and Lansing MI, respectively. The 100 mile threshold is of comparable distance to the UK hurdle employed by DO (180 km. = 112 miles), while 1000 miles maintains the median concept.[9]

Having introduced the distance threshold, we can now return to a more subtle point regarding the confidence band intervals. Selecting the 10th and 990th observations is a simplification of the DO procedure. DO instead repeat the density procedure for each random draw. DO then optimally identify the ranking that when applied uniformly across distances would yield exact 5% and 95% confidence bands. DO report that the 10th and 990th extreme values for 1k random draws are representative outcomes for their UK sample.

In general, however, the DO confidence bands are specific to the distance threshold analyzed. We may risk over-estimating localization with longer distance thresholds if the uniform extreme values of the 10th and 990th observations are employed; likewise, localization may be under-estimated at shorter distances. This is unfortunate as a single appropriate distance threshold

[8] While not expressed in the notation, UC and LC are calculated for the density metric in question (e.g., employment, counts). The bilateral bands are employed for the county-industry densities. DO also analyze local confidence intervals; our exposition focuses only on global confidence levels.

[9] More precisely, the 50th percentile falls between 900 and 1003 miles depending upon the weighting employed. The 25th and 10th percentiles are between 517 and 560 miles and 264 and 275 miles, respectively.

cannot be determined, but the procedure is clearly too intensive to re-estimate the bands every time the distance threshold is adjusted with so many industry pairs.

To address this issue, we repeated the full DO confidence band procedure for the 1000 x 1000 sample. The exact extreme value for the 1000, 500, 250, and 100 mile thresholds are the 6-7th, 9-10th, 10-11th, and 18-19th observations, respectively.[10] Said another way, the 5% confidence band for the 1000 mile horizon falls at the 0.6% uniform percentile, not at the 1% percentile that is appropriate for 250 miles. Likewise, the 100 mile confidence bands are narrower at 1.8%. Given this, we calculated another set of DO metrics that adjusted the extreme values for the 1000 and 100 mile distributions appropriately. We apply this 6th and 18th extreme value rule to all 55 permutations, not just the 1000 x 1000 draw on which it was calculated. Both adjusted DO metrics are extremely close to those applying uniform 1% and 99% rules (correlations of 0.98 and above). Our reported results maintain the uniform 1% and 99% rule for convenience, and the other estimations are available upon request.

Finally, it is important to note that the random-plant baseline is criticized by DO (2008) as being of limited interest for coagglomeration studies as it fails to separate joint-localization (e.g., two industries locate near each other because they share a need for coastal access) from co-localization (e.g., two industries locate near each other for input-output rationales). DO instead undertake pairwise comparisons for a small, selected number of vertically-dependent industries to provide a specific test of input-output rationales. Our goal is to test several co-agglomeration forces through the universe of manufacturing industry pairs. We therefore chose the consistent baseline that could be replicated for all pairwise combinations. We also tested the joint-localization concerns through natural advantages distributions discussed below.

### 2.3.3 Descriptive Statistics

App. Table 3A presents descriptive statistics for the bilateral firm employments and counts with the four thresholds. 87% and 62% of industry pairs exhibit excessive coagglomeration by the 250 and 100 mile thresholds, respectively. Most industry pairs exhibit excessive coagglomeration when employing a 1000 mile threshold. The share of industries showing global localization and dispersion of less than 0.001 rises from 0% to 18% as distance thresholds are shortened.

App. Table 3B lists the fifteen most-coagglomerated industry pairs for the DO metrics at the 1000 mile and 250 mile thresholds. Textile and apparel industries rank very high regardless of the thresholds; the same is true in the EG rankings contained in Table 2 of the main paper. When employing the longer threshold, the DO measure finds substantial concentration between textiles industries (SIC 22) and metal industries (SIC 33-35). The latter industries are heavily concentrated in Michigan and surrounding states. This falls within the 1000 mile distance from the localized textile and apparel activity in the Carolinas and Georgia. The EG metric never identifies coagglomeration between these industry pairs, as their activity is very localized in

[10] These percentiles are the same for the capped distribution except that the 250 mile threshold falls 11-12th.

separate states and cities. The DO metric identifies them as having global localization when a 1000 mile threshold is employed, as 1000 miles is sufficient to bridge the two regions. On the other hand, these industry pairs do not exhibit global localization when a 250 mile threshold is employed.[11]

The most dispersed industry pair using the DO metric at 250 miles is Guided Missiles and Space Vehicles (376) and Pulp Mills (261). The greatest DO dispersion measured with the 1000 mile threshold of 0.746 is for Guided Missiles and Space Vehicles (376) and Broadwoven Mills, Wool (223).

App. Table 3C provides the correlations of the DO and EG metrics. The correlation of EG and DO using a 250 mile threshold is substantial at 0.4; it is lower at 0.1 with a 1000 mile threshold. Similar correlations hold for rank orderings. App. Tables 3D-3F provide correlations of the DO metrics calculated over different distance thresholds. The correlations clearly decline as the distance thresholds are widened. App. Tables 3G and 3H provide correlations of the DO metrics calculated with different metrics. There is a very high correlation, usually above 0.9, between the different techniques. This provides comfort in the robustness of our results to how we approximated the DO metric.

# 3 US Coagglomeration Determinants

We use industry attributes to design coagglomeration-oriented metrics that mirror each of Marshall's three theories of industry agglomeration: (1) labor market pooling, (2) proximity to input suppliers or industrial customers to save on transportation costs, and (3) intellectual or technology spillovers. We also build measures of expected coagglomeration descending from common dependency on specific natural advantages. App. Table 1 documents the summary statistics for these metrics, and App. Table 4 lists the extreme pairwise values. A condensed version of this section appears in the main text.

## 3.1 Labor Market Pooling

One of Marshall's theories of industrial location is that firms locate near one another to shield workers from the vicissitudes of firm-specific shocks. Workers are willing to accept lower wages in locations where other firms stand by ready to hire them (see Diamond and Simon (1990) for evidence and Krugman (1991) for a formalization). Rotemberg and Saloner (2000) present an alternative theory in which workers gain not because of insurance from shocks, but because multiple firms protect workers against ex post appropriation of investments in human capital. Both theories predict that plants that use the same type of workers will locate near one another.

[11] Census region 5 hosts approximately 16% of total manufacturing activity, 64% of textiles (SIC22), and 11% of metals industries (SIC33-35). Census regions 3 and 4 host 29% of total manufacturing, 3% of textiles, and 42% of metals industries. Region 5 includes DC, DE, GA, MD, NC, SC, VA, and WA. Regions 3 and 4 include KS, IA, IL, IN, MI, MN, MO, ND, NE, OH, SD, and WI.

Combes and Duranton (2006) model how start-ups may locate near older firms to hire away their workers.

To test the labor pooling theory, we construct a metric of the similarity in the occupational labor requirements for pairwise industries. We build from the 1987 National Industry-Occupation Employment Matrix (NIOEM) published by the Bureau of Labor Statistics (BLS). The NIOEM provides industry-level employments (at the national level) in 277 occupations. We convert the occupational employment counts into occupational percentages for each industry and map the BLS industries to the SIC3 framework. 52 of the 185 broadly-defined BLS industries are within manufacturing. Each SIC3 industry is assumed to possess the same occupational composition of employment as that of the NIOEM industry to which it belongs.[12]

Our metric of labor similarity, $LaborCorrelation_{ij}$, is a vector correlation of occupational percentages between two industries. $LaborCorrelation_{ij}$ averages 0.47 across the pairs of manufacturing industries, with a range of -0.05 to 1.00. The least correlated industry pair is Logging (241) and Aircrafts and Parts (372) at -0.046. The perfect correlation maximum value reflects that some NIOEM industries map to two or more SIC3 industries; the empirical specifications below are robust to this multiplicity. The most correlated industry pair, not by construction, is Motor Vehicles and Equipment (371) and Motorcycles, Bicycles, and Parts (375) at 0.984. Finally, note that the labor pooling metrics are symmetrical for a pairwise industry combination $i$,$j$: $LaborCorrelation_{ij} = LaborCorrelation_{ji}$. This is not generally the case for the next two factors discussed, where directional flows are evident.

## 3.2 The Presence of Suppliers and Customers

Marshall (1920) also argues that transportation costs should induce plants to locate close to their inputs, close to their customers, or most likely at some point optimally trading off distance between inputs and customers. To test this theory, we construct metrics of the importance of customer or supplier relationships for pairwise industries. We build our metrics from the 1987 Benchmark Input-Output Accounts published by the Bureau of Economic Analysis. The "Use of Commodities by Industries" table provides commodity-level make and use flows for very detailed industries at the national level, which we aggregate to the SIC3 framework. While some commodities can partly be produced by other industries than the one associated with these commodities, we ignore this distinction and therefore interpret the table's estimates as how much of an industry's production is used as an input to other industries.

We define $Input_{i \leftarrow j}$ as the share of industry $i$'s inputs that come from industry $j$, and $Output_{i \rightarrow j}$ as the share of industry $i$'s outputs that go to industry $j$. These measures run from 0 (no input or output purchasing relationship exists) to 1 (full dependency on the paired

[12] The BLS has recently released a 1983-1998 longitudinal version of the NIOEM. Users should note that the occupations employed in the standardized version differ slightly from those in the 1987 NIOEM we employ. Metrics calculated from the new panel are very close to those used in this paper.

industry). These shares are calculated relative to all input-output flows, including those to non-manufacturing industries or to final consumers.

The strongest relative customer or input dependency is Leather Tanning and Finishing's (311) purchases from Meat Products (201) at 0.39. The highest absolute customer dependency (with a relative share of 23%) is Misc. Plastics Products (308) purchases from Plastic Materials and Synthetics (282). The strongest relative output or supplier dependency is Public Building and Related Furniture's sales to Motor Vehicles and Equipment (371) at 82%. The highest absolute supplier dependency (with a relative share of 32%) is Plastic Materials and Synthetics (282) sales to Misc. Plastics Products (308). Approximately 70% of pairwise combinations have an input-output dependency of less than 0.01%.

This construction results in four potential metrics for a pairwise industry $i$,$j$ combination: $Input_{i \leftarrow j}$, $Input_{j \leftarrow i}$, $Output_{i \rightarrow j}$, and $Output_{j \rightarrow i}$. Unlike the labor pooling metrics, customer and supplier flows are not symmetrical ($Input_{i \leftarrow j} \neq Input_{j \rightarrow i}$). Moreover, the flows between the plastics industries highlight how differences in industry size and the importance of flows to or from non-manufacturing industries and final consumers result in asymmetries between pairwise customer and supplier dependencies ($Input_{i \leftarrow j} \neq Output_{j \rightarrow i}$). To operationalize these metrics for the pairwise coagglomeration regressions, we take either the maximum or the mean of the *Input* and *Output* relationships for the pairwise $i$,$j$ combination. We also examine jointly the input-output role by calculating means and maximums across all four metrics.

## 3.3 Intellectual or Technology Spillovers

Firms may also locate where they are likely to learn from other firms. This learning can take the form of workers learning skills from one another (as Marshall argued) or industrial innovators copying each other (as Saxenian (1994) reports for Silicon Valley). Firms will group near one another either because of the gains from continued presence or because the idea leading to the opening of a new establishment came from an existing concentration of employment in nearby plants. To test this third theory, we develop two metrics of intellectual spillovers that focus specifically on the sourcing of technological innovations. The primary metric is derived from technology flow matrices developed by Scherer (1984); the second metric is derived from patent citations.

Of Marshall's three theories, intellectual spillovers are the most difficult to quantify and to assess empirically. We first note that our metrics focus only on technology spillovers. Other intellectual or information spillovers may exist between industries that are not captured by our design, although technology sourcing is a very important form of knowledge sharing for the manufacturing sector. Second, the discussion below highlights that technology flows are not mutually exclusive to Marshall's first two theories. Technologies embodied in products and machinery can be transferred directly through input-output exchanges. Likewise, industries that share similar labor pools may also be industries between which there is a greater possibility for

intellectual spillovers. Our empirical exercises attempt to isolate technology spillovers by jointly testing with these other two factors, but it is important to note that intellectual spillovers do occur within these channels, too.

### 3.3.1 Scherer (1984) Technology Flows

Scherer (1984) develops a technology flow matrix that estimates the extent to which R&D activity in one industry flows out to benefit another industry. This technology transfer occurs either through a supplier-customer relationship between these two industries or through the likelihood that patented inventions obtained in one industry will find applications in the other industry. We develop two metrics, $TechIn_{i \leftarrow j}$ and $TechOut_{i \rightarrow j}$, for these technology flows that mirror *Input* and *Output* described above. These dependencies are again directional in nature and are calculated relative to total technology flows that include non-manufacturing industries and government R&D. The strongest relative technology flows are associated with Plastic Materials and Synthetics (282) and its relationships to Misc. Plastics Products (308), Tires and Inner Tubes (301), and Industrial Organic Chemicals (286).

The raw technology flows are taken from Table 20.1 of Scherer (1984). Each entry in that table is a dollar amount of 1974 R&D spending in a given industry that is estimated to flow out to benefit another industry. We convert the 38 manufacturing industries reported by Scherer (1984) to the SIC3 framework by apportioning entries through total value of shipments (obtained from the 1987 Census of Manufacturing). For instance, if $T^*_{mn}$ is the entry in Scherer's table corresponding to the dollar flow of benefits from industry $m$ to industry $n$, and $j$ (resp., $i$) is a three-digit industry that is part of industry group $m$ (resp., $n$) and accounts for a fraction $w_j$ (resp., $w_i$) of all shipments in that industry group, then $T_{ji} = w_i w_j \, T^*_{mn}$.

