# The heterogeneous determinants of coagglomeration: A differenced perspective*

Kristian Behrens† Rachel Guillain‡

April 27, 2025
—Final version—

**Abstract**

The colocation patterns of industry pairs and the locations of their business functions—such as management, R&D, or production—have been considered mostly separately. We show that considering them jointly allows to better identify the determinants of agglomeration by exploiting heterogeneity across both industry pairs and business functions. While the coagglomeration of production is likely more sensitive to buyer-supplier links and labor market pooling, the coagglomeration of management and R&D is likely more sensitive to shared knowledge. We empirically document that industry pairs that share more knowledge colocate more their knowledge-intensive business functions, whereas industry pairs with stronger buyer-supplier links and a more similar workforce colocate more their production business functions. High transport costs lead to less coagglomeration, and input-output links and knowledge spillovers are more important when there are fewer multiunit firms.

**Keywords:** coagglomeration; business functions; Marshallian agglomeration forces; transport costs; microgeographic data.

**JEL Classification:** R12; L60.

*This paper is a substantially modified and revised version of the CEPR Discussion Paper #11884. It has been prepared for the special issue of the *Japanese Economic Review* honoring Masahisa Fujita. We thank the guest editors, Gilles Duranton and Tomoya Mori, as well as two anonymous reviewers for their comments. We further thank Ferdinando Monte, Nate Baum-Snow, Théophile Bougna, Mark Brown, Don Davis, Amit Khandelwal, Bill Kerr, Fabian Lange, Julien Martin, Richard Shearmur, Will Strange, and seminar and conference participants in numerous places for comments and suggestions. We are grateful to Richard Shearmur and Mario Polèse for sharing the special census tabulations from Statistics Canada; to Bill Kerr for sharing the patent citations data; and to Mathieu Steijn, Hans Koster, and Frank Van Oort for sharing the technological relatedness data. Behrens acknowledges financial support from the CRC Program of the Social Sciences and Humanities Research Council of Canada for the funding of the *Canada Research Chair in Regional Impacts of Globalization*.

†*Corresponding author:* Department of Economics, Université du Québec à Montréal (ESG-UQAM), Canada; and CEPR, UK. E-mail: behrens.kristian@uqam.ca

‡Université de Bourgogne, France. E-mail: rachel.guillain@u-bourgogne.fr.

*"No city is really a one-industry town, not even Hollywood or the Silicon Valley. Neither is any city simply a share of the diverse national population. New York's diversity is different from that of Los Angeles." (Helsley and Strange, 2014, p.1064)*

*"While specialisation continues to be an important feature of the urban system [...] cities are increasingly distinguished by their functional specialisation (i.e., in management and services versus production) rather than by their sectoral specialisation (i.e., in one particular sector of activity versus another one)." (Duranton and Puga, 2005, p.344)*

# 1 Introduction

The most salient feature of the economic landscape is how much locations differ in their size and density of economic activity. An almost equally salient feature is how much their horizontal composition varies in terms of industry mix. Because some industries are ubiquitous, whereas others are geographically concentrated in a few places, some areas are specialized in a narrow set of industries while others have a broad economic tissue. A third salient feature is that—conditional on their industry mix—locations differ in their vertical composition: different locations play different roles in the value chain. Following secular changes in technology and the internal organization of firms (e.g., Fujita and Gokan, 2005), business functions—production, management, or R&D—have become geographically more separated. Thus, locations differ both horizontally in their industry mix (e.g., Helsley and Strange, 2014) and vertically in the business functions they perform (e.g., Duranton and Puga, 2005).

There is no shortage of theories explaining why economic activity agglomerates, why different industries tend to colocate, and why locations tend to specialize in different functions along the value chain.[1] Broadly speaking, geographic concentration patterns and functional specialization are the outcome of firms' efforts to minimize the costs of moving goods, people, and ideas. As these three costs are jointly relevant to firms' location choices, identifying empirically which matter more for the agglomeration of industries is challenging (Combes and Gobillon, 2015). In an important contribution, Ellison et al. (2010) show how coagglomeration patterns can be leveraged to better identify the drivers of agglomeration. The underlying idea is that the coagglomeration of industry pairs—as compared to the agglomeration of individual industries—allows to exploit industry-pair specific variation in the importance of the underlying mechanisms. For example, industry $i$ can have strong input-output relationships with industry $j$ but use a fairly different set of knowledge; whereas it can use a similar set of knowledge than industry $k$ but not have strong input-output relationships with it. These

[1] See, e.g., Fujita et al. (2001), Fujita and Thisse (2002), Duranton and Puga (2004), Abdel-Rahman and Anas (2004), Duranton and Puga (2005), Helsley and Strange (2014), Behrens and Robert-Nicoud (2015), and Davis and Dingel, 2020).

industry-pair specific variations can be exploited to better understand whether and how much input-output relationships, knowledge sharing, and various labor market interactions drive the geographic colocation of industries.

In this paper, we make two contributions to the literature. First, we replicate and extend existing findings on coagglomeration using detailed microgeographic manufacturing data for Canada. Leveraging a unique feature of our dataset—namely that plants are observed to operate in multiple industries—we show that coagglomeration can occur within the boundaries of establishments. We document that plants in industry $i$ are more likely to report secondary activities in industry $j$ if the industries have stronger input-output, labor market, or knowledge links. This 'within-plant coagglomeration' does not reflect externalities between plants and does not show up in traditional coagglomeration measures. To our knowledge, its implications for estimating the determinants of agglomeration have not been investigated until now. Leveraging a second unique feature of our dataset—detailed transport costs estimated from microdata—we further show that industry pairs with higher transport cost are systematically less coagglomerated, and that transport costs matter only for input-output relationships. This result corroborates key findings of the new economic geography (Fujita et al., 2001). Finally, we investigate the role of multiplant firms and show that the agglomeration mechanisms operate more strongly for industries that have fewer multiunit firms. Labor market pooling and knowledge sharing are less important for coagglomeration patterns of industry pairs with many multiunit firms, thus suggesting that the latter are less dependent on (external) agglomeration economies. This result echoes previous findings on the role of industrial organization for agglomeration (e.g., Rosenthal and Strange, 2010) and links to the corporate finance literature on internal markets in large firms (e.g., Lamont, 1997; Maksimovic and Phillips, 2002).

Second, we push further the idea of using heterogeneity in observed coagglomeration patterns to better identify the sources of agglomeration economies. While previous studies have estimated the coagglomeration of total employment (e.g., Ellison et al., 2010; Faggio et al., 2017), we leverage industry-occupation data to highlight heterogeneity in the coagglomeration of business functions. The underlying idea is that business functions benefit to varying degrees from the different sources of agglomeration economies. While production is, for example, sensitive to the presence of buyer-supplier links, these are relatively less important for the location of management or R&D which are more sensitive to shared knowledge. As knowledge dissipates quickly with distance, we would expect that coagglomeration of management and R&D occurs at shorter distances than coagglomeration of production. This is what we find in the data. We further show that the average effects we estimate for total employment mask substantial heterogeneity. Estimating a model where the effects of input-output links, labor market pooling, and knowledge spillovers can vary between management and R&D and production reveals that input-output links and labor market pooling matter more for production, whereas knowledge sharing matters more for management and R&D. Put differently, as ex-

pected knowledge sharing matters relatively more for the coagglomeration of relatively more knowledge-intensive business functions.

Our paper is related to the literatures on coagglomeration and functional specialization.[2] The location patterns of industries *and* functions have been considered mostly separately until now. While there is a growing literature on the dynamics of agglomeration, heterogeneity across industries and plants, and changes in the determinants of (co)agglomeration (e.g., Rigby and Brown, 2015; Faggio et al., 2017; Diodato et al., 2018; and Steijn et al., 2022), the empirical literature at the intersection of agglomeration and functional specialization is thin. Audretsch and Feldman (1996) document that R&D activities in US industries are geographically concentrated but mostly orthogonal to the concentration of production. Faggio et al. (2017, 2020) dissect coagglomeration patterns along dimensions that reflect organizational differences across industries. Their analysis does, however, not directly speak to the issue of functional specialization across space. Closer to our analysis, Bade et al. (2015) document the coagglomeration patterns of broad functions using German data. Their analysis is, however, geographically more aggregated and purely descriptive. Gabe and Abel (2012, 2016) investigate the coagglomeration of occupations that share the same knowledge base. Last, in the strategic management literature Alcácer (2006) analyzes firm-level colocation patterns of subsidiaries by separating R&D, production, and sales. Though very interesting, his analysis focuses however on one particular industry—cellular handsets—only.

The remainder of the paper is organized as follows. Section 2 explains how we measure coagglomeration, describes our data, and shows some descriptive statistics about coagglomeration patterns. Section 3 explains our flow- and similarity-based proxies for the Marshallian determinants of coagglomeration. Section 4 provides evidence for the determinants of coagglomeration using total employment of industries, discusses several new controls, and pays specific attention to the coagglomeration within plants and the role of transport costs. Section 5 deals with functional specialization and looks at heterogeneity in coagglomeration patterns in terms of the colocation of broad types of occupations. It also provides evidence for the differences between industry pairs with few or many multi-plant firms. Last, Section 6 concludes. We relegate technical developments and additional results to the appendix.

[2]For empirical studies on the coagglomeration of industries see, among others, Duranton and Overman (2005, 2008), Ellison et al. (2010), Howard et al. (2016), Faggio et al. (2017, 2020), Diodata et al. (2018), and Steijn et al. (2022). While these studies all deal with manufacturing, Kolko (2010) is one of the rare studies that looks at the coagglomeration of service industries in the US. The empirical literature on functional specialization is thinner (see, e.g., Duranton and Puga, 2001, 2005; Davis and Dingel, 2020).

# 2 Coagglomeration patterns

We briefly explain how we measure coagglomeration, what data we use, and report some descriptive statistics on the patterns and extent of coagglomeration in Canada.

## 2.1 Measuring coagglomeration

Following the literature, we measure coagglomeration in two different ways. First, we leverage plant-level point-pattern data to compute the continuous measure proposed by Duranton and Overman (2005, 2008). Let $\ell_{p(i)}$ denote the employment of plant $p$ in industry $i$ and $\ell_{q(j)}$ the employment of plant $q$ in industry $j \neq i$. The Duranton-Overman (henceforth, DO) coagglomeration measure for the industry pair $ij$ at distance $d$ is given by:

$$\widehat{k}_{ij}(d) = \frac{1}{h \sum_{p=1}^{n_i} \sum_{q=1}^{n_j} \ell_p \ell_q} \sum_{p=1}^{n_i} \sum_{q=1}^{n_j} \ell_p \ell_q f\left(\frac{d - d_{pq}}{h}\right), \tag{1}$$

where $d_{pq}$ denotes the (great circle) distance between plants $p$ and $q$; $n_i$ and $n_j$ are the numbers of plants in industries $i$ and $j$; $f$ is a Gaussian kernel function; and $h$ is a bandwidth parameter, set according to Silverman's rule. As explained in Appendix A.3, we can construct confidence bands for the measures (1) using bootstrapping techniques that control for multiple hypothesis testing as in Duranton and Overman (2005). We do this to assess the statistical significance of the coagglomeration patters that we document in Section 2.2.

The measure (1) can be viewed as the (kernel-smoothed) probability distribution function (PDF) of bilateral distances between plants in industries $i$ and $j$. Since we are interested in the degree of coagglomeration of plants up to some distance $d$, we use as intuitive metric for coagglomeration the cumulative distribution function (CDF) of (1). Formally, we construct:

$$\widehat{K}_{ij}(\overline{d}) = \sum_{d \leq \overline{d}} \widehat{k}_{ij}(d). \tag{2}$$

The measure (2) can be interpreted as the (kernel-smoothed) probability that a pair of employees working in two different plants—drawn randomly from industries $i$ and $j$—work less than $\overline{d}$ kilometers away from one another. The larger this measure, the more coagglomerated is employment of the industry pair $ij$ up to distance $\overline{d}$.[3]

Our second coagglomeration measure follows Ellison et al. (2010) who propose a variant of the discrete (and aspatial) Ellison-Glaeser measure of coagglomeration. Computing this

[3]Our measures and counterfactuals are constructed at the worker level, not the plant level. The latter could be done by giving all plants the same weights, i.e., by letting $\ell_p = \ell_q = \overline{\ell}$ for all observations. Doing so would give substantially less weight to (pairs of) large plants. Whether this is desirable or not is a priori unclear and depends on the context of the analysis. Duranton and Overman (2005, p. 1095) find that "when weighting for employment, fewer industries are localized but those that are deviate more strongly from randomness". Their results between weighted and unweighted estimates are, however, relatively similar.

measure requires to select a spatial unit $c$. The Ellison-Glaeser-Kerr (henceforth, EGK) coagglomeration measure is then given by:

$$\widehat{\gamma}_{ij} = \frac{\sum_{c=1}^{C}(s_{ic} - x_c)(s_{jc} - x_c)}{1 - \sum_{c=1}^{C}(x_c)^2}, \tag{3}$$

where $c = 1, 2, \ldots, C$ indexes spatial units; $s_{ic}$ denotes the share of industry $i$ located in $c$; and $x_c$ is the overall share of employment in $c$. We use this measure mainly as a sanity check—most of the literature has used it instead of the computationally more demanding DO measure (e.g., Faggio et al., 2017; Steijn et al., 2022). For reasons that will become clear later, our preferred measure of coagglomeration is however the DO measure.

Before proceeding, one remark is in order. Our coagglomeration measures (2) and (3) are, by construction, symmetric. However, it may not be ideal to measure the colocation of industries symmetrically if the degree of localization between the industries in the pair differ. If the spatial patterns of industries show central location patterns, then more ubiquitous industries will be found where more localized industries are located but not vice versa (see, e.g., Mori et al., 2008).[4] While we acknowledge the potential upside of looking at asymmetric coagglomeration patterns, we stick with the standard symmetric measures in what follows as those have been the most widely used.

## 2.2 Data and descriptives

Our main dataset is the Scott's Business Register, which provides information on the detailed location of most manufacturing plants in Canada. These data have been used before and are the best alternative to Statistic Canada's confidential establishment-level data (see Behrens and Bougna, 2015; Behrens et al., 2018). Appendixes A.1 and A.2 provide additional information. For each plant, we have information on its location, industry, and an estimate of total on-site employment. Figure A.1 in the appendix illustrates the granularity of our data to measure the coagglomeration patterns of industries. These data can be used directly to compute (1) and (2). For the latter, we use census divisions as the geographic units (more on this later). We compute coagglomeration measures for 86 4-digit industries for the years 2001, 2003, and 2005, which yields a total of $3,655 \times 3 = 10,965$ unique industry pairs.[5]

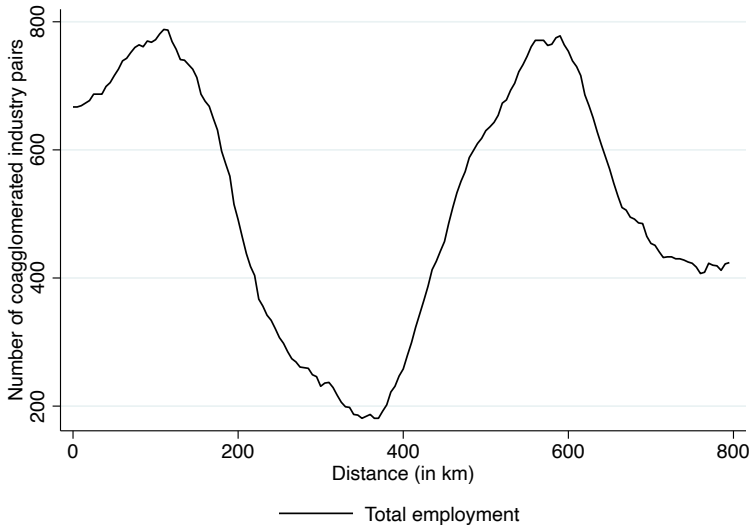
[4] We thank Tomoya Mori for bringing this point to our attention. See also Billings and Johnson (2016, pp.7–8) for a discussion on symmetric vs asymmetric measures of industrial colocation. They find that *"colocalization is not symmetric when comparing 5411 Legal Services with 7224 Drinking Places. When we assign 5411 Legal Services to be industry $j$ and 7224 Drinking Places to be industry $k$, we find significant colocalization. This relationship indicates that lawyers are located significantly closer to bars than the overall population of all industries. Interestingly, the reverse case of bars locating proximate to lawyers is not significantly colocalized. In this case, Drinking Places locate towards a wide range of potential customers, not just lawyers."*

[5] We use 800 kilometers as the cutoff distance (Behrens and Bougna, 2015) and a step size of 5 kilometers. To speed up computations and deal with prohibitive memory requirements, we use the binning approximation

Exhibit (a) of Figure 1 depicts the number of significantly coagglomerated industry pairs by distance according to the DO measure (see Appendix A.3 for a discussion of how we assess the statistical significance). As shown, the number of significantly coaggomerated manufacturing industry pairs increases until about 100 kilometers, and then decreases regularly until about 400 kilometers. It then increases again until a second peak is reached at about 550–600 kilometers. The latter corresponds roughly to the distance between the two largest Canadian metropolitan areas, Toronto and Montréal. The first peak in exhibit (a) of Figure 1 shows that manufacturing industries tend to coagglomerate most often at intermediate distances and less often at very short ones. This pattern suggests that input-output linkages, which are less sensitive to distance than knowledge sharing or common labor pools, may be a primary driver of coagglomeration patterns.[6]

Figure 1: Descriptives for coagglomeration patterns.

(a) # of significantly coagglomerated industry pairs in 2001




(b) Summary statistics

| | DO coagglomeration measure | | |
|---|---|---|---|
| | coagglomerated | random | codispersed |
| 2001 | 0.587 | 0.266 | 0.147 |
| 2003 | 0.530 | 0.336 | 0.134 |
| 2005 | 0.461 | 0.376 | 0.163 |
| | **EGK coagglomeration measure** | | |
| | Total employment | | |
| | Average | Min | Max |
| 2001 | -0.003 | -1.108 | 1.294 |
| 2003 | -0.001 | -0.700 | 1.234 |
| 2005 | -0.001 | -0.807 | 1.306 |

*Notes:* The top panel reports shares of industry pairs. Statistical significance based on 200 bootstrap replications and global confidence bands (see Appendix A.3 for details). The bottom panel reports the average, minimum, and maximum values for the EGK measures of coagglomeration computed at the census division level.

The top panel in exhibit (b) of Figure 1 summarizes the shares of coagglomerated, random, and codispersed industry pairs for total employment in 2001, 2003, and 2005. As shown, a substantial share of industry pairs is coagglomerated, even if the number of coagglomerated pairs decreases between 2001 and 2005 (about 12 percentage points fewer coagglomerated industry pairs). Conversely, random colocation patterns become more prevalent (about 10 percentage points more). Codispersion patterns remain fairly stable at around 15%. These trends echo findings by Behrens and Bougna (2015), who document that the number of significantly lo-

proposed by Scholl and Brenner (2015). This is especially important when constructing the confidence bands—which are based on 200 Monte Carlo replications (see Appendix A.3 for details).

[6]This observation is consistent with the peak in Canadian shipping patterns at about 100 kilometers, as documented by Behrens and Brown (2018) using Canadian trucking microdata.

calized manufacturing industries has decreased in Canada between 2001 and 2009. Here, we report similar trends for the coagglomeration of industry pairs.

Last, the bottom panel in exhibit (b) of Figure 1 provides summary statistics for the EGK coagglomeration measure at the census division level. As shown and as usual in the literature, it is centered around zero and displays some right skew in its distribution.[7]

# 3 Determinants of coagglomeration

Following a long tradition that dates back to Marshall in 1890, we consider proxies for the gains from geographic concentration related to the costs of moving goods, people, and ideas. More precisely, we consider that industry pairs are linked by customer-supplier relationships as they buy and sell intermediate goods from one another (input-output linkages); by the similarity of the types of workers they require (labor market pooling); and by planned and unplanned information exchanges (knowledge spillovers). As moving goods, people, and ideas is costly, industries tend to coagglomerate to reduce these costs. We now provide a quick overview of the data we use and the variables we construct (details are relegated to Appendix A.4). We distinguish between flow-based and similarity-based measures of the Marshallian agglomeration forces. The former capture directly some form of exchange between industry pairs, whereas the latter capture some form of 'likeness' between industries. We also discuss the respective contribution of the two measures in shaping coagglomeration patterns.

## 3.1 Flow-based covariates

**Input-output linkages.** Industries buy and sell from each other, and these input-output linkages may drive coagglomeration. We construct the shares of inputs ($\omega_{ij}^{\text{in}}$) and outputs ($\omega_{ij}^{\text{out}}$) that industry $i$ buys from or sells to industry $j$. Following previous work, we make the measure symmetric by taking the maximum over pairs, i.e., over $ij$ and $ji$.[8] Our measure of the strength of input-output links between industries $i$ and $j$ combines input and output shares as follows:

$$\text{input-output linkages}_{ij} = \max\{\omega_{ij}^{\text{in}}, \omega_{ji}^{\text{in}}, \omega_{ij}^{\text{out}}, \omega_{ji}^{\text{out}}\}.$$

[7] Our minimum and maximum values for the EGK measures are larger than those reported by Ellison et al. (2010) or Faggio et al. (2017) for the US and the UK, respectively. This may be related to either differences in the size of spatial units or to the fact that Canada's economy is 10 times smaller than the US economy and 2 times smaller than the UK economy.

[8] The DO and EGK measures are symmetric between pairs $ij$ and $ji$. Since we include only one of them in the regressions, we need to make sure that the results are not sensitive to which pair $ij$ or $ji$ is included. This requires that the measures be symmetric (see footnote 4 for a discussion on the potential shortcomings of symmetric coagglomeration measures). Alternatively, we could take the average, which makes little difference.

We will also use the strength of input links (input linkages$_{ij} \equiv \max\{\omega_{ij}^{\text{in}}, \omega_{ji}^{\text{in}}\}$) and of output links (output linkages$_{ij} \equiv \max\{\omega_{ij}^{\text{out}}, \omega_{ji}^{\text{out}}\}$) separately in robustness checks.