### 3.3.2 Patent Citation Flows

The NBER Patent Data File was originally compiled by Hall *et al.* (2001). This dataset offers detailed records for all patents granted by the United States Patent and Trademark Office (USPTO) from January 1975 to December 1999. Each patent record provides information about the invention (e.g., technology classification, citations of prior art) and the inventors submitting the application (e.g., name, city). Patent citation patterns can be informative about technology diffusion and knowledge exchanges. Griliches (1990) and Jaffe *et al.* (2000) further discuss employing patent citations in this context.

We construct our second knowledge spillovers metric through the patent citations. We restrict the citations data to be citations where both the citing and cited patents are a) applied for after 1975 and b) filed within the US. This sample includes 4,467,625 citations. These citations are first collapsed into a citation matrix using the USPTO technology categories, over 400 in number. Combining the work of Johnson (1999), Silverman (1999) and Kerr (2008), concordances are developed between the USPTO classification scheme and SIC3 industries (a

probabilistic mapping).

The resulting metrics estimate the extent to which technologies associated with industry $i$ cite technologies associated with industry $j$, and vice versa. These $PatIn_{ij}$ and $PatOut_{ij}$ are normalized by total citations for the industries. In practice, there is little directional difference between $PatIn_{ij}$ and $PatOut_{ij}$ due to the extensive number of citations within a single technology field, in which case the probabilistic citing and cited industry distributions are the same. These patent-based metrics have the advantage of covering the 1975-2000 period, but inventor-to-inventor communication patterns represent a subset of the technology flows Scherer (1984) attempts to encompass.

We primarily use the patent citations data to construct the UK instrument for technology flows in the US. As further noted below, using the same technology-to-industry concordance structurally relates the US and UK citation matrices. Thus, it is better to use the UK citation matrices with the Scherer (1984) technology flows.

## 3.4 Shared Natural Advantages

Some regions simply possess better natural environments for certain industries, and agglomeration can follow from these natural cost advantages. Desert areas are inadequate hosts to the logging industry, and areas with cheap electricity attract aluminum producers. Coagglomeration may be observed if two industries are attracted to the same natural advantages, even if the industries would not otherwise have interacted through Marshallian forces. For example, coastal access is independently important for ship building and oil refining industries.

To model the shared interest of two industries on certain natural advantages, we first develop a predicted spatial distribution for each manufacturing industry based upon local cost advantages and industry traits. This work follows Ellison and Glaeser (1999), who model sixteen state-level characteristics that afford natural advantages in terms of natural resources, transportation costs, and labor inputs. Combining these cost differences with each industry's intensity of factor use, Ellison and Glaeser (1999) estimate a spatial distribution of manufacturing activity that would be expected due to cost differences alone (plus population distributions). They find that 20% of observed state-industry manufacturing activity can be explained through these mostly exogenous local factors.[13]

The starting point is a model that assumes that average state-industry profits take the form,

$$\log \pi_{is} = \alpha_0 \log(pop_s) + \alpha_1 \log(mfg_s) - \delta_i \sum\nolimits_{\varrho} \beta_{\varrho} y_{\varrho s} z_{\varrho i},$$

where $i$ indexed industries, $s$ indexes states, and $\varrho$ indexes inputs used in production processes. Profits are increasing in state populations and manufacturing activity, measured in shares, and

[13] Ellison and Glaeser (1999) suggest that this 20% share likely under-estimates the true portion of spatial agglomeration that can be explained through mostly fixed characteristics. Kim (1999) estimates natural regional advantages over a 100-year period.

decreasing in input expenditures. $y_{\varrho s}$ is the cost of input $\varrho$ in state $s$, and $z_{\varrho i}$ is the intensity with which industry $i$ uses input $\varrho$. The specification includes multiplicative industry dummies $\delta_i$ to account for the fact that observed cost differences will affect location decisions more in some industries than in others. These industry shifters are constrained to be non-negative in the estimations. This approach assumes that effects on industry profitability of differences in input costs are proportional to intensities at which inputs are used. Under this model, the expected state-industry shares are:

$$E(S_{is}) = \frac{pop_s^{\alpha_0} mfg_s^{\alpha_1} \exp(-\delta_i \Sigma_\varrho \beta_\varrho y_{\varrho s} z_{\varrho i})}{\sum_{\acute{s}} pop_{\acute{s}}^{\alpha_0} mfg_{\acute{s}}^{\alpha_1} \exp(-\delta_i \Sigma_\varrho \beta_\varrho y_{\varrho \acute{s}} z_{\varrho i})}.$$

Ellison and Glaeser (1999) estimate this relationship using realized state-industry employments.

One can then back out the expected spatial distributions of industries. The spatial shares are constrained to mimic the overall distribution of manufacturing activity. Difference from simple population shares emerge through an interaction of industry intensity for an input combined with spatial concentration of that input. These expected shares are calculated at the four-digit SIC level, which we then aggregate to the SIC3 level of our analysis. We employ these state-industry expected spatial distributions to calculate expected coagglomeration levels for industry pairs. For the EG metric, this is a very straightforward calculation. We simply substitute the predicted industry shares for the actual industry shares in the above formula. The pairwise correlation between expected and actual coagglomeration using this technique is 0.2.[14]

Creating a similar natural advantages baseline for DO is more challenging. Data constraints prevent extending the Ellison and Glaeser (1999) estimations to the county level. Such an extension would have afforded an exact natural advantages replication to our "lumpy" coagglomeration index derived above. Instead, we must continue to employ the state-level expected distributions, which are then apportioned among counties within a state through the observed distribution of manufacturing activity. The density procedures are then applied as described above. Constructing this baseline unfortunately also requires extensive computation time, as the metrics are specific to industry pairs, limiting the permutations that can be run.

The pairwise correlation between expected and actual DO coagglomeration using this technique is 0.4. The higher correlation vis-a-vis the EG natural advantages extends from two sources. First, the more continuous horizon does help identify clustering along natural advantages across state borders (e.g., neighboring coastal states in New England). A second reason is mechanical. Some of the observed DO coagglomeration descends from plant distances in neighboring counties within states. The apportionment procedure unfortunately mechanically relates these.

[14] Glaeser and Kerr (2008) extend the natural advantages estimation approach to the city-industry level, closely following the non-linear least squares approach of Ellison and Glaeser (1999). We employ the state-level estimates in this paper to provide full coverage of the US, but the results are robust to employing expected coagglomeration calculated through PMSA placements, too.

# 4 UK Instrumental Variables

The above US metrics are useful for examining correlations in the data regarding the determinants of coagglomeration. A clear interpretation of the results, however, is limited by concerns of reverse causality. Take our observed importance of customer relationships as an example. Our exposition suggests firms are choosing their geographic locations to be near their customers in order to minimize transportation costs. An alternative explanation of the findings, however, is that these firm locations are determined by other factors (e.g., historical accidents). After these locations are determined, firms choose to sell to nearby industries. These sales are subsequently reflected in the BEA Benchmark Input-Output Accounts, leading to our observed correlations.

To recover a causal assessment, we first develop instruments for our explanatory variables from equivalent data in the UK. Their sources and construction mirror those described for the US and are outlined below. The identifying assumption is that the observed input-output, labor pooling, and technology sourcing relationships among industries in the UK are correlated with the natural inter-industry dependencies but are orthogonal to any endogenous industry inter-dependencies present in the US data that arise from reverse causality.

It is important to note that reliable use of these UK instruments depends upon controlling for coagglomeration due to natural advantages. Otherwise, endogenous Marshallian linkages due to two industries locating near a shared resource could arise simultaneously in both the UK and US. Continuing with the example above, oil refining industries and ship building could start employing similar labor independently in both the US and UK if they are often co-located in major ports. The expected coagglomeration in the US due to shared natural advantages is exogenous and therefore does not require an instrument.

## 4.1 Labor Market Pooling

The UK does not publish a detailed equivalent of the BLS' National Industry-Occupation Employment Matrix. To construct a similar matrix for the UK, we pooled six years of the UK Labour Force Survey (LFS), akin to the US Current Population Survey. We then developed matrices of the occupation-by-industry distribution of currently employed workers by summing over the survey. The included surveys are March-May 2001, June-August 2002, September-November 2003, December 2004-February 2005, and April-June 2006. This pooled dataset contains 224,528 employed workers out of 520,952 respondents; 42,948 work in manufacturing. We maintained the occupation codes Soc2km (353 classifications) and Sc2kmmn (84 classifications) at their detailed level for estimating labor similarities. We mapped the industry code Indm92m (461 classifications, 265 in manufacturing) into the SIC3 system.[15]

[15] We employ a later period than our typical 1987 date to increase the available LFS sample size and questionnaire detail. The period starts after occupation classifications changed in 2001. The staggered surveys avoid double counting as one-fifth of the LFS' respondents rotate out each quarter. From 2005, the data collection periods shift from (mar-may, jun-aug, sep-nov, dec-feb) to (jan-mar, apr-jun, jul-sep, oct-dec).

### 4.2 Presence of Suppliers and Customers

The input-output matrices are taken from Maskus *et al.* (1994) and Maskus and Webster (1995). These researchers began with the 1989 Input-Output Balance for the United Kingdom, published by the Central Statistical Office, London, in 1992. The original table contained 102 sectors; Maskus *et al.* (1994) aggregated the table into 80 sectors that formed the least common denominator with the US tables they were also employing. These tables again include flows out of the manufacturing sector that are used for normalizations. We map the 80 Maskus *et al.* sectors that corresponded to the SIC3 system. The empirical analysis is robust to introducing additional controls regarding this multiplicity. We further consider the UK instruments in regressions that drop SIC3 pairs within the same SIC2.[16]

### 4.3 Intellectual or Technology Spillovers

The UK technology flows matrices are calculated through the NBER patent citations data. The construction mirrors the US citation development documented above, except that we limit the raw sample to those citations where both the citing and cited patents are filed from the UK. 28,134 citations from 1975-1999 are used for these metrics. It is important to note that UK citations are converted from the USPC classification system to the SIC3 framework using the same technology concordances as used for converting the US data. Using the UK citations as an instrument for spillovers measured through the US citations will deliver overstated first-stages by construction. The UK citations are better suited as instruments for the Scherer (1984) technology flow matrices.[17]

## 5 US Spatial Instrumental Variables

In addition to the UK instrument, we develop a second instrument set for labor market pooling and input-output relationships through the spatial distribution of plants in the US. The idea behind these instruments is simple — identify locations where industry A is absent to calculate industry B's traits, and vice versa. We then combine these traits to estimate the likely Marshallian dependency between A and B.

[16] This restriction circumvents a limitation of the UK input-output tables. We explicitly exclude intra-industry flows at the SIC3 level from the US input-output tables. In several cases, we are required to map the same UK industry to multiple SIC3 industries within an SIC2. In these cases, we are not able to distinguish flows across these SIC3 industries from intra-industry flows.

[17] The core element of the USPC-to-industry concordances comes from Canadian data that jointly classified patents into technologies and industries. Thus combining the UK citations with the industry concordances is still excludable for an instrument of US technology flows. By themselves, the UK citations can serve as instruments for US citations when using just the USPC codes; it is the industry conversion that introduces the common structural forms.

## 5.1 Labor Market Pooling

We calculate an industry-occupation matrix for each PMSA from the 1990 Census IPUMS 5% state sample. We choose the state sample over the MSA sample to maximize the number of respondent observations.[18] We only include employed, civilian workers between the ages of 25 and 65 living outside of group quarters. IPUMS has 83 industries in manufacturing that we map to the SIC3 framework. From the 1987 Census of Manufacturing, we order PMSAs by the relative presence of each industry (compared to all manufacturing activity). We choose the 25 PMSAs where industry A is least present (ideally entirely absent) and industry B is most present to calculate industry B's occupational needs. After doing the mirror image calculation, we construct the labor similarity correlation between industries A and B as described above.

Several points regarding this procedure stand-out. First, the occupation characterizations for each industry differ across pairwise industry combinations. This is not the case when the NIOEM data are employed. Moreover, this variation exists even within the same IPUMS industry where it is necessary to map into multiple SIC3 industries. Second, one can improve the IPUMS' sample size used to calculate these labor inputs by imposing minimum bars — that is, we consider all PMSAs with less than 0.1% of industry A's employment as equal candidates for measuring industry B's occupational needs. We have tested several levels of this bar and found robust results.

One reservation about our PMSA-based procedure is that manufacturing activity within cities may have different traits from rural areas. We therefore also calculate the metric using the four Census regions — measuring industry A's traits in the census region where B is least agglomerated relative to total manufacturing — and find very similar results. A second possible reservation is that the bottom 25 cities for industry A may represent different levels of economic activity depending upon the size of the industry. Accordingly, we also test a specification that uses the cities that represent the bottom 1% of industry A's activity relative to all manufacturing to measure industry B's traits. This approach also yields similar results. We have further confirmed that the results are robust to considering separately college and non-college educated workers.

## 5.2 Presence of Suppliers and Customers

For spatial input-output instruments, we employ the "material inputs trailers" from the 1987 Census of Manufacturing. This form asks plants to list the materials that they use and their expenditures. The form is customized by SIC, so that "typical" input needs for industries are listed for ease of entry. That is, logs would be a listed entry for pulp mills but not for space craft. Plants are also allowed to enter additional inputs that they use, but we only know the total of these write-in amounts in the trailers (and not the input type itself). These inputs are

[18] We are able to link 228 IPUMS MSAs to the Census data for this analysis.

at the five-digit SIC level, which we aggregate to the SIC3 level.

We again calculate industry B's input dependence on industry A through the PMSAs and Census regions where industry A is least present. The dependencies are relative to all plant inputs, including non-manufacturing and unspecified industries. The results are robust to using other denominator variants (e.g., manufacturing only, specified inputs only).

Several important notes should again be made. One result of the "typical" input needs format is that we have a greater number of zero-valued observations than building directly from the BEA input-output data. For example, a few computers are sold to meat packing industries, but this will not be reflected in the trailers. This is not important given the very large number of actual material flows that are almost zero (more than 70%). Second, we use this input-based instrumental variable to instrument for total input-output flows; it is not feasible to calculate similar regional measures of output flows.[19]

Finally, a conceptual distinction between the labor and materials instruments should be noted. Our procedure has better application for labor pooling as the intermediary of occupations exists. In choosing locations where industry B is not present, no constraint is placed on the occupational choices made by industry A. There are no intermediaries for input-output flows, however, and our procedure thus partially biases us against finding an input flow from industry B. The observed robustness of the IV construction mitigates this concern.

# 6 Extended Empirical Estimations

Appendix Tables 6-8 provide multiple permutations on the results presented in Tables 3-5 of the main paper. We report bootstrapped standard errors in the main paper to account properly for the generated natural advantages regressor. In these appendix tables, we report robust standard errors so that we can maintain consistency across specifications (e.g., weighting). The robust and bootstrapped standard errors are very similar.

## 6.1 OLS Extended Outcomes

App. Tables 6A-6D provide extended results for the OLS specifications that employ the EG metric. App. Tables 6A and 6B begin by testing the robustness of Table 4's multivariate specifications with and without natural advantages. The first five columns of each table continue with the pairwise maximum concept for the explanatory variables, while the second five columns consider pairwise means instead. Within each group of five, the second estimation tests splitting inputs and outputs, the third tests excluding industry pairs within the same SIC2, the fourth

[19] Turning the material inputs trailers around, for example, suffers in that sales may have come from other regions, especially for low-weight products. We have also looked at the "product trailers" where plants detail their outputs (e.g., Bernard *et al.* 2008). These data cannot support our needs, however, as the "typical" outputs are almost entirely within the same SIC3 as the plant. Again, write-in responses are not specified.

tests including industry fixed effects (FE), and the fifth tests weighing by the employment size of the pairwise industry combination.

The industry FE specifications produce the most dissimilar results and thus warrant further discussion. Specifically, we create "consolidated" FE for the pairwise industries. We have a single vector of industry dummies ranging from 201 to 399. 202 is coded as one if either industry A or B is industry 202, and zero otherwise. This approach is more appropriate than separate FE for industries A and B, as industry ordering within the SIC system is arbitrary, and we consider only unique industry pairs. We have investigated the observed rise in the labor coefficient. It is due to industry combinations within SIC 22. The simple correlation without SIC 22 between labor and the EG state index is 0.070 (0.010), which increases to just 0.142 (0.015) when the industry FE are introduced. As our central conclusions hold in specifications that exclude industry pairs within the same SIC2, we do not believe this aberration is important. We also note that introducing industry FE does not increase the labor coefficient when controlling for natural advantages.

App. Tables 6C and 6D test the robustness of the OLS results to other techniques for calculating the EG metric. We also consider substitutions of the patent-based technology metric for the Scherer technology metric. The most substantive differences to the specifications reported in the main paper are the weakness of the patent-based metric in the state-level specifications and the weakness of labor similarity in the county-level specifications. Overall, however, the results are very similar regardless of how the EG metric is defined.

App. Tables 6E and 6F repeat for the DO metric the analysis presents in App. Tables 6A and 6B for the EG metric. We do not report pairwise mean specifications, however, in favor of presenting both the 1000 mile and 250 mile thresholds. Similar to the univariate correlations noted in Table 3 of the main paper, multivariate estimations find industries with dissimilar labor needs are coagglomerated when the distance threshold is set at 1000 miles. As noted above, this finding is due to neighboring clusters at moderate distance intervals. Some shifts with the industry FE are again evident.

App. Tables 6G and 6H extend the DO estimations to consider three other variations on the DO metric: bilateral firm counts, county-industry employments, and county-industry firm counts. The results are fairly consistent across the metric design, with some residual weakness evident for labor similarities with firm count-based metrics. The patent-based measures are also consistently weaker with the DO specifications than those with Scherer technology flows.

Finally, unreported specifications test the robustness of how the DO global localization index $\Gamma_{ij}$ is aggregated. $\Gamma_{ij} = 0$ for the specifications reported in the main paper if an industry pair is not globally localized, regardless of the extent to which the industry pair is or is not globally dispersed ($\Psi_{ij}$). To ensure that this construction was reasonable for our study, we first looked at parallel regressions that employ $\Psi_{ij}$ as the dependent variable, with $\Psi_{ij} = 0$ if an industry is globally localized. The coefficients suggests that Marshallian factors are found to reduce

dispersion. We also looked at a combined index $\tilde{\Gamma}_{ij}^{Emp}$:

$$\begin{aligned}\tilde{\Gamma}_{ij}^{Emp} \equiv\ & \Gamma_{ij}^{Emp} \text{ if } \Gamma_{ij}^{Emp} > 0 \\ & -\Psi_{ij} \text{ otherwise}\end{aligned}$$

This combined index also delivers similar outcomes to our core estimations.

## 6.2 IV Extended Outcomes

App. Table 7A provides the first-stage statistics for the UK-based instrument. The individual first stages are all very strong, and this strength carries over into specifications that simultaneously instrument for labor and input-outputs. We consistently pass relevant tests regarding weak instruments. This is not true, however, in our attempts to further instrument for technology flows, as the technology IV does not adequately distinguish itself from input-outputs. We therefore limit the IV analysis to just labor and input-outputs in the main paper.

App. Table 7B provides extended IV results for the EG metric. The results are consistent across the different spatial demarcations with the exception of the labor metrics in the county-level estimations. This weakness is similar to the OLS outcomes noted above. App. Table 7C provides the triple IV results. We report these for consistency with our earlier NBER working paper, but we no longer emphasis them. Triple IVs with the DO metric display weaknesses similar to those with the EG metric and are not reported.

App. Table 7D provides extended DO outcomes for the dual instruments. The most remarkable feature is the strength of the labor similarities metric over moderate distance thresholds. The explanatory power of the input-output metric is more consistent. We focus attention on the results employing the 250 mile threshold.

Tables 8A-8D provide similar materials for the US Spatial IV regressions. The estimations are fairly similar to the UK-based IV results. The most prominent differences are the more reasonable magnitudes for labor pooling regardless of the DO distance threshold employed. Regressions instrumenting for technology also perform better, but we remain skeptical of the strength of the technology spillovers instrument.

# References


[1] Arzaghi, Mohammad, and Henderson, J. Vernon (2008): "Networking off Madison Avenue," *Review of Economic Studies*, forthcoming.

[2] Autor, David, Kerr, William and Kugler, Adriana (2007): "Does Employment Protection Reduce Productivity? Evidence from U.S. States," *The Economic Journal*, 117, 189–217.

[3] Bernard, Andrew, Redding, Stephen, and Schott, Peter (2008): "Multi-Product Firms and Product Switching," Working Paper.

[4] Combes, Pierre-Philippe and Duranton, Gilles (2006): "Labour Pooling, Labour Poaching, and Spatial Clustering," *Regional Science and Urban Economics*, 36, 1-28.

[5] Davis, Steven, Haltiwanger, John and Schuh, Scott (1996): *Job Creation and Destruction.* Cambridge, MA: MIT Press.

[6] Diamond, Charles and Simon, Curtis (1990): "Industrial Specialization and the Returns to Labor," *Journal of Labor Economics*, 8, 175–201.

[7] Dumais, Guy, Ellison, Glenn and Glaeser, Edward L. (2002): "Geographic Concentration as a Dynamic Process," *Review of Economics and Statistics*, 84, 193–204.

[8] Dunne, Timothy, Roberts, Mark and Samuelson, Larry (1989): "Patterns of Firm Entry and Exit in U.S. Manufacturing Industries," *The RAND Journal of Economics*, 19, 495–515.

[9] Duranton, Gilles and Overman, Henry (2005): "Testing for Localization Using Micro-Geographic Data," *Review of Economic Studies*, 72, 1077–1106.

[10] Duranton, Gilles and Overman, Henry (2008): "Exploring the Detailed Location Patterns of UK Manufacturing Industries Using Microgeographic Data," *Journal of Regional Science*, 48:1, 313–343 (previously CEPR Working Paper 5858).

[11] Ellison, Glenn and Glaeser, Edward (1997): "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach, " *Journal of Political Economy*, 105, 889–927.

[12] Ellison, Glenn and Glaeser, Edward (1999): "The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration?", *American Economic Review Papers and Proceedings*, 89, 311–316.

[13] Ellison, Glenn, Glaeser, Edward and Kerr, William (2009): "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns," Working Paper. Previous version released as NBER Working Paper 13068, CES Working Paper 07-13, and HBS Working Paper 07-064. Second revision circulated in August 2008.

[14] Ellison, Glenn, Glaeser, Edward and Kerr, William (2006): "The Impact of the SIC-NAICS Conversion on Industrial Organization Metrics: Evidence Building from Establishment Data," Census Bureau Technical Paper.

[15] Glaeser, Edward and Kerr, William (2008): "Local Industrial Conditions and Entrepreneurship: How Much of the Spatial Distribution Can We Explain?," *Journal of Economics and Management Strategy*, forthcoming.

[16] Griliches, Zvi (1990): "Patent Statistics as Economic Indicators: A Survey," *Journal of Economic Literature,* 28, 1661–1707.

[17] Hall, Bronwyn, Jaffe, Adam and Trajtenberg, Manuel (2001): "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools," NBER Working Paper 8498.

[18] Harrison, Brett and Kominers, Scott (2008): "A Concentration-Based Index of Industrial Agglomeration," Working Paper.

[19] Jaffe, Adam, Trajtenberg, Manuel and Fogarty, Michael (2000): "Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors," *Papers and Proceedings, American Economic Review*, 90, 215–218.

[20] Johnson, Daniel (1999): "150 Years of American Invention: Methodology and a First Geographic Application," Wellesley College Economics Working Paper 99-01.

[21] Kerr, William (2008): "Ethnic Scientific Communities and International Technology Diffusion," *The Review of Economics and Statistics*, 90, 518–537.

[22] Kerr, William and Nanda, Ramana (2008): "Democratizing Entry: Banking Deregulations, Financing Constraints, and Entrepreneurship," *Journal of Financial Economics*, forthcoming.

[23] Kim, Sukkoo (1999): "Regions, Resources, and Economic Geography: Sources of U.S. Regional Comparative Advantage, 1880-1987," *Regional Science and Urban Economics*, 29, 123–137.

[24] Kolko, Jed (2000): "Essays on Information Technology, Cities, and Location Choices," Harvard University Ph.D. Thesis.

[25] Krugman, Paul (1991): *Geography and Trade*. Cambridge, MA: MIT Press.

[26] Marshall, Alfred (1920): *Principles of Economics*. London, UK: MacMillan and Co.

[27] Maskus, Keith, Sveikauskas, C. and Webster, Allan (1994): "The Composition of the Human Capital Stock and Its Relation to International Trade: Evidence from the U.S. and Britain," *Weltwirtschaftliches Archiv*, 1994, Band 130, Heft 1.

[28] Maskus, Keith and Webster, Allan (1995): "Factor Specialization in U.S. and U.K. Trade: Simple Departures from the Factor-Content Theory," *Swiss Journal of Economics and Statistics*, 1.

[29] McGuckin, Robert and Peck, Suzanne (1992): "Manufacturing Establishments Reclassified into New Industries: The Effect of Survey Design Rules," Center for Economic Studies 92-14 (U.S. Bureau of the Census).

[30] Rotemberg, Julio and Saloner, Garth (2000): "Competition and Human Capital Accumulation: A Theory of Interregional Specialization and Trade," *Regional Science and Urban Economics*, 30, 373–404.

[31] Saxenian, AnnaLee (1994): *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Cambridge, MA: Harvard University Press.

[32] Scherer, Frederic M. (1984): "Using Linked Patent Data and R&D Data to Measure Technology Flows," in Griliches, Zvi (ed.) *R & D, Patents and Productivity*. Chicago, IL: The University of Chicago Press.

[33] Silverman, Brian (1999): "Technological Resources and the Direction of Corporate Diversification: Toward an Integration of the Resource-Based View and Transaction Cost Economics," *Management Science*, 45, 1109–1124.