**Worker flows.** Workers move between industries and this 'job hopping' (Fallick et al., 2006) may increase productivity by allowing for better matching and the transfer of human capital. As a direct measure of labor market flows, we compute an index of observed labor mobility across manufacturing industries using US public use microdata. We use moves between manufacturing industries to construct a matrix that contains the share of workers leaving industry $i$ to move to industry $j$, move$_{ij}$. Industries with a larger value of move$_{ij}$ are likely to be more similar in terms of their labor requirements. To obtain a symmetric measure, we take the maximum between $ij$ and $ji$ as follows:

$$\text{worker flows}_{ij} = \max\{\text{move}_{ij}, \text{move}_{ji}\}.$$

**Knowledge flows.** Formal and informal knowledge spills across industries, and these forms of knowledge spillovers are likely drivers of coagglomeration. We follow previous work by Kerr (2008) and Ellison et al. (2010) and construct a measure of knowledge flows using US patent citations data. Our proxy for knowledge flows is based on the shares of patents that industries $i$ or $j$ cite (citations$_{ij}$) and which originate from the other industry. Our proxy for knowledge flows is the maximum of the shares of patent citations between $ij$ and $ji$ as follows:

$$\text{knowledge flows}_{ij} = \max\{\text{citations}_{ij}, \text{citations}_{ji}\}.$$

## 3.2 Similarity-based covariates

**Input-output correlation.** Industries that are more similar in terms of their customer or supplier base may coagglomerate. We construct a measure of input-output similarity by computing the correlation between the vectors of input shares (corr input shares$_{ij}^{\text{mfg}}$) or output shares (corr output shares$_{ij}^{\text{mfg}}$) of industries $i$ and $j$ with all manufacturing industries. We then take the maximum between the similarity of the industries' input- and output structures as follows:

$$\text{input-output correlation}_{ij} = \max\{\text{corr input shares}_{ij}^{\text{mfg}}, \text{corr output shares}_{ij}^{\text{mfg}}\}.$$

**Occupational correlation.** Industries that are more similar in terms of the workers they require may coagglomerate due to labor market considerations. We construct a measure of occupational employment similarity of the workforce in the different industries. To this end, we use US industry data that break down employment by occupations. Computing the vector of occupation shares for each industry, we then computed the correlation coefficient between the vectors of occupational shares of industries $i$ and $j$, occupational correlation$_{ij}$. This measure is by construction symmetric.

**Technological relatedness.** Industries that are more similar in the technologies they use may coagglomerate because of knowledge and learning externalities. We construct a measure of 'technological relatedness' (e.g., Scherer, 1984; Audretsch et Feldman, 1996) from Canadian patent office data. For each patent, we have a stochastic concordance between technology classes and industries in which the patent is likely to be used. We compute the correlation coefficient (at the patent level) between the probabilities that patents are used by industries $i$ and $j$. This yields our measure technological relatedness$_{ij}$, which is by construction symmetric.

Appendix Table A.1 provides summary statistics for our measures, as well as details on the data sources, industry level of aggregation, and available years. Appendix Table A.2 shows that, as expected, our Marshallian covariates are all positively correlated to some degree.

## 3.3 Flows vs similarity

The extant literature mixes flow-based and similarity-based measures, although these are conceptually different.[9] Consider, e.g., input-output linkages. We can either directly measure exchanges between industries from input-output tables, which tell us how many inputs flow between industries; or we can measure how similar industries are in terms of the inputs they generally require. While the former directly captures costly interactions across space, the latter captures more vaguely these interactions via some idea of 'relatedness'. A particularly telling illustration can be given in terms of labor market pooling. The literature usually constructs a measure of labor market similarity by computing the correlation between occupation shares in the two industries. The underlying idea is that two industries with a high correlation in their employment structure hire 'from the same labor pool' and thus will tend to make similar location choices. This does, however, not imply that there is any fundamental reason that the industries benefit from being close to one another. Imagine for example that all IT workers like mild California winters and end up clustering in the San José area. Two industries that rely on IT workers will end up coagglomerated to hire from the same labor pool, but if the IT workers had more heterogeneous preferences and were split across a larger number of places the industries might be less coagglomerated. The usual story one has in mind is that there is better matching or more job hopping in thick labor markets, but this story is only vaguely related to the occupational similarity metric used in the regression analysis. A flow-based measure of how workers move between firms in the two industries provides a more direct measure of job

[9]There is little discussion in the literature on how results depend on flow-vs-similarity based measures. Faggio et al. (2017, p.89) provide one of the rare short discussions on this issue. They state that *"the labor pooling proxy is based on the correlation between the two industries' occupation mixes, while the input sharing variable is constructed using maximum flows in the sector pair. This could imply that our approach is skewed toward picking up significant input-output linkages only for highly coagglomerated industries, while the labor pooling measure could be more significant in other parts of the distribution."*

hopping or, eventually, matching.

In what follows, we take no stand on which measures are better or worse and provide estimates using both sets of measures. As a reference point, we will also provide estimates which are based on a mix between flow- and similarity-based measures, as these specifications have been extensively used before. We refer to these estimates as 'Ellison-Glaeser-Kerr' (EGK) estimates (see Ellison et al., 2010; Faggio et al., 2017). Last, we will also see that we can estimate an 'all dressed' model where both sets of covariates are jointly included. This is possible since, as stated above, those measures capture different aspects of the agglomeration mechanisms.

# 4 Empirical analysis

## 4.1 Estimating equation

We follow the literature and regress our two sets of coagglomeration measures on our proxies for the Marshallian agglomeration forces. More precisely, we follow Ellison et al. (2010) and Faggio et al. (2017) and run regressions of the following form:

$$\text{coagglo}_{ij,t} = \alpha_{io}\text{io}_{ij,t} + \alpha_{lm}\text{lm}_{ij,t} + \alpha_{ks}\text{ks}_{ij,t} + \mathbf{X}_{ij,t}\beta + \delta_t + \epsilon_{ij,t}, \tag{4}$$

where $\text{coagglo}_{ij,t}$ denotes the coagglomeration of industry pair $ij$ in period $t$; $\mathbf{X}_{ij,t}$ is a vector of time-varying industry-pair controls; and $\delta_t$ are cohort (year) fixed effects. Our key variables of interest are the three Marshallian agglomeration forces related to input-output linkages ($\text{io}_{ij,t}$), labor market pooling ($\text{lm}_{ij,t}$), and knowledge sharing ($\text{ks}_{ij,t}$).[10]

In what follows, we estimate equation (4) in univariate and multivariate settings. Concerning the latter, we use three combinations of our Marshallian covariates. First, we use as our baseline specification the one proposed by Ellison et al. (2010) and Faggio et al. (2017), which includes input-output linkages, occupational correlation, and knowledge flows (the 'EGK' specification). Second, we use a specification that includes only flow measures (input-output linkages, labor flows, and knowledge flows). We refer to this as the 'flow' specification. Last, we analoguously use a specification that includes only similarity measures (input-output correlation, occupational correlation, and technological relatedness). We refer to this as the 'similarity' specification.

We will also consider the interactions of our three Marshallian agglomeration forces with

[10]Since coagglomeration patterns vary slowly, running panel regressions makes little sense. We furthermore do not have enough time-series variation to reasonably run those regressions (see Steijn et al., 2022, for panel regressions using a longer time series of coagglomeration measures constructed from US county business patterns data). In what follows, we mostly report pooled regressions, but our main results also hold in cross-sectional regressions, which we relegate to robustness checks.

transport costs and estimate regressions of the following form:

$$\begin{aligned}\text{coagglo}_{ij,t} &= \alpha_{io}\text{io}_{ij,t} + \alpha_{lm}\text{lm}_{ij,t} + \alpha_{ks}\text{ks}_{ij,t} + \mathbf{X}_{ij,t}\beta + \gamma\text{AVTC}_{ij,t} + \delta_t \\ &\quad + \alpha_{io}\text{io}_{ij,t} \times \text{AVTC}_{ij,t} + \alpha_{lm}\text{lm}_{ij,t} \times \text{AVTC}_{ij,t} + \alpha_{ks}\text{ks}_{ij,t} \times \text{AVTC}_{ij,t} + \epsilon_{ij,t}. \quad (5)\end{aligned}$$

Transport costs are important to understand the geographic structure of manufacturing industries, both in economic geography models and in the empirical investigation of agglomeration patterns. Yet, although they are included as a control in most analyses, we rarely have direct measures that allow to estimate a specification such as (5).[11] As pointed out by Combes and Gobillon (2015) this should, however, be done, especially for input-output linkages whose strength is affected by transport costs.

## 4.2 Controls

We now briefly describe our controls (more details are given in Appendix A.5). Our 'industry $ij$ pair' variables $\mathbf{X}_{ij,t}$ include controls related to: (i) the pair's dissimilarity in terms of the input shares they buy from and the output shares they sell to primary industries; and (ii) the pair's dissimilarity in terms of the input shares they buy from and the output shares they sell to business services industries. These measures control for how dissimilar the industries are in their primary- and business services input and output profiles. They thus control for the propensity of industries to colocate spuriously if both require the same access to primary inputs or business services (Faggio et al., 2017).[12]

While the foregoing controls are relatively standard in the literature, we add three additional controls that we believe are important. Since these controls are less used in the literature—or not used at all—we briefly comment on our rationale for including them.

**Share of multiunit firms.** Usually, we do not worry about whether or not plants split their activity across several locations. Yet, the more an industry has firms with multiple plants, the more likely it is that firms can spatially separate occupations such as management and R&D

[11] Following Ellison and Glaeser (1999), Ellison et al. (2010, p.1203) use *"state level characteristics that afford natural advantages in terms of* [...] *transportation costs"* to model a counterfactual industry distribution based on a battery of state-level characteristics. Faggio et al. (2017) compute a 'transport dissimilarity' index between industries based on the industries' sourcing patterns from transport-related sectors. This index is similar to the dissimilarity control for primary inputs and business services used in our analysis.

[12] Results from the US and the UK suggest that there is little interaction between natural advantage and Marshallian factors. Controlling for natural advantage has little effect on the coefficients of the Marshallian forces. Faggio et al. (2017, p.86) write: *"Interestingly, controlling for the dissimilarity [of use of natural resources or business services] proxies does not change in any meaningful way the three Marshallian coefficients, suggesting that access to natural resources and nonmanufacturing industries does not bias the results in simple models without the additional controls."* Ellison et al. (2010, p.1206) state that: *"natural advantages and Marshallian factors are mostly orthogonal to one another."*

from occupations such as production. If these types of occupations differ in how intensively they rely on the different agglomeration forces, the coagglomeration patterns of industries may be different. Furthermore, it is well known that multiplant firms are significantly larger than single-plant firms (Bernard and Jensen, 1995, Table 4), and that larger firms are likely to be less dependent on agglomeration economies (Rosenthal and Strange, 2003), which may thus again affect colocation patterns. In what follows we create a control that is the maximum of the share of plants in multiunit firms in both industries as follows:

$$\text{multiplant share}_{ij,t} = \max\{\text{multiplant share}_{i,t}, \text{multiplant share}_{j,t}\}.$$

**Transport costs.** We leverage high-quality time-varying industry-level ad valorem transport costs estimated from commodity flow survey microdata (see Behrens et al., 2018, for details). Our industry-pair measure of transport cost is defined as follows:

$$\text{Ad valorem transport costs}_{ij,t} = \max\{\text{AV transport costs}_{i,t}, \text{AV transport costs}_{j,t}\}.$$

**Within-plant coagglomeration.** Our last control is conceptually important and deserves a longer discussion. While the literature has measured coagglomeration across plants, *the most obvious form of geographic colocation of economic activity is within plants.* Few establishments do only one thing. On the contrary, many establishments do multiple 'related' things. This may be due to a host of unobserved locational and technical features such as complementarities and indivisibilities in the production process, or the existence of specific infrastructure and institutions. We think it is reasonable to assume that most unobserved features that drive the coagglomeration of industries across plants also drive the coagglomeration of industries within plants.[13] Thus, the latter—provided we can observe it—may be a good proxy to control for a host of otherwise unobservable features that drive geographic colocation for reasons other than the forces we are interested in.

Conceptually, assume that there are unobserved factors $Z_{ij,t}$ that are related to the coagglomeration of industries $i$ and $j$. The 'true' regression model is

$$\text{coagglo}_{ij,t} = \alpha_{io}\text{io}_{ij,t} + \alpha_{lm}\text{lm}_{ij,t} + \alpha_{ks}\text{ks}_{ij,t} + \mathbf{X}_{ij,t}\beta + \mathbf{Z}_{ij,t}\gamma + \delta_t + \epsilon_{ij,t}. \tag{6}$$

[13] Assume that industries $i$ and $j$ both benefit from a local trade association that promotes these industries and has an effect on productivity. If the existence of the chamber of commerce is independent of input-output linkages, labor market pooling, or knowledge spillovers between $i$ and $j$, this poses no problem for our estimations, but it is very likely that the industries $i$ and $j$ the chamber promotes have some Marshallian relationships. They could, e.g., benefit from technological complementarities, common input suppliers, or labor-force specificities. Our assumption is that the local trade association's positive effects will at least partly map into more within-plant coagglomeration (which is correlated with IO linkages, labor market pooling, and knowledge flows). Controlling for the within-plant coagglomeration is hence going to pick up some of these unobserved factors.

If the omitted factors $\mathbf{Z}_{ij,t}$ are correlated with our Marshallian covariates, we have classical omitted variables bias and the estimates of $\alpha_{io}$, $\alpha_{lm}$, and $\alpha_{ks}$ are not consistent. Assume that the unobserved factors $\mathbf{Z}_{ij,t}$ also lead plants to do 'related things' in house, i.e.,

$$\text{within-plant share}_{ij,t} = \mathbf{Z}_{ij,t}\theta + \widetilde{\alpha}_{io}\text{io}_{ij,t} + \widetilde{\alpha}_{lm}\text{lm}_{ij,t} + \widetilde{\alpha}_{ks}\text{ks}_{ij,t} + \nu_{ij,t}. \tag{7}$$

Since we observe within-plant share$_{ij,t}$ in our data, it may serve as a proxy for the unobserved $\mathbf{Z}_{ij,t}$. Assuming that $\theta \neq 0$, we can estimate

$$\begin{aligned} \text{coagglo}_{ij,t} \;=\; & \left(\alpha_{io} - \frac{\widetilde{\alpha}_{io}}{\theta}\right)\text{io}_{ij,t} + \left(\alpha_{lm} - \frac{\widetilde{\alpha}_{lm}}{\theta}\right)\text{lm}_{ij,t} + \left(\alpha_{ks} - \frac{\widetilde{\alpha}_{ks}}{\theta}\right)\text{ks}_{ij,t} \\ & + \frac{\gamma}{\theta}\text{within-plant share}_{ij,t} + \mathbf{X}_{ij,t}\beta + \delta_t + \left(\epsilon_{ij,t} - \frac{\nu_{ij,t}}{\theta}\right) \end{aligned} \tag{8}$$

to get consistent estimates of our Marshallian covariates, provided that the compound error term $\epsilon_{ij,t} - \frac{\nu_{ij,t}}{\theta}$ is uncorrelated with our explanatory variables.[14]

In our data, we observe each plant's primary activity and up to four secondary activities. We can hence compute for each industry $i$ the share of establishments in that industry that report secondary activities in industry $j$. We call this measure the within-plant share. As before, we construct a symmetric measure between industries $i$ and $j$ by taking the maximum across the industries:

$$\text{within-plant share}_{ij,t} = \max\{\text{within-plant share}_{i,t}, \text{within-plant share}_{j,t}\}.$$

Our within-plant share variable is correlated at 0.116, 0.121, and 0.118 with the EGK coagglomeration measure in 2001, 2003, and 2005; and at 0.146, 0.155, and 0.149 with the DO coagglomeration measure, respectively (all significant at 1%). This suggests that unobserved features that may drive the coagglomeration of industries within plants may also drive the coagglomeration of industries across plants. Table 1 further shows that our within-plant share is strongly and significantly correlated with all of our six Marshallian proxies, both in univariate and in multivariate regressions. As just explained, this likely reflects unobserved technological complementarities and indivisibilities and suggests that this variable may be important to control for these unobservables in our regressions. Table B.4 in the appendix presents results from a battery of robustness checks. Our results in Table 1 are robust across all specifications.

In what follows, we include our within-plant share in our analysis to capture a number of unobserved factors that drive coagglomeration of industries $i$ and $j$ but are unrelated to Marshallian externalities across plants. The results from Table B.4 show that, as expected, including this control will reduce the coefficients on the Marshallian variables. As we will see this effect is however not strong enough to reduce the coefficients we estimate for the Marshallian determinants to zero.

[14] If we assume that $\nu_{ij,t}$ is uncorrelated with the Marshallian covariates, conditional on $\mathbf{X}_{ij,t}$, then $\epsilon_{ij,t} - \frac{\nu_{ij,t}}{\theta}$ is uncorrelated too.

Table 1: Within-establishment coagglomeration.

| | (1) (Univariate) | (2) (EGK) | (3) (Flow) | (4) (Similarity) | (5) (EGK) | (6) (Flow) | (7) (Similarity) |
|---|---|---|---|---|---|---|---|
| Input-output linkages | 0.467$^a$ | 0.338$^a$ | 0.201$^a$ | | 0.338$^a$ | 0.201$^a$ | |
| | (0.027) | (0.021) | (0.017) | | (0.022) | (0.017) | |
| Input-output correlation | 0.592$^a$ | | | 0.268$^a$ | | | 0.277$^a$ |
| | (0.016) | | | (0.015) | | | (0.016) |
| Labor flows | 0.355$^a$ | | 0.281$^a$ | | | 0.294$^a$ | |
| | (0.014) | | (0.015) | | | (0.016) | |
| Occupational correlation | 0.491$^a$ | 0.355$^a$ | | 0.157$^a$ | 0.356$^a$ | | 0.147$^a$ |
| | (0.015) | (0.013) | | (0.011) | (0.014) | | (0.011) |
| Knowledge flows | 0.139$^a$ | 0.068$^a$ | 0.049$^a$ | | 0.066$^a$ | 0.046$^a$ | |
| | (0.020) | (0.011) | (0.008) | | (0.011) | (0.008) | |
| Technological relatedness | 0.209$^a$ | | | 0.104$^a$ | | | 0.115$^a$ |
| | (0.012) | | | (0.010) | | | (0.010) |
| Ad valorem transport costs | | | | | -0.007 | 0.010 | -0.047$^a$ |
| | | | | | (0.012) | (0.007) | (0.007) |
| Multiplant share | | | | | 0.002 | -0.040$^a$ | -0.011 |
| | | | | | (0.010) | (0.008) | (0.008) |
| Controls $ij$ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Cohort FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 10,292 | 10,292 | 9,729 | 9,729 | 10,045 | 9,506 | 9,506 |
| $R^2$ | | 0.373 | 0.286 | 0.284 | 0.374 | 0.296 | 0.289 |

*Notes:* Results for unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. All regressions include cohort fixed effects. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. The dependent variables is the share of secondary NAICS codes in industry $j$ reported by plants in industry $i$. Column (1) summarizes six independent univariate regressions. Industry-pair controls $ij$ are included. Huber-White robust standard errors in parentheses. $^a$p<0.01, $^b$p<0.05, $^c$p<0.1.

## 4.3 Results

Table 2 reports our baseline results for the DO and EGK coagglomeration metrics. We construct the DO measure using 15 kilometers as the distance cutoff, and the EGK metric using census divisions as the spatial unit of analysis.[15] Table 2 shows that the coagglomeration patterns are positively and significantly associated with most of the flow-based and similarity-based measures of our Marshallian covariates, both for the DO and the EGK metrics. Input-output linkages and occupational correlation are always positive and significant across all specifications we estimate, both in univariate and multivariate regressions. These results are in line with the literature and suggest that input-output linkages and labor market pooling are robustly associated with coagglomeration. Labor flows are generally positively and significantly related to coagglomeration using both the DO and EGK metrics, though the effect does not survive the inclusion of our controls (see columns (6) and (13)).

[15]We verified that the DO results are generally robust to reasonable alternative distance cutoffs, with the magnitude of agglomeration effects decreasing with distance. The EGK results are robust to alternatively using census metropolitan areas as larger spatial units.

Table 2: Baseline coagglomeration regressions.

| | (a) DO coagglomeration measure | | | | | | | (b) EGK coagglomeration measure | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) (Univar) | (2) (EGK) | (3) (Flow) | (4) (Similarity) | (5) (EGK) | (6) (Flow) | (7) (Similarity) | (8) (Univar) | (9) (EGK) | (10) (Flow) | (11) (Similarity) | (12) (EGK) | (13) (Flow) | (14) (Similarity) |
| Input-output linkages | 0.086$^a$ | 0.032$^a$ | 0.054$^a$ | | 0.027$^b$ | 0.059$^a$ | | 0.094$^a$ | 0.063$^a$ | 0.085$^a$ | | 0.045$^a$ | 0.068$^a$ | |
| | (0.012) | (0.011) | (0.016) | | (0.011) | (0.015) | | (0.013) | (0.013) | (0.017) | | (0.013) | (0.018) | |
| Input-output correlation | 0.153$^a$ | | | 0.116$^a$ | | | 0.047$^a$ | 0.104$^a$ | | | 0.049$^a$ | | | 0.009 |
| | (0.010) | | | (0.016) | | | (0.016) | (0.009) | | | (0.017) | | | (0.018) |
| Labor flows | 0.143$^a$ | | 0.101$^a$ | | | -0.005 | | 0.058$^a$ | | 0.033$^a$ | | | -0.003 | |
| | (0.010) | | (0.011) | | | (0.011) | | (0.008) | | (0.010) | | | (0.011) | |
| Occupational correlation | 0.152$^a$ | 0.121$^a$ | | 0.122$^a$ | 0.049$^a$ | | 0.049$^a$ | 0.111$^a$ | 0.087$^a$ | | 0.078$^a$ | 0.060$^a$ | | 0.061$^a$ |
| | (0.010) | (0.011) | | (0.013) | (0.011) | | (0.013) | (0.010) | (0.011) | | (0.014) | (0.012) | | (0.015) |
| Knowledge flows | 0.023$^c$ | 0.005 | 0.024$^c$ | | -0.010 | 0.003 | | 0.057$^a$ | 0.041$^a$ | 0.036$^a$ | | 0.035$^a$ | 0.029$^b$ | |
| | (0.012) | (0.012) | (0.014) | | (0.010) | (0.012) | | (0.012) | (0.012) | (0.013) | | (0.012) | (0.013) | |
| Technological relatedness | -0.048$^a$ | | | -0.131$^a$ | | | -0.075$^a$ | 0.005 | | | -0.023$^b$ | | | -0.025$^b$ |
| | (0.011) | | | (0.011) | | | (0.010) | (0.010) | | | (0.011) | | | (0.011) |
| Within-plant share | | | | | 0.067$^a$ | 0.132$^a$ | 0.134$^a$ | | | | | 0.061$^a$ | 0.099$^a$ | 0.102$^a$ |
| | | | | | (0.011) | (0.015) | (0.015) | | | | | (0.011) | (0.015) | (0.014) |
| Ad valorem transport costs | | | | | -0.160$^a$ | -0.154$^a$ | -0.152$^a$ | | | | | -0.004 | 0.003 | -0.004 |
| | | | | | (0.009) | (0.009) | (0.009) | | | | | (0.009) | (0.010) | (0.010) |
| Multiplant share | | | | | -0.275$^a$ | -0.287$^a$ | -0.266$^a$ | | | | | -0.033$^a$ | -0.043$^a$ | -0.024$^b$ |
| | | | | | (0.009) | (0.009) | (0.010) | | | | | (0.011) | (0.011) | (0.011) |
| Controls $ij$ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Cohort FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 10,292 | 10,292 | 9,729 | 9,729 | 10,045 | 9,506 | 9,506 | 10,292 | 10,292 | 9,729 | 9,729 | 10,045 | 9,506 | 9,506 |
| $R^2$ | | 0.075 | 0.069 | 0.089 | 0.206 | 0.209 | 0.214 | | 0.020 | 0.010 | 0.011 | 0.024 | 0.015 | 0.015 |

*Notes:* The dependent variable in columns (1)–(7) is the DO $K$-density evaluated at 15 kilometers distance. The dependent variable in columns (8)–(14) is the EGK coagglomeration measure computed at the census division level. The sample is all unique industry pairs $ij$ obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. Sample sizes vary across specifications because of missing covariates. We exclude small industry pairs where at least one industry has less than 30 plants. All variables are standardized. All regressions include cohort fixed effects. Industry-pair controls $ij$ (dissimilarity in industry-level input- and output shares with primary and business services industries) are included, except in the univariate regressions in columns (1) and (8). We report Huber-White robust standard errors in parentheses. $^a p < 0.01$, $^b p < 0.05$, $^c p < 0.1$.