**App. Table 1: Descriptive Statistics for Pairwise Coagglomeration Regressions**

| | Mean | Stand. Dev. | Minimum | Maximum |
|---|---|---|---|---|
| | | *A. Pairwise EG Coaggl. Measures* | | |
| EG State Total Empl. Coaggl. | 0.000 | 0.013 | -0.065 | 0.207 |
| EG PMSA Total Empl. Coaggl. | 0.000 | 0.006 | -0.025 | 0.119 |
| EG County Total Empl. Coaggl. | 0.000 | 0.003 | -0.018 | 0.080 |
| EG State Firm Birth Empl. Coaggl. | 0.000 | 0.015 | -0.082 | 0.259 |
| EG Expected Coaggl. Due to Natural Advantages | 0.000 | 0.001 | -0.008 | 0.022 |
| | | *B. Marshallian Factors* | | |
| Labor Correlation | 0.470 | 0.226 | -0.046 | 1.000 |
| Input-Output Maximum | 0.007 | 0.029 | 0.000 | 0.823 |
| Input-Output Mean | 0.002 | 0.010 | 0.000 | 0.240 |
| Input Maximum | 0.005 | 0.019 | 0.000 | 0.392 |
| Input Mean | 0.002 | 0.010 | 0.000 | 0.196 |
| Output Maximum | 0.005 | 0.026 | 0.000 | 0.823 |
| Output Mean | 0.002 | 0.013 | 0.000 | 0.411 |
| Scherer R&D Tech Maximum | 0.005 | 0.026 | 0.000 | 0.625 |
| Scherer R&D Tech Mean | 0.002 | 0.010 | 0.000 | 0.263 |
| Patent Citation Tech Maximum | 0.015 | 0.025 | 0.000 | 0.400 |
| Patent Citation Tech Mean | 0.007 | 0.014 | 0.000 | 0.203 |

Notes: See Table 1.

**App. Table 2A: Levels of EG Agglomeration and Coagglomeration 1972-1997**

| | 1972 | 1977 | 1982 | 1987 | 1992 | 1997 |
|---|---|---|---|---|---|---|
| | | | *A. State-Level Total Employment* | | | |
| EG Agglomeration Index γ Mean | 0.0398 | 0.0399 | 0.0392 | 0.0368 | 0.0351 | 0.0342 |
| EG Coagglomeration Index γc Mean | 0.0003 | 0.0003 | 0.0002 | 0.0004 | 0.0003 | 0.0003 |
| EG Coagglomeration Index γc SD | 0.0150 | 0.0139 | 0.0140 | 0.0133 | 0.0129 | 0.0124 |
| | | | *B. PMSA-Level Total Employment* | | | |
| EG Agglomeration Index γ Mean | 0.0298 | 0.0292 | 0.0286 | 0.0285 | 0.0271 | 0.0299 |
| EG Coagglomeration Index γc Mean | 0.0003 | 0.0003 | 0.0002 | 0.0003 | 0.0002 | 0.0002 |
| EG Coagglomeration Index γc SD | 0.0086 | 0.0075 | 0.0069 | 0.0061 | 0.0054 | 0.0060 |
| | | | *C. State-Level Employment in Firm Births* | | | |
| EG Agglomeration Index γ Mean | 0.0290 | 0.0022 | 0.0121 | 0.0107 | 0.0158 | 0.0285 |
| EG Coagglomeration Index γc Mean | 0.0001 | 0.0003 | 0.0003 | 0.0005 | 0.0004 | 0.0003 |
| EG Coagglomeration Index γc SD | 0.0193 | 0.0172 | 0.0177 | 0.0150 | 0.0187 | 0.0181 |

Notes: Measures of EG industrial agglomeration and coagglomeration calculated from the Census of Manufacturers. Estimates include all manufacturing SIC3 industries, except those listed in the text, for 134 observations per year.

**App. Table 2B: Correlation of EG Coagglomeration Index**

| | 1972 | 1977 | 1982 | 1987 | 1992 |
|---|---|---|---|---|---|
| 1977 | 0.953 | | | | |
| 1982 | 0.891 | 0.944 | | | |
| 1987 | 0.841 | 0.889 | 0.936 | | |
| 1992 | 0.791 | 0.840 | 0.895 | 0.959 | |
| 1997 | 0.740 | 0.789 | 0.832 | 0.890 | 0.941 |

Notes: See App. Table 2A. EG Coagglomeration Index measured through state total employments for each industry.

**App. Table 2C: Inter-Industry Averages of 1987 EG Pairwise Coagglomerations**

| | 20 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20 | 0.002 | | | | | | | | | | | | | | | | | | |
| 22 | -0.003 | 0.102 | | | | | | | | | | | | | | | | | |
| 23 | 0.000 | 0.021 | 0.012 | | | | | | | | | | | | | | | | |
| 24 | 0.003 | 0.012 | 0.003 | 0.013 | | | | | | | | | | | | | | | |
| 25 | -0.001 | 0.016 | 0.002 | 0.002 | 0.000 | | | | | | | | | | | | | | |
| 26 | 0.001 | 0.012 | 0.004 | 0.006 | -0.001 | 0.005 | | | | | | | | | | | | | |
| 27 | 0.001 | -0.001 | 0.001 | -0.003 | -0.004 | 0.000 | 0.004 | | | | | | | | | | | | |
| 28 | 0.001 | 0.004 | 0.001 | 0.000 | -0.002 | 0.002 | 0.001 | 0.007 | | | | | | | | | | | |
| 29 | 0.004 | -0.018 | -0.003 | 0.000 | -0.006 | -0.002 | 0.001 | 0.008 | 0.013 | | | | | | | | | | |
| 30 | -0.001 | 0.003 | -0.003 | -0.001 | 0.001 | 0.000 | -0.001 | 0.001 | -0.001 | 0.002 | | | | | | | | | |
| 31 | 0.000 | -0.005 | 0.006 | -0.001 | -0.003 | 0.005 | 0.006 | 0.001 | -0.003 | -0.003 | 0.019 | | | | | | | | |
| 32 | 0.001 | 0.001 | 0.002 | 0.000 | -0.002 | 0.000 | 0.000 | 0.003 | 0.006 | 0.001 | -0.001 | 0.003 | | | | | | | |
| 33 | -0.001 | -0.012 | -0.006 | -0.002 | -0.004 | 0.001 | -0.001 | 0.001 | 0.002 | 0.004 | -0.003 | 0.004 | 0.010 | | | | | | |
| 34 | -0.001 | -0.014 | -0.007 | -0.004 | -0.002 | -0.002 | 0.000 | -0.001 | 0.001 | 0.002 | -0.003 | 0.000 | 0.005 | 0.004 | | | | | |
| 35 | 0.000 | -0.011 | -0.006 | -0.003 | -0.002 | -0.001 | 0.001 | -0.001 | 0.000 | 0.001 | -0.001 | -0.001 | 0.003 | 0.004 | 0.001 | | | | |
| 36 | 0.001 | -0.007 | -0.001 | -0.002 | 0.000 | -0.003 | 0.001 | -0.002 | 0.000 | 0.000 | 0.002 | -0.001 | -0.001 | 0.000 | 0.000 | 0.000 | | | |
| 37 | -0.001 | -0.017 | -0.008 | -0.001 | 0.001 | -0.004 | -0.002 | -0.004 | 0.000 | -0.002 | -0.008 | -0.002 | 0.004 | 0.004 | 0.001 | -0.001 | -0.004 | | |
| 38 | -0.002 | -0.010 | 0.005 | -0.005 | -0.003 | -0.002 | 0.006 | -0.003 | -0.005 | -0.005 | 0.009 | -0.003 | -0.005 | -0.002 | 0.000 | 0.002 | -0.004 | 0.008 | |
| 39 | -0.001 | -0.007 | 0.005 | -0.004 | -0.004 | 0.000 | 0.005 | -0.001 | -0.003 | -0.003 | 0.010 | -0.002 | -0.002 | -0.001 | 0.000 | 0.003 | -0.006 | 0.012 | 0.014 |

Notes: Table entries are the weighted-average pairwise SIC3 coagglomerations within the pairwise SIC2 cell. EG Coagglomeration Index measured through state total employments for each industry.

**App. Table 3A: Descriptive Statistics for Pairwise DO Coagglomeration Measures**

| | Industry Count | Relevant Industries (non-zero) | | |
|---|---|---|---|---|
| | | Mean | Stand. Dev. | Maximum |
| DO Global Localization Emp. Coaggl., 1000 mi. | 7371 | 0.133 | 0.073 | 0.454 |
| DO Global Dispersion | 10 | 0.592 | 0.078 | 0.746 |
| DO Localization and Dispersion < 0.001 | 21 | | | |
| DO Global Localization Count Coaggl., 1000 mi. | 7304 | 0.086 | 0.061 | 0.372 |
| DO Global Dispersion | 77 | 0.272 | 0.095 | 0.591 |
| DO Localization and Dispersion < 0.001 | 92 | | | |
| DO Expected Global Localization Coaggl., 1000 mi. | 7381 | 0.181 | 0.027 | 0.256 |
| DO Global Dispersion | 0 | | | |
| DO Localization and Dispersion < 0.001 | 0 | | | |
| DO Global Localization Emp. Coaggl., 500 mi. | 7235 | 0.051 | 0.039 | 0.437 |
| DO Global Dispersion | 146 | 0.145 | 0.058 | 0.400 |
| DO Localization and Dispersion < 0.001 | 134 | | | |
| DO Global Localization Count Coaggl., 500 mi. | 6858 | 0.037 | 0.030 | 0.335 |
| DO Global Dispersion | 523 | 0.088 | 0.027 | 0.207 |
| DO Localization and Dispersion < 0.001 | 264 | | | |
| DO Expected Global Localization Coaggl., 500 mi. | 7381 | 0.083 | 0.015 | 0.136 |
| DO Global Dispersion | 0 | | | |
| DO Localization and Dispersion < 0.001 | 0 | | | |
| DO Global Localization Emp. Coaggl., 250 mi. | 6429 | 0.017 | 0.019 | 0.283 |
| DO Global Dispersion | 952 | 0.042 | 0.029 | 0.307 |
| DO Localization and Dispersion < 0.001 | 530 | | | |
| DO Global Localization Count Coaggl., 250 mi. | 6228 | 0.014 | 0.015 | 0.204 |
| DO Global Dispersion | 1153 | 0.030 | 0.010 | 0.071 |
| DO Localization and Dispersion < 0.001 | 708 | | | |
| DO Expected Global Localization Coaggl., 250 mi. | 7381 | 0.029 | 0.010 | 0.077 |
| DO Global Dispersion | 0 | | | |
| DO Localization and Dispersion < 0.001 | 6 | | | |
| DO Global Localization Emp. Coaggl., 100 mi. | 4557 | 0.006 | 0.008 | 0.124 |
| DO Global Dispersion | 2824 | 0.012 | 0.016 | 0.213 |
| DO Localization and Dispersion < 0.001 | 1327 | | | |
| DO Global Localization Count Coaggl., 100 mi. | 4394 | 0.006 | 0.007 | 0.094 |
| DO Global Dispersion | 2987 | 0.009 | 0.004 | 0.030 |
| DO Localization and Dispersion < 0.001 | 945 | | | |
| DO Expected Global Localization Coaggl., 100 mi. | 6802 | 0.007 | 0.005 | 0.043 |
| DO Global Dispersion | 579 | 0.016 | 0.024 | 0.210 |
| DO Localization and Dispersion < 0.001 | 828 | | | |

Notes: See Table 1. DO coagglomeration metrics are calculated from the 1997 Census of Manufacturers. The distance threshold for determining global localization or dispersion is adjusted across DO row groupings.

**App. Table 3B: Highest DO Pairwise Coagglomerations**

| Rank | Industry 1 | Industry 2 | Coaggl. |
|---|---|---|---|
| | *A. DO Index using 1997 Firm Employments, 1000 mi. Threshold* | | |
| 1 | Carpets and Rugs (227) | Yarn and Thread Mills (228) | 0.454 |
| 2 | Broadwoven Mills, Fiber (222) | Yarn and Thread Mills (228) | 0.450 |
| 3 | Knitting Mills (225) | Yarn and Thread Mills (228) | 0.427 |
| 4 | Yarn and Thread Mills (228) | Metalworking Machinery (354) | 0.426 |
| 5 | Yarn and Thread Mills (228) | Blast Furnace, Basic Steel Products (331) | 0.414 |
| 6 | Broadwoven Mills, Fiber (222) | Carpets and Rugs (227) | 0.410 |
| 7 | Yarn and Thread Mills (228) | Metal Forgings and Stampings (346) | 0.404 |
| 8 | Broadwoven Mills, Fiber (222) | Metalworking Machinery (354) | 0.403 |
| 9 | Broadwoven Mills, Fiber (222) | Knitting Mills (225) | 0.403 |
| 10 | Carpets and Rugs (227) | Metalworking Machinery (354) | 0.397 |
| 11 | Carpets and Rugs (227) | Blast Furnace, Basic Steel Products (331) | 0.394 |
| 12 | Knitting Mills (225) | Carpets and Rugs (227) | 0.394 |
| 13 | Narrow Fabric Mills (224) | Yarn and Thread Mills (228) | 0.391 |
| 14 | Carpets and Rugs (227) | Tires and Inner Tubes (301) | 0.388 |
| 15 | Broadwoven Mills, Cotton (221) | Yarn and Thread Mills (228) | 0.388 |
| | *B. DO Index using 1997 Firm Employments, 250 mi. Threshold* | | |
| 1 | Broadwoven Mills, Fiber (222) | Yarn and Thread Mills (228) | 0.283 |
| 2 | Carpets and Rugs (227) | Yarn and Thread Mills (228) | 0.262 |
| 3 | Broadwoven Mills, Fiber (222) | Carpets and Rugs (227) | 0.226 |
| 4 | Broadwoven Mills, Cotton (221) | Yarn and Thread Mills (228) | 0.219 |
| 5 | Broadwoven Mills, Cotton (221) | Carpets and Rugs (227) | 0.218 |
| 6 | Footwear Cut Stock (313) | Costume Jewelry and Notions (396) | 0.217 |
| 7 | Jewelry, Silverware, Plated Ware (391) | Costume Jewelry and Notions (396) | 0.208 |
| 8 | Knitting Mills (225) | Yarn and Thread Mills (228) | 0.200 |
| 9 | Broadwoven Mills, Fiber (222) | Knitting Mills (225) | 0.190 |
| 10 | Broadwoven Mills, Cotton (221) | Broadwoven Mills, Fiber (222) | 0.175 |
| 11 | Textile Finishing (226) | Yarn and Thread Mills (228) | 0.163 |
| 12 | Footwear Cut Stock (313) | Jewelry, Silverware, Plated Ware (391) | 0.157 |
| 13 | Handbags (317) | Costume Jewelry and Notions (396) | 0.151 |
| 14 | Broadwoven Mills, Cotton (221) | Knitting Mills (225) | 0.149 |
| 15 | Women’s and Misses’ Outerwear (233) | Costume Jewelry and Notions (396) | 0.149 |

Notes: See App. Table 3A.