Technological relatedness is the only measure that is either insignificant or negatively associated with coagglomeration. The two Marshallian covariates that show some differences between the EGK and DO measures are input-output correlation and knowledge spillovers. Using the EGK metric yields robustly positive coefficients for knowledge spillovers, although their magnitude is smaller than for input-output linkages and labor market pooling. They are, however, mainly insignificant when using the DO measure. That knowledge spillovers display weaker effects than the other covariates is a usual finding in the literature on agglomeration and coagglomeration (see, e.g., Rosenthal and Strange, 2023; Jofre-Monseny at al., 2011; Ellison et al., 2010; Faggio et al., 2017). Input-output correlation has larger and more significant effects using the DO measures than the EGK measures. Overall, our results show that flow- and similarity-based measures provide a relatively coherent picture, though some results are sensitive to the measures used. Generally, similarity-based measures provide slightly stronger results than flow-based measures, at least in terms of model fit as measured by the $R^2$. The latter one is, quite surprisingly, much higher using the DO rather than the EGK based measures.

Turning to our three additional controls, note first that the within-plant share is always positively and significantly correlated with coagglomeration, whereas the multiplant share is always negatively and significantly correlated with coagglomeration, irrespective of the coagglomeration metric. Thus, industries with a larger share of multiplant firms are less coagglomerated than those with fewer multiplants, all other things being equal. This already points to the idea that multiplant firms— which tend to be large firms—are less dependent on agglomeration economies, a point we return to later.

Turning to transport costs, there is a sharp difference between the DO and EGK metrics. While industry pairs with high transport costs are less coagglomerated according to the DO metric, transport costs are never significantly associated with coagglomeration in the EGK metric. This may be due to the size of the spatial scale of census divisions and the a-spatial nature of the EGK index (recall that the spatial arrangement of census divisions can be modified without affecting the value of the EGK index).

Last, observe that our three additional controls—within-plant agglomeration, transport costs, and the share of multiplant firms—have sizable bite as they boost the $R^2$ quite a bit. Since they are correlated with our Marshallian covariates, they tend to reduce the importance of the latter, without however fully offsetting their effect.

Appendix Tables B.5 and B.6 summarize the results of a battery of robustness checks. The latter include regressions where we: (i) use all industries, including small ones with less than 30 plants; (ii) exclude industry pairs within the same NAICS 3-digit industries (to deal with the potential problem that some of our variables are only available at this level of industrial aggregation); (iii) use the years 2001, 2003, and 2005 separately as cross sections; (iv) use separate input- and output linkages; (v) use a make-based measure of patent citations (instead of a use-based measure); (vi) use the US technological relatedness measure by Steijn et al.

(2022) as an alternative to ours; (vii) run IV regressions where we instrument the Canadian input-output linkages with their US counterparts; (viii) exclude the within-plant share control; and (ix) use all six Marshallian covariates simultaneously.

Tables B.5 and B.6 show that our results are remarkably robust across specifications. The only differences are that when we use input- and output linkages separately, the former are negative in Table B.5 and insignificant in Table B.6. Output linkages, on the other hand, are strongly associated with coagglomeration patterns. The IV regressions in columns (10) and (11) yield somewhat larger estimates for input-output linkages, both for the DO and EGK coagglomeration measures. Excluding the within-plant share from the regressions leads to larger estimates for input-output linkages and labor market pooling, but leaves the estimates for knowledge spillovers largely unchanged. Last, including all six measures yields similar results for the DO and EGK measures: input-output linkages and occupational correlation have positive estimates, whereas technological relatedness is negative. While knowledge spillovers are positive using the EGK measure, they are insignificant using the DO measure.

## 4.4 The effects of transport costs

Since we have detailed measures of transport costs, we next investigate how they interact with the estimates for our Marshallian covariates. Table 3 summarizes the results. As shown, transport costs are always negatively associated with coagglomeration: high transport cost pairs are less coagglomerated. Column (1) shows that the interaction with input-output linkages is, surprisingly, negative. This result holds across columns (1) to (4). This suggests that there is less coagglomeration for industries that buy or sell a lot from another when transport costs are high. Column (4) further shows that the within-plant share is less linked to coagglomeration for high transport cost industries.

To understand the counterintuitive result that high transport cost pairs appear less coagglomerated when they have strong input-output linkages, column (5) shows that the within-plant share increases with the interaction between input-output linkages and transport costs. Combined with the results from column (4), one interpretation of this finding is that firms tend to internalize input-output linkages when those become quite costly. As this does by construction not show up in the coagglomeration measures—which embody only cross-industry activities that occur in separate plants—this may explain our result.

Last, and reassuringly, columns (2) and (3) show that transport costs for goods only matter for input-output linkages: all interaction terms with proxies for labor market pooling or knowledge spillovers are insignificant.

Table 3: Transport costs and coagglomeration.

| | (1) (EGK) | (2) (EGK) | (3) (Flows) | (4) (EGK) | (5) (Within) |
|---|---|---|---|---|---|
| Ad valorem transport costs (AVTC) | -0.143$^a$ | -0.144$^a$ | -0.133$^a$ | -0.137$^a$ | -0.026$^b$ |
| | (0.009) | (0.015) | (0.011) | (0.009) | (0.012) |
| Input-output linkages | 0.088$^a$ | 0.089$^a$ | 0.207$^a$ | 0.073$^a$ | 0.263$^a$ |
| | (0.020) | (0.021) | (0.029) | (0.021) | (0.035) |
| Input-output linkages × AVTC | -0.042$^a$ | -0.042$^a$ | -0.090$^a$ | -0.032$^a$ | 0.050$^b$ |
| | (0.009) | (0.009) | (0.015) | (0.009) | (0.024) |
| Occupational correlation | 0.047$^a$ | 0.046$^b$ | | 0.044$^a$ | 0.357$^a$ |
| | (0.011) | (0.019) | | (0.011) | (0.014) |
| Occupational correlation × AVTC | | 0.000 | | | |
| | | (0.008) | | | |
| Labor flows | | | -0.025 | | |
| | | | (0.022) | | |
| Labor flows_share × AVTC | | | 0.012 | | |
| | | | (0.011) | | |
| Knowledge flows | -0.010 | -0.010 | 0.018 | -0.010 | 0.065$^a$ |
| | (0.010) | (0.024) | (0.026) | (0.010) | (0.011) |
| Knowledge flows × AVTC | | 0.001 | -0.009 | | |
| | | (0.015) | (0.016) | | |
| Within-plant share | 0.070$^a$ | 0.070$^a$ | 0.133$^a$ | 0.107$^a$ | |
| | (0.010) | (0.010) | (0.015) | (0.017) | |
| Within-plant share × AVTC | | | | -0.020$^a$ | |
| | | | | (0.007) | |
| Observations | 10,045 | 10,045 | 9,506 | 10,045 | 10,045 |
| $R^2$ | 0.208 | 0.208 | 0.212 | 0.208 | 0.377 |

*Notes:* The dependent variable in columns (1)–(4) is the DO $K$-density at 15 kilometers distance; and the within-plant share in column (5). Results are for all unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. Sample sizes vary across specifications because of missing covariates. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All regressions include cohort fixed effects, industry-pair controls $ij$ (dissimilarity in industry-level input- and output shares with primary and business services industries) and the share of multiplant firms (not reported). Huber-White robust standard errors in parentheses. $^a p < 0.01$, $^b p < 0.05$, $^c p < 0.1$.

# 5 Heterogeneous coagglomeration patterns

A recent literature has shown that coagglomeration patterns are heterogeneous along many dimensions, with subsets of industry pairs exhibiting different relations with Marshall's agglomeration forces. Faggio et al. (2017, 2020), for example, show that splits into subsets yield results in line with the idea that industry pairs coagglomerate to exploit the agglomeration forces they are more sensitive to. For example, more dynamic and skilled industries—with high plant turnover and a more educated workforce—are more sensitive to knowledge spillovers than to input-output linkages. The latter matter more for mature and more low-skilled industries.

Another literature has shown that the geographic distribution of employment differs also

functionally, with places specializing in different types of activity (e.g., Duranton and Puga, 2005). Within industries, some higher-level activities—such as research and development or management—tend to occur in specific places where they can benefit from knowledge spillovers and specific types of workers. Other activities, such as production or more routine tasks, tend to occur elsewhere where they can benefit from specific Marshallian agglomeration forces such as input-output linkages.

At a fundamental level, both of these literatures show that the sensitivity of the location of economic activity—either in terms of industry or function—to Marshall's agglomeration forces varies, and that (co)agglomeration patterns reflect this differential sensitivity. A direct corollary, which quite surprisingly has not been investigated until now, is that industry pairs that are more sensitive to some Marshallian force should show stronger colocation of their functions that are relatively more sensitive to that agglomeration force. If, for example, management or R&D are more sensitive to knowledge spillovers than production, industry pairs that display large knowledge flows should colocate more their knowledge-intensive functions. Conversely, if production is more sensitive to input-output linkages, we should see that industry pairs that display large input-output linkages should colocate their production functions more.

## 5.1 Industry-by-occupation employment data and descriptives

To operationalize the foregoing ideas requires that we measure the coagglomeration of industries by broad occupational employment categories. However, computing these coagglomeration patterns, especially at fine spatial scales, is challenging. Ideally, we would use detailed establishment-level data that report the establishment's industry, its employment numbers by occupational categories, and its precise location. Confidentiality issues related to cell sizes, however, prevent us from having access to such data.[16] We thus combine two different datasets to approximate the distribution of manufacturing employment by broad occupational categories at a microgeographic scale.

The first dataset we use are special census tabulations (from the 1996 and 2001 censuses) that split employment by industry and broad occupation at the census division (henceforth, CD) level. The first advantage of these data is that they report employment numbers for 7 broad categories and 142 sectors, more detail than is usually available at the CD level.[17] We retain two broad employment types for our analysis (see Appendix A.6 for details): Management and

[16]The annual survey of manufacturers only reports employment by broad educational categories. One could use confidential linked employer-employee data, but these would not be available at a fine geographic scale.

[17]The sectoral aggregation level is somewhere between NAICS 3- and 4-digit. We use a cross-walk to map those data to NAICS 3- and 4-digit industries. Where only 3-digit data are available, we map the same numbers to all 4-digit industries within the same 3-digit industry. In robustness checks, we exclude those industries to mitigate concerns about these imputations.

R&D (henceforth, MRD); and production-related employment.[18]

The second important advantage of these special tabulations is that they separately report rural and urban parts of the CDs (see Appendix A.7 for details).[19] The drawback of these data is that they are residence based, though we are interested in the colocation of employment at the work place.

To obtain a workplace-based picture of the employment distribution, we combine the special census tabulation data with the Scott's data that we documented before. The Scott's data allow us to spatially allocate the residence-based census occupations-by-industry data to plants, thus approximating the geographic distribution of employment at the workplace. More precisely, let $L_{p(i)}$ denote the total employment of plant $p$ in industry $i$, and let $\theta^o_{p(i)}$ denote the share of occupation $o$ in the plant. We thus have $\ell^o_{p(i)} = \theta^o_{p(i)} \times L_{p(i)}$ workers in occupation $o$ at the site of the plant. We do not observe $\theta^o_{p(i)}$ at the plant level. We hence approximate it as follows. First, we assign the occupation shares from each census division in the special tabulation to all of its postal code centroids.[20] Then, we average these shares over all postal codes within a 25 km radius around each plant (90% of Canadians commute less than 25 kilometers to work) to obtain the average occupation shares $\theta^o_{i,25km}$.[21] We finally use these industry-location specific shares to break down plant-level employment by occupation:

$$\ell^o_{p(i)} = \theta^o_{i,25km} \times L_{p(i)} \tag{9}$$

We use (9) to compute (1) and (3), where we replace total employment $\ell_{p(i)}$ by our constructed employment $\ell^o_{p(i)}$ in occupation $o$. The rest of the construction of the coagglomeration measures proceeds as before.

Figure 2 shows that there is substantial heterogeneity in coagglomeration of occupations accross industry pairs. Management and R&D (MRD) is substantially more coagglomerated at short distances (up to 80 km) than production. This suggests that the drivers of coagglomeration are likely to be different across both occupation types. The pattern substantiated by

[18] We choose to aggregate management and R&D into a single category. We view this category as that of the 'knowledge intensive' business functions. We acknowledge that the coagglomeration mechanisms of headquarters (management) and R&D may be different along other dimensions. Unfortunately, we cannot really explore these other dimensions using our data.

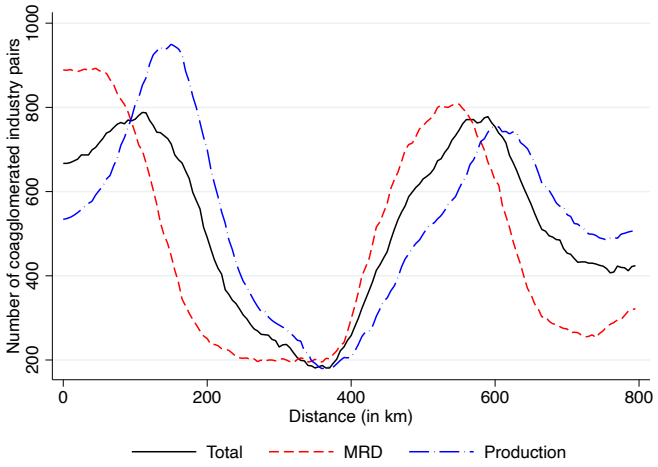
[19] Distinguishing between rural and urban parts is important as cities increasingly specialize by functions rather than by industries (Duranton and Puga, 2005). Hence, plants in the same CD are likely to do different things and have a different employment mix depending on whether they are located in rural or urban parts of the CD. Our data allow to distinguish these.

[20] This allows to work at a much finer within-census division geography. There are more than 800 thousand postal codes in Canada. Especially in cities, these provide a very precise geographic location.

[21] See Statistics Canada (2006; catalogue number 97-561-XCB2006010) for the commuting figures. We apply an exponential distance decay with exponent -0.01 (Ahlfeldt et al., 2015) to account for gravity in commuting flows. Our results do not change much if we just use the occupation shares of the CD the establishment is located in. Since we focus in what follows only on 'management and R&D' and 'production', our shares do not sum to one.

Figure 2 suggests in particular that MRD may be more sensitive to agglomeration economies that exhibit a steep spatial decay—such as knowledge spillovers—whereas production may be sensitive to agglomeration economies that operate at a much broader spatial scale—such as input-output linkages. Production appears coagglomerated in more industry pairs at a spatial scale where input-output links matter more, whereas MRD appears coagglomerated in more industry pairs at a spatial scale where knowledge spillovers matter more. Our subsequent regression analysis will substantiate these informal claims more formally.

Figure 2: # of significantly coagglomerated industry pairs in 2001.




The top panel of Table 4 summarizes the shares of coagglomerated, random, and codispersed industry pairs using MRD and production employment in 2001, 2003, and 2005. Observe that production appears generally more coagglomerated than MRD, and that both decrease over time.[22] Yet, production employment seems to become less coagglomerated more quickly over that period—which witnessed the entry of China into the WTO—than management and R&D. The bottom panel of Table 4 provides summary statistics for the EGK coagglomeration measure at the CD level. As before, they are centered around zero and display some right skew in their distribution.

## 5.2 A 'differenced' perspective

In Section 4, we assumed that coagglomeration is driven uniformly by input-output linkages, labor market pooling, and knowledge spillovers across industries and occupations. We now show that the foregoing regressions mask substantial heterogeneity. To this end, we extend and complement the analysis of Faggio et al. (2017) by looking at differences across industry

[22] To make sense of the seemingly contradictory results between Figure 2 and Table 4, note that the table reports the number of coagglomerated industry pairs for all distances. There are more coagglomerated industry pairs in production across all distances, whereas there are more coagglomerated industry pairs in MRD at short distances.

Table 4: Summary statistics of coagglomeration measures.

| | DO coagglomeration measure | | | | | |
|---|---|---|---|---|---|---|
| | Management and Research | | | Production | | |
| | coagglomerated | random | codispersed | coagglomerated | random | codispersed |
| 2001 | 0.559 | 0.282 | 0.159 | 0.605 | 0.242 | 0.153 |
| 2003 | 0.485 | 0.348 | 0.167 | 0.593 | 0.274 | 0.134 |
| 2005 | 0.480 | 0.338 | 0.182 | 0.508 | 0.326 | 0.167 |
| | **EGK coagglomeration measure** | | | | | |
| | Management and Research | | | Production | | |
| | Average | Min | Max | Average | Min | Max |
| 2001 | -0.002 | -0.618 | 0.901 | 0.004 | -1.658 | 2.517 |
| 2003 | -0.001 | -0.587 | 0.965 | 0.003 | -0.840 | 1.460 |
| 2005 | -0.002 | -0.665 | 1.383 | 0.005 | -0.966 | 1.281 |

*Notes:* The top panel reports the shares of industry pairs that are statistically significantly coagglomerated, random, or codispersed based on the DO measure of coagglomeration. We assess the statistical significance of the patterns using 200 bootstrap replications and global confidence bands (see Appendix A.3 for details). The bottom panel reports the average, minimum, and maximum values for the EGK measures of coagglomeration computed at the census division level. These measures have no interpretation in terms of statistical significance.

pairs in terms of functional employment patterns. More precisely, we now look separately at the determinants of coagglomeration for management and R&D employment and for production employment. We conjecture that MRD employment is more sensitive to knowledge spillovers, whereas production employment is more sensitive to input-output linkages. Which employment type is more sensitive to labor market pooling is a priori unclear.

To begin with, we first replicate our baseline regressions using separately the coagglomeration measures constructed from the MRD employment shares and the production employment shares. Tables B.7 and B.8 in the appendix reports the results using both the DO and the EGK measures of coagglomeration. The differences between the coagglomeration of the two occupational employment types appear minor. There are no obvious large differences between both tables when using directly these coagglomeration measures in the regressions. One explanation for the absence of notable differences could be that there are indivisibilities in the colocation of employment types. To some degree, the colocation or MRD is strongly correlated with the colocation of production, and our coagglomeration measures are not able to pick up the small differences in patterns. We hence provide additional evidence using the difference in the coagglomeration of employment in MRD relative to the employment in production. A first simple way to do so is to pool the coagglomeration measures for both employment types and to interact the Marshallian covariates with a MRD dummy. Table 5 shows that MRD employment is systematically more coagglomerated than production employment. A second set of findings is that: (i) input-output linkages and labor market pooling are somewhat less important for MRD employment than for production employment; and (ii) knowledge spillovers and technological relatedness are somewhat more important for MRD employment than for production employment. Most estimates of the interaction terms are, however, imprecisely measured.

Table 5: Coagglomeration regressions (pooled, MRD and production).

| | (1) (Univar) | (2) (EGK) | (3) (Flow) | (4) (Similarity) | (5) (EGK) | (6) (Flow) | (7) (Similarity) |
|---|---|---|---|---|---|---|---|
| MRD employment dummy | 0.141$^a$ | 0.183$^a$ | 0.154$^a$ | 0.222$^a$ | 0.159$^a$ | 0.137$^a$ | 0.221$^a$ |
| | (0.015) | (0.025) | (0.018) | (0.028) | (0.037) | (0.033) | (0.041) |
| Input-output linkages | 0.084$^a$ | 0.029$^a$ | 0.047$^a$ | | 0.027$^a$ | 0.053$^a$ | |
| | (0.012) | (0.011) | (0.015) | | (0.010) | (0.014) | |
| Input-output linkages × MRD | -0.017 | -0.009 | -0.014 | | -0.015 | -0.014 | |
| | (0.017) | (0.016) | (0.021) | | (0.014) | (0.019) | |
| Input-output correlation | 0.152$^a$ | | | 0.121$^a$ | | | 0.052$^a$ |
| | (0.010) | | | (0.015) | | | (0.015) |
| Input-output correlation × MRD | -0.026$^c$ | | | -0.028 | | | -0.025 |
| | (0.013) | | | (0.020) | | | (0.020) |
| Labor flows | 0.143$^a$ | | 0.113$^a$ | | | 0.008 | |
| | (0.010) | | (0.011) | | | (0.011) | |
| Labor flows × MRD | -0.026$^c$ | | -0.022 | | | -0.023 | |
| | (0.014) | | (0.014) | | | (0.015) | |
| Occupational correlation | 0.161$^a$ | 0.139$^a$ | | 0.140$^a$ | 0.065$^a$ | | 0.064$^a$ |
| | (0.010) | (0.011) | | (0.013) | (0.011) | | (0.013) |
| Occupational correlation × MRD | -0.034$^b$ | -0.033$^b$ | | -0.031$^c$ | -0.036$^b$ | | -0.030$^c$ |
| | (0.014) | (0.014) | | (0.018) | (0.016) | | (0.018) |
| Knowledge flows | 0.023$^c$ | 0.003 | 0.023$^c$ | | -0.012 | 0.001 | |
| | (0.012) | (0.012) | (0.014) | | (0.010) | (0.012) | |
| Knowledge flows × MRD | 0.008 | 0.013 | 0.008 | | 0.012 | 0.009 | |
| | (0.018) | (0.017) | (0.020) | | (0.015) | (0.017) | |
| Technological relatedness | -0.073$^a$ | | | -0.158$^a$ | | | -0.100$^a$ |
| | (0.012) | | | (0.012) | | | (0.011) |
| Technological relatedness × MRD | 0.029$^c$ | | | 0.042$^a$ | | | 0.042$^a$ |
| | (0.016) | | | (0.016) | | | (0.015) |
| Within-plant share | | | | | 0.062$^a$ | 0.126$^a$ | 0.133$^a$ |
| | | | | | (0.010) | (0.015) | (0.015) |
| Within-plant share × MRD | | | | | 0.014 | 0.001 | -0.006 |
| | | | | | (0.015) | (0.021) | (0.021) |
| Ad valorem transport costs | | | | | -0.159$^a$ | -0.150$^a$ | -0.146$^a$ |
| | | | | | (0.009) | (0.009) | (0.009) |
| Ad valorem transport costs × MRD | | | | | -0.002 | -0.008 | -0.009 |
| | | | | | (0.012) | (0.013) | (0.013) |
| Multiplant share | | | | | -0.292$^a$ | -0.305$^a$ | -0.280$^a$ |
| | | | | | (0.009) | (0.009) | (0.009) |
| Multiplant share × MRD | | | | | 0.015 | 0.017 | 0.008 |
| | | | | | (0.012) | (0.012) | (0.013) |
| Controls $ij$ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Cohort FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 20,584 | 20,584 | 19,458 | 19,458 | 20,090 | 19,012 | 19,012 |
| $R^2$ | | 0.073 | 0.068 | 0.089 | 0.210 | 0.212 | 0.218 |

*Notes:* The dependent variable is the DO $K$-density at 15 kilometers distance. MRD is a dummy variable with value 1 for management and R&D employment and 0 for MRD employment. Results are for all unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All regressions include cohort fixed effects. Huber-White robust standard errors in parentheses. $^a$p<0.01, $^b$p<0.05, $^c$p<0.1.