**App. Table 3C: EG and DO Coaggl. Correlations**

| | DO Empl. Loc. Coaggl. 1000 mi. | DO Empl. Loc. Coaggl. 250 mi. | EG State Employment Coaggl. | EG PMSA Employment Coaggl. | EG County Employment Coaggl. |
|---|---|---|---|---|---|
| DO Global Localization Emp. Coaggl., 250 mi. | 0.408 | | | | |
| EG State Total Empl. Coaggl. | 0.177 | 0.559 | | | |
| EG PMSA Total Empl. Coaggl. | 0.153 | 0.424 | 0.611 | | |
| EG County Total Empl. Coaggl. | 0.179 | 0.331 | 0.515 | 0.796 | |
| EG State Birth Empl. Coaggl. | 0.186 | 0.281 | 0.333 | 0.241 | 0.219 |

Notes: See Table 1. Pairwise correlations calculated over sample used for 1987 pairwise coagglomeration regressions.

**App. Table 3D: Correlation of DO Employment Coaggl. Index**

| | DO Global Loc. Emp. Coaggl. 1000 mi. | DO Global Loc. Emp. Coaggl. 500 mi. | DO Global Loc. Emp. Coaggl. 250 mi. |
|---|---|---|---|
| DO Global Localization Emp. Coaggl., 500 mi. | 0.799 | | |
| DO Global Localization Emp. Coaggl., 250 mi. | 0.408 | 0.752 | |
| DO Global Localization Emp. Coaggl., 100 mi. | 0.082 | 0.371 | 0.798 |

**App. Table 3E: Correlation of DO Count Coaggl. Index**

| | DO Global Loc. Count Coaggl. 1000 mi. | DO Global Loc. Count Coaggl. 500 mi. | DO Global Loc. Count Coaggl. 250 mi. |
|---|---|---|---|
| DO Global Localization Count Coaggl., 500 mi. | 0.855 | | |
| DO Global Localization Count Coaggl., 250 mi. | 0.442 | 0.757 | |
| DO Global Localization Count Coaggl., 100 mi. | 0.160 | 0.464 | 0.896 |

**App. Table 3F: Correlation of DO Expected Employment Coaggl. Index**

| | DO Global Loc. Emp. Expected Coaggl. 1000 mi. | DO Global Loc. Emp. Expected Coaggl. 500 mi. | DO Global Loc. Emp. Expected Coaggl. 250 mi. |
|---|---|---|---|
| DO Expected Global Localization Coaggl., 500 mi. | 0.777 | | |
| DO Expected Global Localization Coaggl., 250 mi. | 0.159 | 0.481 | |
| DO Expected Global Localization Coaggl., 100 mi. | -0.139 | 0.142 | 0.821 |

**App. Table 3G: Correlation of DO Coaggl. Indices, 1000 Mile**

| | DO Global Loc. With Bilateral Firm Employments | DO Global Loc. With Bilateral Firm Counts | DO Global Loc. With Cnty-Ind. Employments |
|---|---|---|---|
| DO Global Localization Bilateral Firm Counts | 0.893 | | |
| DO Global Localization County-Industry Empl. | 0.911 | 0.753 | |
| DO Global Localization County-Industry Counts | 0.887 | 0.996 | 0.747 |

**App. Table 3H: Correlation of DO Coaggl. Indices, 250 Mile**

| | DO Global Loc. With Bilateral Firm Employments | DO Global Loc. With Bilateral Firm Counts | DO Global Loc. With Cnty-Ind. Employments |
|---|---|---|---|
| DO Global Localization Bilateral Firm Counts | 0.836 | | |
| DO Global Localization County-Industry Empl. | 0.888 | 0.674 | |
| DO Global Localization County-Industry Counts | 0.832 | 0.994 | 0.675 |

**App. Table 4A: Highest Labor Correlation Metrics**

| Industry 1 | Industry 2 | Labor Cor. |
|---|---|---|
| Motor Vehicles and Equipment (371) | Railroad Equipment (374) | 0.984 |
| Motor Vehicles and Equipment (371) | Motorcycles, Bicycles, and Parts (375) | 0.984 |
| Motor Vehicles and Equipment (371) | Miscellaneous Transportation Equipment (379) | 0.984 |
| Musical Instruments (393) | Toys and Sporting Goods (394) | 0.979 |
| Toys and Sporting Goods (394) | Pens, Pencils, Office & Art Suppliers (395) | 0.979 |

**App. Table 4B: Lowest Labor Correlation Metrics**

| Industry 1 | Industry 2 | Labor Cor. |
|---|---|---|
| Logging (241) | Aircrafts and Parts (372) | -0.046 |
| Logging (241) | Engines and Turbines (351) | -0.029 |
| Logging (241) | Motor Vehicles and Equipment (371) | -0.029 |
| Logging (241) | Guided Missiles, Space Vehicles, Parts (376) | -0.029 |
| Logging (241) | Metalworking Machinery (354) | -0.021 |

**App. Table 4C: Highest Relative Customer Dependencies Metrics**

| Using Industry | Source Industry | Input Vol. | Input Share |
|---|---|---|---|
| Leather Tanning and Finishing (311) | Meat Products (201) | 872 | 0.392 |
| Sawmills and Planing Mills (242) | Logging (241) | 6811 | 0.360 |
| Leather Gloves and Mittens (315) | Leather Tanning and Finishing (311) | 58 | 0.345 |
| Yarn and Thread Mills (228) | Plastics Materials and Synthetics (282) | 2154 | 0.309 |
| Wood Containers (244) | Sawmills and Planing Mills (242) | 548 | 0.271 |

**App. Table 4D: Highest Absolute Customer Dependencies Metrics**

| Using Industry | Source Industry | Input Vol. | Input Share |
|---|---|---|---|
| Misc. Plastics Products (308) | Plastics Materials and Synthetics (282) | 13,999 | 0.229 |
| Motor Vehicles and Equipment (371) | Metal Forgings and Stampings (346) | 11,378 | 0.055 |
| Plastics Materials and Synthetics (282) | Industrial Organic Chemicals (286) | 9903 | 0.243 |
| Fabricated Structural Metal Products (344) | Blast Furnace and Basic Steel Products (331) | 7607 | 0.196 |
| Metal Forgings and Stampings (346) | Blast Furnace and Basic Steel Products (331) | 7011 | 0.249 |

**App. Table 4E: Highest Relative Supplier Dependencies Metrics**

| Source Industry | Using Industry | Output Vol. | Output Share |
|---|---|---|---|
| Public Building and Related Furniture (253) | Motor Vehicles and Equipment (371) | 1681 | 0.823 |
| Cement, Hydraulic (324) | Concrete, Gypsum, and Plaster Products (327) | 3380 | 0.819 |
| Primary Nonferrous Metals (333) | Nonferrous Rolling and Drawing (335) | 5750 | 0.504 |
| Metal Cans and Shipping Containers (341) | Beverages (208) | 5768 | 0.491 |
| Logging (241) | Sawmills and Planing Mills (242) | 6811 | 0.440 |

**App. Table 4F: Highest Absolute Supplier Dependencies Metrics**

| Source Industry | Using Industry | Output Vol. | Output Share |
|---|---|---|---|
| Plastics Materials and Synthetics (282) | Misc. Plastics Products (308) | 13,999 | 0.322 |
| Metal Forgings and Stampings (346) | Motor Vehicles and Equipment (371) | 11,378 | 0.401 |
| Industrial Organic Chemicals (286) | Plastics Materials and Synthetics (282) | 9903 | 0.179 |
| Blast Furnace and Basic Steel Products (331) | Fabricated Structural Metal Products (344) | 7607 | 0.153 |
| Blast Furnace and Basic Steel Products (331) | Metal Forgings and Stampings (346) | 7011 | 0.141 |

**App. Table 4G: Highest Relative Technology Input Dependencies Metrics**

| Using Industry | Source Industry | Input Vol. | Input Share |
|---|---|---|---|
| Misc. Plastics Products (308) | Plastics Materials and Synthetics (282) | 104 | 0.217 |
| Rubber and Plastics Footwear (302) | Plastics Materials and Synthetics (282) | 1 | 0.200 |
| Tires and Inner Tubes (301) | Plastics Materials and Synthetics (282) | 48 | 0.165 |
| Fabricated Rubber Products (306) | Plastics Materials and Synthetics (282) | 11 | 0.131 |
| Hose, Belting, Gaskets, and Packing (305) | Plastics Materials and Synthetics (282) | 5 | 0.116 |

**App. Table 4H: Highest Absolute Technology Input Dependencies Metrics**

| Using Industry | Source Industry | Input Vol. | Input Share |
|---|---|---|---|
| Misc. Plastics Products (308) | Plastics Materials and Synthetics (282) | 104 | 0.217 |
| Tires and Inner Tubes (301) | Plastics Materials and Synthetics (282) | 48 | 0.165 |
| Plastics Materials and Synthetics (282) | Industrial Organic Chemicals (286) | 24 | 0.040 |
| Aircrafts and Parts (372) | Computers and Office Equipment (357) | 21 | 0.039 |
| Petroleum Refining (291) | Computers and Office Equipment (357) | 19 | 0.043 |

**App. Table 4I: Highest Relative Technology Supplier Dependencies Metrics**

| Source Industry | Using Industry | Output Vol. | Output Share |
|---|---|---|---|
| Plastics Materials and Synthetics (282) | Misc. Plastics Products (308) | 104 | 0.172 |
| Textile Finishing (226) | Misc. Plastics Products (308) | 2 | 0.146 |
| Ordnance and Accessories (348) | Guided Missiles, Space Vehicles, Parts (376) | 3 | 0.133 |
| Broadwoven Mills, Fiber (222) | Misc. Plastics Products (308) | 2 | 0.086 |
| Industrial Organic Chemicals (286) | Plastics Materials and Synthetics (282) | 24 | 0.081 |

**App. Table 4J: Highest Absolute Technology Supplier Dependencies Metrics**

| Source Industry | Using Industry | Output Vol. | Output Share |
|---|---|---|---|
| Plastics Materials and Synthetics (282) | Misc. Plastics Products (308) | 104 | 0.172 |
| Plastics Materials and Synthetics (282) | Tires and Inner Tubes (301) | 48 | 0.080 |
| Industrial Organic Chemicals (286) | Plastics Materials and Synthetics (282) | 24 | 0.081 |
| Computers and Office Equipment (357) | Aircrafts and Parts (372) | 21 | 0.018 |
| Computers and Office Equipment (357) | Petroleum Refining (291) | 19 | 0.017 |

**App. Table 4K: Highest Expected EG Coagglomeration from Natural Advantages**

| Industry 1 | Industry 2 | Coaggl. |
|---|---|---|
| Cement, Hydraulic (324) | Primary Nonferrous Metals (333) | 0.022 |
| Cement, Hydraulic (324) | Blast Furnace and Basic Steel Products (331) | 0.016 |
| Industrial Inorganic Chemicals (281) | Primary Nonferrous Metals (333) | 0.014 |
| Blast Furnace and Basic Steel Products (331) | Primary Nonferrous Metals (333) | 0.012 |
| Industrial Inorganic Chemicals (281) | Cement, Hydraulic (324) | 0.011 |

**App. Table 4L: Highest Expected DO Coaggl. from Shared Natural Advantages, 1000 mi.**

| Industry 1 | Industry 2 | Coaggl. |
|---|---|---|
| Paperboard Mills (263) | Blast Furnace and Basic Steel Products (331) | 0.256 |
| Broadwoven Fabric Mills (222) | Paperboard Mills (263) | 0.249 |
| Paperboard Mills (263) | Iron and Steel Foundries (332) | 0.246 |
| Paper Mills (262) | Blast Furnace and Basic Steel Products (331) | 0.245 |
| Broadwoven Fabric Mills (222) | Blast Furnace and Basic Steel Products (331) | 0.244 |

**App. Table 4M: Highest Expected DO Coaggl. from Shared Natural Advantages, 250 mi.**

| Industry 1 | Industry 2 | Coaggl. |
|---|---|---|
| Watches, Clocks, Watchcases, and Parts (387) | Jewelry, Silverware, and Plated Ware (391) | 0.077 |
| Periodicals (272) | Watches, Clocks, Watchcases, and Parts (387) | 0.075 |
| Periodicals (272) | Jewelry, Silverware, and Plated Ware (391) | 0.072 |
| Aircraft and Parts (372) | Watches, Clocks, Watchcases, and Parts (387) | 0.070 |
| Books (273) | Watches, Clocks, Watchcases, and Parts (387) | 0.068 |

**App. Table 5: Natural Advantages and Marshallian Correlations**

| | EG Natural Advantages | DO Natural Advantages 1000 mi. | DO Natural Advantages 250 mi. | Labor Correlation | Input-Output | Technology Flows Scherer |
|---|---|---|---|---|---|---|
| DO Natural Advantages 1000 mi. | 0.119 | | | | | |
| DO Natural Advantages 250 mi. | 0.083 | 0.159 | | | | |
| Labor Correlation | 0.191 | 0.007 | 0.111 | | | |
| Input-Output | 0.077 | 0.068 | -0.031 | 0.131 | | |
| Technology - Scherer | 0.132 | 0.028 | -0.005 | 0.125 | 0.327 | |
| Technology - Patents | 0.055 | 0.059 | 0.128 | 0.226 | 0.252 | 0.286 |

Notes: See Table 1. Pairwise correlations calculated over sample used for pairwise coagglomeration regressions.

## App. Table 6A: Extended EG OLS Results without Natural Advantages

| Dependent variable is EG Coaggl. Index calculated with state total emp. | Pairwise Maximum Regressions | | | | | Pairwise Mean Regressions | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Base Estimation | Separate Input & Output | Exclude Pairs in Same SIC2 | Include Industry FE | Weight by Pairwise Ind. Size | Base Estimation | Separate Input & Output | Exclude Pairs in Same SIC2 | Include Industry FE | Weight by Pairwise Ind. Size |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Labor Correlation | 0.146 | 0.142 | 0.110 | 0.349 | 0.121 | 0.135 | 0.134 | 0.108 | 0.322 | 0.112 |
| | (0.014) | (0.014) | (0.012) | (0.037) | (0.014) | (0.013) | (0.013) | (0.012) | (0.034) | (0.014) |
| Input-Output | 0.149 | | 0.108 | 0.137 | 0.135 | 0.185 | | 0.117 | 0.169 | 0.163 |
| | (0.032) | | (0.024) | (0.031) | (0.034) | (0.037) | | (0.022) | (0.036) | (0.034) |
| Input | | 0.109 | | | | | 0.116 | | | |
| | | (0.026) | | | | | (0.030) | | | |
| Output | | 0.095 | | | | | 0.098 | | | |
| | | (0.035) | | | | | (0.036) | | | |
| Technology Flows | 0.112 | 0.096 | 0.050 | 0.079 | 0.095 | 0.125 | 0.121 | 0.032 | 0.093 | 0.108 |
| Scherer R&D | (0.031) | (0.030) | (0.022) | (0.026) | (0.024) | (0.035) | (0.036) | (0.025) | (0.030) | (0.030) |
| $R^2$ | 0.077 | 0.084 | 0.031 | 0.144 | 0.080 | 0.097 | 0.098 | 0.031 | 0.157 | 0.097 |
| Observations | 7381 | 7381 | 7000 | 7381 | 7381 | 7381 | 7381 | 7000 | 7381 | 7381 |

Notes: See Table 4. Regression of pairwise EG Coagglomeration Index on determinants of industrial co-location. Coagglomeration measures are calculated from the 1987 Census of Manufacturers using state total employments for each industry. Columns 3 and 8 exclude SIC3 pairwise combinations within the same SIC2. Columns 4 and 9 include consolidated industry fixed effects where indicators take unit value for either pairwise industry. App. Table 6C repeats Column 1 with alternative coagglomeration metrics. Variables are transformed to have unit standard deviation for interpretation. Robust standard errors are reported in parentheses.

**App. Table 6B: Extended EG OLS Results with Natural Advantages**

| Dependent variable is EG Coaggl. Index calculated with state total emp. | Pairwise Maximum Regressions | | | | | Pairwise Mean Regressions | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Base Estimation | Separate Input & Output | Exclude Pairs in Same SIC2 | Include Industry FE | Weight by Pairwise Ind. Size | Base Estimation | Separate Input & Output | Exclude Pairs in Same SIC2 | Include Industry FE | Weight by Pairwise Ind. Size |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Natural Advantages | 0.163 | 0.162 | 0.172 | 0.123 | 0.161 | 0.159 | 0.158 | 0.172 | 0.124 | 0.157 |
| | (0.015) | (0.015) | (0.015) | (0.014) | (0.016) | (0.015) | (0.015) | (0.015) | (0.014) | (0.016) |
| Labor Correlation | 0.118 | 0.114 | 0.085 | 0.297 | 0.091 | 0.108 | 0.107 | 0.083 | 0.269 | 0.083 |
| | (0.013) | (0.013) | (0.012) | (0.038) | (0.014) | (0.013) | (0.013) | (0.012) | (0.035) | (0.014) |
| Input-Output | 0.146 | | 0.110 | 0.141 | 0.134 | 0.180 | | 0.119 | 0.173 | 0.161 |
| | (0.032) | | (0.024) | (0.031) | (0.033) | (0.037) | | (0.022) | (0.036) | (0.034) |
| Input | | 0.106 | | | | | 0.112 | | | |
| | | (0.026) | | | | | (0.029) | | | |
| Output | | 0.093 | | | | | 0.097 | | | |
| | | (0.035) | | | | | (0.037) | | | |
| Technology Flows Scherer R&D | 0.096 | 0.079 | 0.046 | 0.072 | 0.079 | 0.110 | 0.106 | 0.029 | 0.087 | 0.093 |
| | (0.031) | (0.030) | (0.022) | (0.027) | (0.024) | (0.035) | (0.036) | (0.025) | (0.031) | (0.030) |
| $R^2$ | 0.103 | 0.110 | 0.059 | 0.157 | 0.105 | 0.121 | 0.121 | 0.060 | 0.170 | 0.121 |
| Observations | 7381 | 7381 | 7000 | 7381 | 7381 | 7381 | 7381 | 7000 | 7381 | 7381 |

Notes: See App. Table 6A.

## App. Table 6C: Extended EG OLS Results with Scherer Technology Flows

| | Dependent Variable is EG Coagglomeration Index | | | |
|---|---|---|---|---|
| | State Total Employment Coagglomeration | PMSA Total Employment Coagglomeration | County Total Employment Coagglomeration | State Firm Birth Employment Coagglomeration |
| | (1) | (2) | (3) | (4) |
| | *A. Excluding Natural Advantages* | | | |
| Labor Correlation | 0.146 | 0.078 | 0.060 | 0.060 |
| | (0.014) | (0.014) | (0.013) | (0.012) |
| Input-Output | 0.149 | 0.125 | 0.101 | 0.086 |
| | (0.032) | (0.024) | (0.019) | (0.026) |
| Technology Flows | 0.112 | 0.098 | 0.067 | 0.054 |
| Scherer R&D | (0.031) | (0.027) | (0.019) | (0.020) |
| $R^2$ | 0.077 | 0.044 | 0.025 | 0.019 |
| | *B. Including Natural Advantages* | | | |
| Natural Advantages | 0.163 | 0.159 | 0.204 | 0.100 |
| | (0.015) | (0.012) | (0.012) | (0.016) |
| Labor Correlation | 0.118 | 0.050 | 0.024 | 0.043 |
| | (0.013) | (0.014) | (0.013) | (0.012) |
| Input-Output | 0.146 | 0.122 | 0.096 | 0.085 |
| | (0.032) | (0.024) | (0.019) | (0.026) |
| Technology Flows | 0.096 | 0.081 | 0.046 | 0.044 |
| Scherer R&D | (0.031) | (0.027) | (0.020) | (0.020) |
| $R^2$ | 0.103 | 0.067 | 0.065 | 0.028 |

Notes: See App. Table 6A. Column 1 repeats the first column of App. Tables 6A and 6B with coagglomeration measured through state total employments for each industry. Columns 2-4 substitute alternative EG metrics of coagglomeration.

## App. Table 6D: Extended EG OLS Results with Patent Technology Flows

| | Dependent Variable is EG Coagglomeration Index | | | |
|---|---|---|---|---|
| | State Total Employment Coagglomeration | PMSA Total Employment Coagglomeration | County Total Employment Coagglomeration | State Firm Birth Employment Coagglomeration |
| | (1) | (2) | (3) | (4) |
| | *A. Excluding Natural Advantages* | | | |
| Labor Correlation | 0.156 | 0.077 | 0.057 | 0.058 |
| | (0.016) | (0.015) | (0.013) | (0.013) |
| Input-Output | 0.185 | 0.145 | 0.112 | 0.097 |
| | (0.037) | (0.026) | (0.019) | (0.028) |
| Technology Flows | -0.001 | 0.046 | 0.044 | 0.032 |
| Patent Citations | (0.013) | (0.015) | (0.013) | (0.013) |
| $R^2$ | 0.066 | 0.037 | 0.023 | 0.017 |
| | *B. Including Natural Advantages* | | | |
| Natural Advantages | 0.172 | 0.166 | 0.208 | 0.105 |
| | (0.016) | (0.013) | (0.012) | (0.016) |
| Labor Correlation | 0.124 | 0.046 | 0.019 | 0.039 |
| | (0.015) | (0.014) | (0.013) | (0.013) |
| Input-Output | 0.176 | 0.136 | 0.101 | 0.092 |
| | (0.036) | (0.026) | (0.018) | (0.028) |
| Technology Flows | -0.001 | 0.046 | 0.044 | 0.031 |
| Patent Citations | (0.013) | (0.015) | (0.012) | (0.013) |
| $R^2$ | 0.095 | 0.064 | 0.065 | 0.027 |

Notes: See App. Table 6C. Estimations substitute technology flows calculated from patent citations.

## App. Table 6E: Extended DO OLS Results without Natural Advantages

| Dependent variable is DO Coaggl. Index calculated with firm employments | Bilateral Firm Employments, Threshold 1000 mi. | | | | | Bilateral Firm Employments, Threshold 250 mi. | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Base Estimation | Separate Input & Output | Exclude Pairs in Same SIC2 | Include Industry FE | Weight by Pairwise Ind. Size | Base Estimation | Separate Input & Output | Exclude Pairs in Same SIC2 | Include Industry FE | Weight by Pairwise Ind. Size |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Labor Correlation | -0.173 | -0.175 | -0.199 | 0.021 | -0.190 | 0.098 | 0.096 | 0.053 | 0.285 | 0.091 |
| | (0.012) | (0.012) | (0.011) | (0.007) | (0.015) | (0.013) | (0.013) | (0.011) | (0.031) | (0.013) |
| Input-Output | 0.113 | | 0.104 | 0.011 | 0.128 | 0.150 | | 0.173 | 0.096 | 0.185 |
| | (0.023) | | (0.020) | (0.005) | (0.022) | (0.035) | | (0.032) | (0.026) | (0.036) |
| Input | | 0.097 | | | | | 0.080 | | | |
| | | (0.019) | | | | | (0.028) | | | |
| Output | | 0.048 | | | | | 0.108 | | | |
| | | (0.022) | | | | | (0.040) | | | |
| Technology Flows | 0.031 | 0.018 | 0.052 | 0.003 | 0.042 | 0.075 | 0.066 | 0.037 | 0.070 | 0.068 |
| Scherer R&D | (0.019) | (0.020) | (0.016) | (0.005) | (0.019) | (0.034) | (0.033) | (0.022) | (0.027) | (0.029) |
| $R^2$ | 0.040 | 0.042 | 0.054 | 0.899 | 0.051 | 0.051 | 0.054 | 0.039 | 0.419 | 0.085 |
| Observations | 7381 | 7381 | 7000 | 7381 | 7381 | 7381 | 7381 | 7000 | 7381 | 7381 |

Notes: See Table 4. Regression of pairwise DO Coagglomeration Index on determinants of industrial co-location. Coagglomeration measures are calculated from the 1997 Census of Manufacturers using bilateral firm employments. The column header indicates the localization threshold employed. Columns 3 and 8 exclude SIC3 pairwise combinations within the same SIC2. Columns 4 and 9 include consolidated industry fixed effects where indicators take unit value for either pairwise industry. App. Table 6G repeats Column 1 with alternative coagglomeration metrics. Maximum values for the pairwise combination are employed. Variables are transformed to have unit standard deviation for interpretation. Robust standard errors are reported in parentheses.

## App. Table 6F: Extended DO OLS Results with Natural Advantages

| Dependent variable is DO Coaggl. Index calculated with firm employments | Bilateral Firm Employments, Threshold 1000 mi. | | | | | Bilateral Firm Employments, Threshold 250 mi. | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Base Estimation | Separate Input & Output | Exclude Pairs in Same SIC2 | Include Industry FE | Weight by Pairwise Ind. Size | Base Estimation | Separate Input & Output | Exclude Pairs in Same SIC2 | Include Industry FE | Weight by Pairwise Ind. Size |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Natural Advantages | 0.437 | 0.437 | 0.433 | 0.408 | 0.567 | 0.251 | 0.252 | 0.253 | 0.428 | 0.156 |
| | (0.013) | (0.013) | (0.014) | (0.033) | (0.015) | (0.012) | (0.012) | (0.013) | (0.027) | (0.013) |
| Labor Correlation | -0.172 | -0.174 | -0.196 | 0.012 | -0.189 | 0.069 | 0.066 | 0.029 | 0.233 | 0.065 |
| | (0.011) | (0.011) | (0.010) | (0.007) | (0.013) | (0.012) | (0.012) | (0.011) | (0.030) | (0.013) |
| Input-Output | 0.084 | | 0.058 | 0.014 | 0.078 | 0.162 | | 0.177 | 0.096 | 0.190 |
| | (0.019) | | (0.015) | (0.005) | (0.016) | (0.035) | | (0.034) | (0.026) | (0.037) |
| Input | | 0.086 | | | | | 0.097 | | | |
| | | (0.018) | | | | | (0.029) | | | |
| Output | | 0.024 | | | | | 0.107 | | | |
| | | (0.017) | | | | | (0.038) | | | |
| Technology Flows Scherer R&D | 0.027 | 0.015 | 0.053 | 0.002 | 0.036 | 0.076 | 0.065 | 0.033 | 0.057 | 0.068 |
| | (0.014) | (0.015) | (0.012) | (0.005) | (0.012) | (0.033) | (0.032) | (0.022) | (0.026) | (0.029) |
| $R^2$ | 0.230 | 0.232 | 0.239 | 0.904 | 0.291 | 0.113 | 0.117 | 0.102 | 0.448 | 0.108 |
| Observations | 7381 | 7381 | 7000 | 7381 | 7381 | 7381 | 7381 | 7000 | 7381 | 7381 |

Notes: See App. Table 6E.