We can obtain stronger evidence by using as the dependent variable not the pooled coagglomeration measures but the difference between the coagglomeration measure for MRD and for production. Formally, starting from (2) we use the relative measures

$$\frac{\widehat{K}_{ij}^{MRD}(\overline{d})}{\widehat{K}_{ij}^{TOT}(\overline{d})}, \quad \frac{\widehat{K}_{ij}^{PRO}(\overline{d})}{\widehat{K}_{ij}^{TOT}(\overline{d})}, \quad \text{and} \quad \frac{\widehat{K}_{ij}^{MRD}(\overline{d})}{\widehat{K}_{ij}^{PRD}(\overline{d})} \tag{10}$$

as our dependent variable, where $TOT$ stands for total employment. In what follows, we report these 'differenced regressions' for the DO measure only.[23] These regressions allow us to investigate whether MRD—which is likely relatively more knowledge intensive—is relatively more coagglomerated for industries with strong knowledge linkages than production employment. We believe that this approach is useful to shed more light on the role of the different agglomeration forces.

Table 6 shows our baseline differenced results. It reveals a very systematic and intuitive pattern: the knowledge flows and technological relatedness variables are relatively more important for the coagglomeration of MRD than for the coagglomeration of production (or for total employment). The reverse holds for the input-output and labor market variables, which are relatively more important for the coagglomeration of production employment than for either MRD or total employment.

Table B.9 in the appendix summarizes results from a battery of robustness checks for the determinants of the relative coagglomeration of MRD vs production. It includes regressions where we: (i) use all industries, including small ones with less than 30 plants; (ii) exclude industry pairs within the same NAICS 3-digit industries (to deal with the potential problem that some of our variables are only available at this level of industrial aggregation); (iii) use the years 2001, 2003, and 2005 separately as cross sections; (iv) use separate input- and output linkages. (v) use a make-based measure of patent citations (instead of a use-based measure) which is the maximum of the shares of patents that industries $i$ or $j$ manufacture and which originate from the other industry; (vi) use the US technological relatedness measure by Steijn et al. (2022) as an alternative to ours [24]; (vii) run IV regressions where we instrument the Canadian input-output linkages with their US counterparts; (viii) exclude the within-plant

[23] By differences we mean differences between coagglomeration measures for different business functions, not a within business functions type of analysis. As shown before, the DO and the EGK measures of coagglomeration provide a fairly similar picture when focusing on the coagglomeration of total employment. However, since the EGK measure (3) can be positive or negative (and thus has an asymptote at zero, around which that measure is centered), it is not suitable to produce 'differenced' results: taking ratios (log differences) makes no sense. Simple differences are also hard to interpret when both terms can change signs. We thus focus on the DO measure in what follows. Taking log differences or ratios of CDFs poses no problem and has a simple interpretation.

[24] See Steijn et al., 2022, for details. Their measure is constructed from the co-occurrences of technology classes of industries in patents. They argue that it is a better measure of knowledge sharing than patent citations. It is, however, unclear to us if it captures 'knowledge sharing' or some other general similarity of technologies.

share control; and (ix) use all six Marshallian covariates simultaneously.[25] Comparing tables 6 and B.9 shows that our results are robust across the different specifications.

## 5.3 Multiplant firms are different

Table 7 provides a last set of results that investigates whether the share of multiplant firms in the industries matter for the determinants of coagglomeration of MRD and production. Following Faggio et al. (2017), we split industry pairs into three categories: (i) pairs where both industries have multiplant shares above the median (high-high pairs); (ii) pairs where both industries have multiplant shares below the median (low-low pairs); and (iii) and mixed pairs, which we henceforth exclude from the analysis.

Table 7 is consistent with existing evidence that multiplant firms are very different from single-plant firms. Multiplant firms are usually large firms that have extensive internal resources that make them potentially less likely to rely on external agglomeration effects.[26] Our results show that industries with fewer multiplant firms (low-low pairs) are more coagglomerated when it comes to knowledge spillovers and occupational similarity. The $R^2$ of the low-low pairs regressions is also significantly larger, thus showing that our variables explain a larger share in the variation of coagglomeration patterns for industry pairs with smaller single-unit firms. In a nutshell, the Marshallian agglomeration forces matter more for smaller firms (e.g., Rosenthal and Strange, 2001). While industry pairs with large shares of multiunit firms (high-high pairs) do not systematically colocate to exploit Marshall's agglomeration economies, industry pairs with small shares of multiunit firms (hence also smaller firms) do colocate to exploit agglomeration economies related to labor market pooling (for production) and knowledge spillovers (for MRD and production). Labor market aspects seem to matter substantially for the production operations of smaller single-plant firms, whereas knowledge spillovers seem to matter substantially for the MRD operations of these smaller firms. Input-output links do not seem to matter for either high-high or low-low multiplant-share industry

[25] In an unreported robustness check, we used the unsmoothed (raw) shares from the census divisions to construct the geographic distribution of employment by occupation. As those shares are highly correlated with the commuting-adjusted smoothed shares from the main analysis, the kernel-smoothed DO measures are very similar and our results barely change.

[26] The literature has documented the existence of internal markets within the boundaries of multiunit firms and conglomerates. There is substantial evidence in the corporate finance literature that resource transfers within multiunit firms take place to mitigate market frictions—credit constraints or hold-up problems with relationship-specific intermediates goods suppliers. Multiunit firms also have internal capital and labor markets (e.g., Lamont, 1997; Maksimovic and Philips, 2002; Tate and Yang, 2015), which affects firm- and plant-level productivity. Strange et al. (2006, p.345) find that *"establishments that meet their needs for specialized labor externally are more likely to cluster than those that meet these needs internally through training."* This suggests that if multiunit firms have more scope for internal training, and larger internal labor markets, they should be less dependent on labor market considerations in the location choices of their plants.

Table 6: Differenced regressions.

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Production / total employment | | | | MRD / total employment | | | | MRD / production employment | | | |
| | (Univar) | (EGK) | (Flow) | (Similarity) | (Univar) | (EGK) | (Flow) | (Similarity) | (Univar) | (EGK) | (Flow) | (Similarity) |
| Input-output linkages | 0.010 | 0.007 | -0.007 | | -0.054$^a$ | -0.028$^a$ | -0.035$^a$ | | -0.044$^a$ | -0.027$^a$ | -0.035$^a$ | |
| | (0.009) | (0.008) | (0.008) | | (0.007) | (0.008) | (0.010) | | (0.008) | (0.007) | (0.009) | |
| Input-output correlation | 0.049$^a$ | | | 0.045$^a$ | -0.089$^a$ | | | -0.055$^a$ | -0.090$^a$ | | | -0.065$^a$ |
| | (0.009) | | | (0.012) | (0.008) | | | (0.012) | (0.008) | | | (0.012) |
| Labor flows | 0.039$^a$ | | 0.034$^a$ | | -0.076$^a$ | | -0.023$^b$ | | -0.073$^a$ | | -0.027$^a$ | |
| | (0.008) | | (0.010) | | (0.008) | | (0.009) | | (0.007) | | (0.009) | |
| Occupational correlation | 0.096$^a$ | 0.067$^a$ | | 0.058$^a$ | -0.091$^a$ | -0.058$^a$ | | -0.042$^a$ | -0.115$^a$ | -0.075$^a$ | | -0.058$^a$ |
| | (0.010) | (0.011) | | (0.011) | (0.008) | (0.010) | | (0.011) | (0.008) | (0.010) | | (0.011) |
| Knowledge flows | -0.008 | -0.015$^c$ | -0.013 | | 0.048$^a$ | 0.045$^a$ | 0.034$^a$ | | 0.041$^a$ | 0.041$^a$ | 0.032$^a$ | |
| | (0.009) | (0.009) | (0.009) | | (0.011) | (0.011) | (0.011) | | (0.011) | (0.011) | (0.011) | |
| Technological relatedness | -0.096$^a$ | | | -0.071$^a$ | -0.002 | | | 0.020$^c$ | 0.028$^a$ | | | 0.033$^a$ |
| | (0.013) | | | (0.014) | (0.010) | | | (0.010) | (0.010) | | | (0.011) |
| Within-plant share | | -0.013 | 0.001 | -0.004 | | 0.016$^c$ | -0.033$^a$ | -0.017 | | 0.020$^b$ | -0.027$^b$ | -0.008 |
| | | (0.010) | (0.012) | (0.012) | | (0.010) | (0.011) | (0.011) | | (0.009) | (0.011) | (0.011) |
| Ad valorem transport costs | | 0.021$^b$ | 0.035$^a$ | 0.037$^a$ | | -0.096$^a$ | -0.103$^a$ | -0.100$^a$ | | -0.106$^a$ | -0.116$^a$ | -0.112$^a$ |
| | | (0.010) | (0.010) | (0.010) | | (0.010) | (0.011) | (0.011) | | (0.010) | (0.010) | (0.010) |
| Multiplant share | | -0.263$^a$ | -0.263$^a$ | -0.246$^a$ | | 0.004 | 0.007 | -0.007 | | 0.172$^a$ | 0.176$^a$ | 0.158$^a$ |
| | | (0.016) | (0.016) | (0.016) | | (0.012) | (0.012) | (0.012) | | (0.015) | (0.015) | (0.015) |
| Controls $ij$ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ |
| Cohort FE | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ |
| Observations | 10,292 | 10,045 | 9,506 | 9,506 | 10,292 | 10,045 | 9,506 | 9,506 | 10,292 | 10,045 | 9,506 | 9,506 |
| $R^2$ | | 0.086 | 0.083 | 0.090 | | 0.080 | 0.085 | 0.087 | | 0.083 | 0.082 | 0.087 |

*Notes:* The dependent variable in columns (1)–(4) is the DO $K$-density of production relative to that of total employment at 15 kilometers distance. In columns (5)–(8) it is the DO $K$-density of MRD relative to that of total employment at 15 kilometers distance. Finally, in columns (9)–(12), it is the DO $K$-density of MRD relative to that of production employment at 15 kilometers distance. Results are for all unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. Sample sizes vary across specifications because of missing covariates. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All regressions include cohort fixed effects. Huber-White robust standard errors in parentheses. $^a p < 0.01$, $^b p < 0.05$, $^c p < 0.1$.