## App. Table 6G: Extended DO OLS Results with Scherer Technology Flows

| | DO Coagglomeration Index, Threshold 1000 mi. | | | | DO Coagglomeration Index, Threshold 250 mi. | | | |
|---|---|---|---|---|---|---|---|---|
| | Bilateral Firm Employments | Bilateral Firm Counts | County-Industry Employments | County-Industry Counts | Bilateral Firm Employments | Bilateral Firm Counts | County-Industry Employments | County-Industry Counts |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | | | | *A. Excluding Natural Advantages* | | | | |
| Labor Correlation | -0.173 | -0.181 | -0.075 | -0.177 | 0.098 | 0.032 | 0.130 | 0.031 |
| | (0.012) | (0.012) | (0.012) | (0.012) | (0.013) | (0.013) | (0.014) | (0.013) |
| Input-Output | 0.113 | 0.116 | 0.100 | 0.116 | 0.150 | 0.126 | 0.161 | 0.128 |
| | (0.023) | (0.022) | (0.022) | (0.022) | (0.035) | (0.031) | (0.036) | (0.031) |
| Technology Flows Scherer R&D | 0.031 | 0.038 | 0.010 | 0.038 | 0.075 | 0.057 | 0.081 | 0.058 |
| | (0.019) | (0.018) | (0.019) | (0.018) | (0.034) | (0.027) | (0.036) | (0.027) |
| $R^2$ | 0.040 | 0.043 | 0.014 | 0.042 | 0.051 | 0.026 | 0.066 | 0.027 |
| | | | | *B. Including Natural Advantages* | | | | |
| Natural Advantages | 0.437 | 0.351 | 0.402 | 0.353 | 0.251 | 0.208 | 0.217 | 0.201 |
| | (0.013) | (0.014) | (0.014) | (0.014) | (0.012) | (0.014) | (0.011) | (0.014) |
| Labor Correlation | -0.172 | -0.180 | -0.074 | -0.177 | 0.069 | 0.008 | 0.105 | 0.008 |
| | (0.011) | (0.011) | (0.011) | (0.011) | (0.012) | (0.012) | (0.013) | (0.012) |
| Input-Output | 0.084 | 0.093 | 0.073 | 0.092 | 0.162 | 0.136 | 0.171 | 0.137 |
| | (0.019) | (0.019) | (0.019) | (0.019) | (0.035) | (0.031) | (0.035) | (0.031) |
| Technology Flows Scherer R&D | 0.027 | 0.036 | 0.007 | 0.036 | 0.076 | 0.058 | 0.082 | 0.059 |
| | (0.014) | (0.014) | (0.014) | (0.014) | (0.033) | (0.027) | (0.035) | (0.026) |
| $R^2$ | 0.230 | 0.166 | 0.175 | 0.166 | 0.113 | 0.069 | 0.113 | 0.067 |

Notes: See App. Table 6E. Column 1 repeats the first column of App. Tables 6E and 6F with coagglomeration measured through firm bilateral employments and 1000 mi. threshold. Column 5 repeats the sixth column of App. Tables 6E and 6F with the 250 mi. threshold. The remaining columns substitute alternative DO metrics.

**App. Table 6H: Extended DO OLS Results with Patent Technology Flows**

| | DO Coagglomeration Index, Threshold 1000 mi. | | | | DO Coagglomeration Index, Threshold 250 mi. | | | |
|---|---|---|---|---|---|---|---|---|
| | Bilateral Firm Employments | Bilateral Firm Counts | County-Industry Employments | County-Industry Counts | Bilateral Firm Employments | Bilateral Firm Counts | County-Industry Employments | County-Industry Counts |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | | | | *A. Excluding Natural Advantages* | | | | |
| Labor Correlation | -0.172 | -0.174 | -0.069 | -0.171 | 0.097 | 0.032 | 0.135 | 0.031 |
| | (0.012) | (0.012) | (0.013) | (0.012) | (0.014) | (0.014) | (0.016) | (0.014) |
| Input-Output | 0.121 | 0.132 | 0.108 | 0.131 | 0.165 | 0.139 | 0.185 | 0.140 |
| | (0.023) | (0.022) | (0.023) | (0.022) | (0.037) | (0.033) | (0.038) | (0.033) |
| Technology Flows | 0.008 | -0.020 | -0.024 | -0.016 | 0.039 | 0.025 | 0.008 | 0.027 |
| Patent Citations | (0.013) | (0.012) | (0.013) | (0.012) | (0.015) | (0.014) | (0.015) | (0.014) |
| $R^2$ | 0.039 | 0.042 | 0.015 | 0.041 | 0.047 | 0.024 | 0.060 | 0.025 |
| | | | | *B. Including Natural Advantages* | | | | |
| Natural Advantages | 0.438 | 0.352 | 0.404 | 0.355 | 0.249 | 0.208 | 0.219 | 0.201 |
| | (0.013) | (0.014) | (0.014) | (0.014) | (0.013) | (0.014) | (0.012) | (0.014) |
| Labor Correlation | -0.167 | -0.170 | -0.065 | -0.167 | 0.074 | 0.013 | 0.115 | 0.012 |
| | (0.011) | (0.011) | (0.012) | (0.011) | (0.014) | (0.013) | (0.015) | (0.013) |
| Input-Output | 0.095 | 0.112 | 0.085 | 0.111 | 0.184 | 0.154 | 0.201 | 0.155 |
| | (0.019) | (0.020) | (0.020) | (0.020) | (0.037) | (0.033) | (0.039) | (0.033) |
| Technology Flows | -0.013 | -0.036 | -0.043 | -0.032 | 0.008 | -0.001 | -0.019 | 0.002 |
| Patent Citations | (0.011) | (0.011) | (0.016) | (0.011) | (0.015) | (0.014) | (0.016) | (0.014) |
| $R^2$ | 0.230 | 0.166 | 0.176 | 0.166 | 0.110 | 0.066 | 0.107 | 0.064 |

Notes: See App. Table 6G. Estimations substitute technology flows calculated from patent citations.

## App. Table 7A: First-Stage Specifications for UK Instruments

| Dependent variable is the explanatory regressor listed in the column header | Univariate First-Stage Specifications | | | Multivariate First-Stages without Technology | | Multivariate First-Stages with Technology | | |
|---|---|---|---|---|---|---|---|---|
| | Labor Correlation | Input-Output | Technology Scherer | Labor Correlation | Input-Output | Labor Correlation | Input-Output | Technology Scherer |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | | | | *A. Excluding Natural Advantages* | | | | |
| UK Labor IV | 0.278 | | | **0.262** | 0.095 | **0.247** | 0.041 | 0.048 |
| | (0.012) | | | **(0.013)** | (0.017) | **(0.013)** | (0.024) | (0.017) |
| UK Input-Output IV | | 0.345 | | 0.070 | **0.323** | 0.064 | **0.300** | 0.159 |
| | | (0.041) | | (0.011) | **(0.042)** | (0.011) | **(0.043)** | (0.051) |
| UK Technology Flows Flows IV | | | 0.237 | | | 0.054 | 0.202 | **0.195** |
| | | | (0.034) | | | (0.012) | (0.089) | **(0.029)** |
| $R^2$, Shea's Partial $R^2$ | 0.077 | 0.119 | 0.056 | 0.045 | 0.070 | 0.040 | 0.010 | 0.004 |
| Minimum Eigenvalue | 587.5 | 946.9 | 415.2 | 125.5 | | | 9.0 | |
| F(1,6998) | 527.5 | 69.5 | 47.6 | | | | | |
| | | | | *B. Including Natural Advantages* | | | | |
| Natural Advantages | 0.131 | 0.020 | 0.014 | 0.134 | 0.013 | 0.135 | 0.015 | 0.018 |
| | (0.010) | (0.008) | (0.010) | (0.010) | (0.008) | (0.010) | (0.008) | (0.011) |
| UK Labor IV | 0.271 | | | **0.252** | 0.094 | **0.238** | 0.040 | 0.047 |
| | (0.012) | | | **(0.013)** | (0.017) | **(0.013)** | (0.029) | (0.017) |
| UK Input-Output IV | | 0.346 | | 0.077 | **0.323** | 0.071 | **0.301** | 0.159 |
| | | (0.041) | | (0.011) | **(0.043)** | (0.012) | **(0.043)** | (0.051) |
| UK Technology Flows Flows IV | | | 0.237 | | | 0.055 | 0.202 | **0.195** |
| | | | (0.034) | | | (0.012) | (0.089) | **(0.029)** |
| $R^2$, Shea's Partial $R^2$ | 0.094 | 0.119 | 0.056 | 0.042 | 0.066 | 0.037 | 0.010 | 0.004 |
| Minimum Eigenvalue | 565.7 | 949.9 | 415.0 | 115.5 | | | 9.1 | |
| F(1,6997) | 504.6 | 69.7 | 47.6 | | | | | |

Notes: First-stage regressions of US pairwise determinants of industrial co-location on similarly constructed UK instruments. All pairwise combinations of manufacturing SIC3 industries are included, except those listed in the text, for 7000 observations. The decline in observations from primary OLS specifications is due to the exclusion of pairwise combinations within the same SIC2. Variable constructions are described in the text. Maximum values for the pairwise combination are employed. Variables are transformed to have unit standard deviation for interpretation. Regressions are unweighted. Robust standard errors are reported in parentheses. Shea's partial R2 is reported for specifications with multiple instruments. The p-values for all F statistics are 0.0000. The 2SLS sizes of the nominal 5% Wald test are 16.38 and 7.03 for single and dual IVs, respectively, at 10%.

**App. Table 7B: Extended EG UK IV Results without Technology Flows**

| | Dependent Variable is EG Coagglomeration Index | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | State Total Empl. Coaggl. OLS | State Total Empl. Coaggl. IV | PMSA Total Empl. Coaggl. OLS | PMSA Total Empl. Coaggl. IV | County Total Empl. Coaggl. OLS | County Total Empl. Coaggl. IV | State Birth Empl. Coaggl. OLS | State Birth Empl. Coaggl. IV |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | | | | *A. Excluding Natural Advantages* | | | | |
| Labor Correlation | 0.108 | 0.140 | 0.033 | 0.151 | 0.029 | 0.047 | 0.042 | 0.187 |
| | (0.012) | (0.057) | (0.013) | (0.063) | (0.013) | (0.061) | (0.012) | (0.059) |
| Input-Output | 0.121 | 0.149 | 0.096 | 0.078 | 0.075 | 0.103 | 0.051 | 0.152 |
| | (0.025) | (0.047) | (0.021) | (0.045) | (0.017) | (0.047) | (0.015) | (0.064) |
| | | | | *B. Including Natural Advantages* | | | | |
| Natural Advantages | 0.173 | 0.173 | 0.162 | 0.149 | 0.221 | 0.225 | 0.103 | 0.083 |
| | (0.015) | (0.018) | (0.011) | (0.014) | (0.013) | (0.017) | (0.017) | (0.019) |
| Labor Correlation | 0.083 | 0.079 | 0.009 | 0.099 | -0.003 | -0.032 | 0.027 | 0.161 |
| | (0.012) | (0.059) | (0.013) | (0.065) | (0.013) | (0.063) | (0.012) | (0.062) |
| Input-Output | 0.122 | 0.191 | 0.096 | 0.114 | 0.076 | 0.157 | 0.051 | 0.168 |
| | (0.025) | (0.049) | (0.020) | (0.046) | (0.017) | (0.050) | (0.015) | (0.066) |

Notes: OLS and IV Regression of pairwise EG Coagglomeration Index on determinants of industrial co-location. Instruments developed through UK data. App. Table 7A documents the first-stage coefficients. Variables are transformed to have unit standard deviation for interpretation. Robust standard errors are reported in parentheses.

## App. Table 7C: Extended EG UK IV Results with Technology Flows

| | Dependent Variable is EG Coagglomeration Index | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | State Total Empl. Coaggl. OLS | State Total Empl. Coaggl. IV | PMSA Total Empl. Coaggl. OLS | PMSA Total Empl. Coaggl. IV | County Total Empl. Coaggl. OLS | County Total Empl. Coaggl. IV | State Birth Empl. Coaggl. OLS | State Birth Empl. Coaggl. IV |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | | | | *A. Excluding Natural Advantages* | | | | |
| Labor Correlation | 0.110 | 0.120 | 0.035 | 0.136 | 0.030 | 0.028 | 0.042 | 0.254 |
| | (0.012) | (0.061) | (0.013) | (0.061) | (0.013) | (0.062) | (0.012) | (0.100) |
| Input-Output | 0.108 | 0.095 | 0.085 | 0.039 | 0.068 | 0.051 | 0.047 | 0.341 |
| | (0.024) | (0.118) | (0.020) | (0.108) | (0.017) | (0.106) | (0.016) | (0.305) |
| Technology Flows Scherer R&D | 0.050 | 0.104 | 0.041 | 0.076 | 0.026 | 0.099 | 0.015 | -0.359 |
| | (0.022) | (0.180) | (0.018) | (0.151) | (0.012) | (0.146) | (0.017) | (0.476) |
| | | | | *B. Including Natural Advantages* | | | | |
| Natural Advantages | 0.172 | 0.174 | 0.161 | 0.149 | 0.221 | 0.225 | 0.103 | 0.078 |
| | (0.015) | (0.018) | (0.011) | (0.014) | (0.013) | (0.017) | (0.017) | (0.021) |
| Labor Correlation | 0.085 | 0.068 | 0.011 | 0.092 | -0.002 | -0.039 | 0.028 | 0.232 |
| | (0.012) | (0.067) | (0.013) | (0.064) | (0.013) | (0.069) | (0.012) | (0.107) |
| Input-Output | 0.110 | 0.162 | 0.086 | 0.097 | 0.070 | 0.139 | 0.048 | 0.365 |
| | (0.024) | (0.135) | (0.020) | (0.113) | (0.017) | (0.129) | (0.016) | (0.320) |
| Technology Flows Scherer R&D | 0.046 | 0.055 | 0.037 | 0.033 | 0.022 | 0.035 | 0.012 | -0.376 |
| | (0.022) | (0.216) | (0.017) | (0.165) | (0.012) | (0.201) | (0.017) | (0.503) |

Notes: See App. Table 7B.

**App. Table 7D: Extended DO UK IV Results without Technology Flows**

| | Dependent Variable is DO Coagglomeration Index Calculated with Bilateral Firm Employments | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1000 mi. Distance Threshold OLS | 1000 mi. Distance Threshold IV | 500 mi. Distance Threshold OLS | 500 mi. Distance Threshold IV | 250 mi. Distance Threshold OLS | 250 mi. Distance Threshold IV | 100 mi. Distance Threshold OLS | 100 mi. Distance Threshold IV |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | | | | *A. Excluding Natural Advantages* | | | | |
| Labor Correlation | -0.202 | 0.171 | -0.055 | 0.667 | 0.051 | 0.549 | 0.038 | 0.190 |
| | (0.011) | (0.059) | (0.011) | (0.074) | (0.011) | (0.071) | (0.011) | (0.058) |
| Input-Output | 0.118 | 0.104 | 0.172 | 0.058 | 0.183 | 0.127 | 0.073 | 0.038 |
| | (0.022) | (0.046) | (0.031) | (0.060) | (0.032) | (0.059) | (0.017) | (0.039) |
| | | | | *B. Including Natural Advantages* | | | | |
| Natural Advantages | 0.432 | 0.431 | 0.410 | 0.362 | 0.254 | 0.210 | 0.564 | 0.561 |
| | (0.014) | (0.014) | (0.012) | (0.013) | (0.012) | (0.016) | (0.021) | (0.021) |
| Labor Correlation | -0.199 | -0.014 | -0.097 | 0.377 | 0.027 | 0.501 | 0.024 | 0.201 |
| | (0.010) | (0.049) | (0.010) | (0.061) | (0.011) | (0.071) | (0.010) | (0.051) |
| Input-Output | 0.072 | 0.080 | 0.136 | 0.109 | 0.186 | 0.164 | 0.092 | 0.125 |
| | (0.017) | (0.040) | (0.027) | (0.051) | (0.033) | (0.059) | (0.020) | (0.038) |

Notes: OLS and IV Regression of pairwise DO Coagglomeration Index on determinants of industrial co-location. Instruments developed through UK data. App. Table 7A documents the first-stage coefficients. Variables are transformed to have unit standard deviation for interpretation. Robust standard errors are reported in parentheses.

## App. Table 8A: First-Stage Specifications for US Spatial Instruments

| Dependent variable is the explanatory regressor listed in the column header | Univariate First-Stage Specifications | | | Multivariate First-Stages without Technology | | Multivariate First-Stages with Technology | | |
|---|---|---|---|---|---|---|---|---|
| | Labor Correlation | Input-Output | Technology Scherer | Labor Correlation | Input-Output | Labor Correlation | Input-Output | Technology Scherer |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | | | | *A. Excluding Natural Advantages* | | | | |
| US Spatial Labor IV | 0.501 | | | **0.499** | 0.045 | **0.490** | 0.019 | -0.022 |
| | (0.010) | | | **(0.010)** | (0.010) | **(0.010)** | (0.009) | (0.012) |
| US Spatial Input-Output IV | | 0.525 | | 0.020 | **0.522** | 0.013 | **0.502** | 0.210 |
| | | (0.049) | | (0.012) | **(0.049)** | (0.012) | **(0.049)** | (0.059) |
| US Spatial Technology Flows IV | | | 0.237 | | | 0.073 | 0.212 | **0.218** |
| | | | (0.034) | | | (0.011) | (0.080) | **(0.029)** |
| $R^2$, Shea's Partial $R^2$ | 0.251 | 0.276 | 0.056 | 0.243 | 0.269 | 0.243 | 0.046 | 0.014 |
| Minimum Eigenvalue | 2341.3 | 2668.0 | 415.2 | 1030.6 | | | 31.7 | |
| F(1,6998) | 2580.0 | 116.8 | 47.6 | | | | | |
| | | | | *B. Including Natural Advantages* | | | | |
| Natural Advantages | 0.087 | 0.013 | 0.014 | 0.087 | 0.008 | 0.088 | 0.010 | 0.020 |
| | (0.009) | (0.007) | (0.010) | (0.009) | (0.007) | (0.009) | (0.007) | (0.011) |
| US Spatial Labor IV | 0.490 | | | **0.489** | 0.044 | **0.479** | 0.018 | -0.025 |
| | (0.010) | | | **(0.010)** | (0.010) | **(0.010)** | (0.009) | (0.012) |
| US Spatial Input-Output IV | | 0.526 | | 0.021 | **0.522** | 0.014 | **0.502** | 0.211 |
| | | (0.049) | | (0.012) | **(0.048)** | (0.012) | **(0.049)** | (0.059) |
| US Spatial Technology Flows IV | | | 0.237 | | | 0.074 | 0.212 | **0.218** |
| | | | (0.034) | | | (0.011) | (0.080) | **(0.029)** |
| $R^2$, Shea's Partial $R^2$ | 0.258 | 0.276 | 0.056 | 0.234 | 0.269 | 0.234 | 0.046 | 0.014 |
| Minimum Eigenvalue | 2234.7 | 2669.5 | 415.0 | 991.3 | | | 31.8 | |
| F(1,6997) | 2434.4 | 117.4 | 47.6 | | | | | |

Notes: See App. Table 7A. Estimations employ US spatial instruments.

**App. Table 8B: Extended EG US-Spatial IV Results without Technology Flows**

| | Dependent Variable is EG Coagglomeration Index | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | State Total Empl. Coaggl. OLS | State Total Empl. Coaggl. IV | PMSA Total Empl. Coaggl. OLS | PMSA Total Empl. Coaggl. IV | County Total Empl. Coaggl. OLS | County Total Empl. Coaggl. IV | State Birth Empl. Coaggl. OLS | State Birth Empl. Coaggl. IV |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | | | | *A. Excluding Natural Advantages* | | | | |
| Labor Correlation | 0.108 | 0.133 | 0.033 | 0.111 | 0.029 | 0.068 | 0.042 | 0.067 |
| | (0.012) | (0.024) | (0.013) | (0.024) | (0.013) | (0.024) | (0.012) | (0.027) |
| Input-Output | 0.121 | 0.177 | 0.096 | 0.129 | 0.075 | 0.115 | 0.051 | 0.111 |
| | (0.025) | (0.042) | (0.021) | (0.035) | (0.017) | (0.031) | (0.015) | (0.053) |
| | | | | *B. Including Natural Advantages* | | | | |
| Natural Advantages | 0.173 | 0.171 | 0.162 | 0.152 | 0.221 | 0.218 | 0.103 | 0.101 |
| | (0.015) | (0.016) | (0.011) | (0.012) | (0.013) | (0.013) | (0.017) | (0.017) |
| Labor Correlation | 0.083 | 0.091 | 0.009 | 0.073 | -0.003 | 0.014 | 0.027 | 0.043 |
| | (0.012) | (0.025) | (0.013) | (0.024) | (0.013) | (0.024) | (0.012) | (0.027) |
| Input-Output | 0.122 | 0.185 | 0.096 | 0.136 | 0.076 | 0.125 | 0.051 | 0.117 |
| | (0.025) | (0.041) | (0.020) | (0.034) | (0.017) | (0.029) | (0.015) | (0.054) |

Notes: OLS and IV Regression of pairwise EG Coagglomeration Index on determinants of industrial co-location. Instruments developed through US spatial data. App. Table 8A documents the first-stage coefficients. Variables are transformed to have unit standard deviation for interpretation. Robust standard errors are reported in parentheses.

## App. Table 8C: Extended EG US-Spatial IV Results with Technology Flows

| | Dependent Variable is EG Coagglomeration Index | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | State Total Empl. Coaggl. OLS | State Total Empl. Coaggl. IV | PMSA Total Empl. Coaggl. OLS | PMSA Total Empl. Coaggl. IV | County Total Empl. Coaggl. OLS | County Total Empl. Coaggl. IV | State Birth Empl. Coaggl. OLS | State Birth Empl. Coaggl. IV |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | | | | *A. Excluding Natural Advantages* | | | | |
| Labor Correlation | 0.110 | 0.134 | 0.035 | 0.110 | 0.030 | 0.068 | 0.042 | 0.068 |
| | (0.012) | (0.025) | (0.013) | (0.025) | (0.013) | (0.024) | (0.012) | (0.028) |
| Input-Output | 0.108 | 0.174 | 0.085 | 0.138 | 0.068 | 0.111 | 0.047 | 0.101 |
| | (0.024) | (0.092) | (0.020) | (0.078) | (0.017) | (0.065) | (0.016) | (0.103) |
| Technology Flows Scherer R&D | 0.050 | 0.007 | 0.041 | -0.021 | 0.026 | 0.009 | 0.015 | 0.022 |
| | (0.022) | (0.144) | (0.018) | (0.123) | (0.012) | (0.107) | (0.017) | (0.132) |
| | | | | *B. Including Natural Advantages* | | | | |
| Natural Advantages | 0.172 | 0.170 | 0.161 | 0.152 | 0.221 | 0.218 | 0.103 | 0.100 |
| | (0.015) | (0.016) | (0.011) | (0.012) | (0.013) | (0.013) | (0.017) | (0.017) |
| Labor Correlation | 0.085 | 0.092 | 0.011 | 0.073 | -0.002 | 0.016 | 0.028 | 0.045 |
| | (0.012) | (0.026) | (0.013) | (0.025) | (0.013) | (0.024) | (0.012) | (0.029) |
| Input-Output | 0.110 | 0.172 | 0.086 | 0.136 | 0.070 | 0.108 | 0.048 | 0.104 |
| | (0.024) | (0.088) | (0.020) | (0.074) | (0.017) | (0.060) | (0.016) | (0.105) |
| Technology Flows Scherer R&D | 0.046 | 0.031 | 0.037 | 0.000 | 0.022 | 0.039 | 0.012 | 0.029 |
| | (0.022) | (0.140) | (0.017) | (0.119) | (0.012) | (0.102) | (0.017) | (0.134) |

Notes: See App. Table 8B.

## App. Table 8D: Extended DO US-Spatial IV Results without Technology Flows

| | Dependent Variable is DO Coagglomeration Index Calculated with Bilateral Firm Employments | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1000 mi. Distance Threshold OLS | 1000 mi. Distance Threshold IV | 500 mi. Distance Threshold OLS | 500 mi. Distance Threshold IV | 250 mi. Distance Threshold OLS | 250 mi. Distance Threshold IV | 100 mi. Distance Threshold OLS | 100 mi. Distance Threshold IV |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | | | | *A. Excluding Natural Advantages* | | | | |
| Labor Correlation | -0.202 | -0.072 | -0.055 | 0.095 | 0.051 | 0.198 | 0.038 | 0.053 |
| | (0.011) | (0.023) | (0.011) | (0.024) | (0.011) | (0.025) | (0.012) | (0.024) |
| Input-Output | 0.118 | 0.166 | 0.172 | 0.230 | 0.183 | 0.203 | 0.073 | 0.095 |
| | (0.022) | (0.029) | (0.031) | (0.049) | (0.032) | (0.049) | (0.017) | (0.033) |
| | | | | *B. Including Natural Advantages* | | | | |
| Natural Advantages | 0.432 | 0.427 | 0.410 | 0.390 | 0.254 | 0.233 | 0.564 | 0.562 |
| | (0.014) | (0.014) | (0.012) | (0.012) | (0.012) | (0.013) | (0.021) | (0.021) |
| Labor Correlation | -0.199 | -0.229 | -0.097 | 0.030 | 0.027 | 0.248 | 0.024 | 0.209 |
| | (0.010) | (0.020) | (0.010) | (0.022) | (0.011) | (0.024) | (0.010) | (0.022) |
| Input-Output | 0.072 | 0.124 | 0.136 | 0.204 | 0.186 | 0.213 | 0.092 | 0.141 |
| | (0.017) | (0.026) | (0.027) | (0.045) | (0.033) | (0.050) | (0.020) | (0.033) |

Notes: OLS and IV Regression of pairwise DO Coagglomeration Index on determinants of industrial co-location. Instruments developed through US spatial data. App. Table 8A documents the first-stage coefficients. Variables are transformed to have unit standard deviation for interpretation. Robust standard errors are reported in parentheses.