Table 7: Multiplant firms and coagglomeration.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Management and R&D | | Production | |
| | (High-High) | (Low-Low) | (High-High) | (Low-Low) |
| Input-output linkages | 0.002 | 0.041 | -0.000 | -0.006 |
| | (0.022) | (0.037) | (0.019) | (0.039) |
| Occupational correlation | 0.089$^a$ | 0.084$^b$ | 0.030 | 0.174$^a$ |
| | (0.028) | (0.036) | (0.024) | (0.048) |
| Knowledge flows | 0.040 | 0.214$^a$ | -0.022 | 0.283$^a$ |
| | (0.033) | (0.071) | (0.033) | (0.074) |
| Within-plant share | -0.004 | -0.000 | 0.020 | -0.026 |
| | (0.042) | (0.021) | (0.038) | (0.019) |
| Ad valorem transport costs | -0.099$^a$ | -0.140$^a$ | -0.083$^a$ | -0.216$^a$ |
| | (0.018) | (0.052) | (0.016) | (0.060) |
| Observations | 590 | 590 | 590 | 590 |
| $R^2$ | 0.246 | 0.397 | 0.201 | 0.343 |

*Notes:* The dependent variable in columns (1)–(2) is the DO $K$-density at 15 kilometers distance for MRD. The dependent variable in columns (3)–(4) is the DO $K$-density at 15 kilometers distance for production. Columns (1) and (3) report results for all industry pairs where each industry is above the median share of multiplant firms. Columns (2) and (4) report results for all industry pairs where each industry is below the median share of multiplant firms. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All regressions include cohort fixed effects. Industry-pair controls $ij$ are included. Huber-White robust standard errors in parentheses. $^a$p<0.01, $^b$p<0.05, $^c$p<0.1.

pairs, which suggests that the results for the whole sample are driven by high-low pairs.

# 6 Conclusion

We believe that—despite the criticisms they sometimes get concerning their usefulness for empirical work (e.g., Combes and Gobillon, 2015)—the coagglomeration patterns of industry pairs are a powerful tool to learn more about the determinants of agglomeration. Recent research has shown that there is substantial heterogeneity in those patterns, and that this heterogeneity can be leveraged to sharpen existing results and gain new insights.

Our paper contributes to this literature by showing how a finer slicing of the data allows to uncover robust patterns: industries that have stronger knowledge links tend to coagglomerate more their functions that are relatively knowledge intensive (management and R&D), whereas industries that have stronger input-output links tend to coagglomerate more their functions that are relatively reliant on these input-output links (production). These relationships become sharp once we take a 'differenced perspective', i.e., once we look at how much more one function is coagglomerated relative to another one (or to total employment). Uncovering these relationships requires, however, to use coagglomeration metrics that allow to be differenced in a meaningful way. The widely used Ellison-Glaeser-Kerr measure does not seem to be

very useful in that respect. We hence believe that the literature should consider more often alternative measures such as that by Duranton-Overman.

Our results further show that transport costs have a strong effect on coagglomeration patterns, thus underscoring their relevance for understanding the spatial structure of production and trade. Though this point is a key message of (new) economic geography models (e.g., Fujita et al., 2001), direct measures are rarely used and considered as 'controls' rather than of direct interest (yet, see Aleksandrova et al. 2020). Furthermore, industrial structure matters substantially, especially firms' ability to break down production across plants and to spatially separate their knowledge-intensive activities from their input-output intensive activities. While this has been emphasized in theoretical models since a long time (e.g., Fujita and Gokan, 2005), it has not really been a part of the analysis of coagglomeration patterns until now.

While our analysis provides a first shot at slicing coagglomeration patterns along functional lines, future work leveraging more detailed geocoded employment-by-workplace-and-occupation data, instead of our constructed data, would be desirable. Such data are, however, rarely available because of confidentiality issues: data at high industrial-occupational resolution are likely to have limited spatial information, whereas spatially granular data usually have limited industrial-occupational resolution. Maybe those problems could be overcome using specific confidential datasets for some countries such as those in Scandinavia. We think that this is a research agenda worth pursuing in the future.

**Statements and Declarations.** The authors have no competing financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

# References

[1] Ahlfeldt, Gabriel M., Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf. 2015. "The economics of density: Evidence from the Berlin Wall." *Econometrica* 83**(6)**: 2127–2189.

[2] Aleksandrova, Ekaterina, Kristian Behrens, and Maria Kuznetsova. 2020. "Manufacturing (co) agglomeration in a transition country: Evidence from Russia." *Journal of Regional Science* 60**(1)**: 88–128.

[3] Abdel-Rahman, Hesham, and Alex Anas. 2004. "Theories of systems of cities." In: J. Vernon Henderson and Jacques-François Thisse (eds.) *Handbook of Regional and Urban Economics*, vol. 4, Elsevier: North-Holland, pp. 2293–2339.

[4] Alcácer, Juan. 2006. "Location choices across the value chain: How activity and capability influence collocation." *Management Science* 52**(10)**: 1457–1471.

[5] Audretsch, David B., and Maryann P. Feldman. 1996. "R&D spillovers and the geography of innovation and production." *American Economic Review* 86**(3)**: 630–640.

[6] Bade, Franz-Josef, Eckhardt Bode, and Eleonora Cutrini. 2015. "Spatial fragmentation of industries by functions." *Annals of Regional Science* 54**(1)**: 215–250.

[7] Behrens, Kristian, and Théophile Bougna. 2015. "An anatomy of the geographical concentration of Canadian manufacturing industries." *Regional Science and Urban Economics* 51: 47–69.

[8] Behrens, Kristian, Mark W. Brown, and Théophile Bougna. 2018. "The world is not yet flat: Transport costs matter!" *Review of Economics and Statistics* 100**(4)**: 712–724.

[9] Behrens, Kristian, and Mark W. Brown. 2018. "Transport costs, trade, and geographic concentration: Evidence from Canada." In: Blonigen, Bruce A. and Wesley W. Wilson (eds.), *Handbook of International Trade and Transportation*. Edward Elgar Publishing, pp.188–235.

[10] Behrens, Kristian, and Rachel Guillain. 2017. "The determinants of coagglomeration: Evidence from functional employment patterns." CEPR Discussion Paper # DP11884.

[11] Behrens, Kristian, and Frédéric Robert-Nicoud. 2015. "Agglomeration theory with heterogeneous agents." In: Duranton, Gilles, J. Vernon Henderson, and William C. Strange (eds.) *Handbook of Regional and Urban Economics, vol. 5*. North-Holland: Elsevier B.V., pp. 171–245.

[12] Bernard, Andrew. B., J. Bradford Jensen. 1995. "Exporters, jobs, and wages in US manufacturing: 1976-1987." *Brookings Papers on Economic Activity. Microeconomics*: pp. 67-119.

[13] Billings Stephan B., and Erik B. Johnson. 2016. "Agglomeration within an urban area." *Journal of Urban Economics*: pp. 13-25.

[14] Combes, Pierre-Philippe, and Laurent Gobillon. 2015. "The empirics of agglomeration economies." In: Duranton, Gilles, J. Vernon Henderson, and William C. Strange (eds.), *Handbook of Regional and Urban Economics, vol.5A*. North-Holland: Elsevier B.V., pp. 247–341.

[15] Davis, Donald R. and Jonathan I. Dingel. 2020. "The comparative advantage of cities." *Journal of International Economics*123: 103291

[16] Diodato, Dario, Frank Neffke, and Neave O'Clery. 2018. "Why do industries coagglomerate? How Marshallian externalities differ by industry and have evolved over time." *Journal of Urban Economics* 106: 1–26.

[17] Duranton, Gilles, and Henry G. Overman. 2008. "Exploring the detailed location patterns of U.K. manufacturing industries using microgeographic data." *Journal of Regional Science* 48**(1)**: 213–243.

[18] Duranton, Gilles, and Henry G. Overman. 2005. "Testing for localization using microgeographic data." *Review of Economic Studies* 72**(4)**: 1077–1106.

[19] Duranton, Gilles, and Diego Puga. 2005. "From sectoral to functional urban specialisation." *Journal of Urban Economics* 57**(2)**: 343–370.

[20] Duranton, Gilles, and Diego Puga. 2004. "Micro-foundations of urban agglomeration economies." In: J. Vernon Henderson and Jacques-François Thisse (eds.) *Handbook of Regional and Urban Economics*, vol. 4, Elsevier: North-Holland, pp. 2063–2117.

[21] Duranton, Gilles, and Diego Puga. 2001. "Nursery cities: Urban diversity, process innovation, and the life cycle of products." *American Economic Review* 91**(5)**: 1454–1477.

[22] Ellison, Glenn D., and Edward L. Glaeser. 1999. "The geographic concentration of industry: Does natural advantage explain agglomeration?" *American Economic Review* 89**(2)**: 311–316.

[23] Ellison, Glenn D., Edward L. Glaeser, and William R. Kerr. 2010. "What causes industry agglomeration? Evidence from coagglomeration patterns." *American Economic Review* 100**(3)**: 1195–1213.

[24] Faggio, Giulia, Olmo Silva, and William C. Strange. 2017. "Heterogeneous agglomeration." *Review of Economics and Statistics* 99**(1)**: 80–94.

[25] Faggio, Giulia, Olmo Silva, and William C. Strange. 2020. "Tales of the city: what do agglomeration cases tell us about agglomeration in general?" *Journal of Economic Geography* 20**(5)**, 1117–1143.

[26] Fallick, Bruce, Charles A. Fleischman, and James B. Rebitzer. 2006. "Job-hopping in Silicon-Valley: Some evidence concerning the microfoundations of a high-technology cluster." *The Review of Economics and Statistics* 88**(3)**: 471–481.

[27] Fujita, Masahisa, and Jacques-François Thisse. 2002. *Economics of Agglomeration: Cities, Industrial Location, and Regional Growth.* Cambridge University Press: Cambridge, MA.

[28] Gabe, Todd M., and Jaison R. Abel. 2016. "Shared knowledge and the coagglomeration of occupations." *Regional Studies* 50**(8)**: 1360–1373.

[29] Gabe, Todd M., and Jaison R. Abel. 2012. "Specialized knowledge and the geographic concentration of occupations." *Journal of Economic Geography* 12**(2)**: 435–453.

[30] Fujita, Masahisa, and Toshitaka Gokan. 2005. "On the evolution of the spatial economy with multi-unit multi-plant firms: the impact of IT development." *Portuguese Economic Journal* 4: 73–105.

[31] Fujita, Masahisa, Paul R. Krugman, and Anthony J. Venables. 2001. *The spatial economy: Cities, regions, and international trade.* Cambridge, MA: MIT Press.

[32] Helsley, Robert W., and William C. Strange. 2014. "Coagglomeration, clusters, and the scale and composition of cities." *Journal of Political Economy* 122**(5)**: 1064–1093.

[33] Howard Emma, Carol Newman, and Finn Tarp. 2016. "Measuring industry coagglomeration and identifying the driving forces." *Journal of Economic Geography* 16**(5)**: 1055–1078.

[34] Jofre-Monseny, Jordi, Raquel Marín-López, and Elisabet Viladecans-Marsal. 2011. "The mechanisms of agglomeration: Evidence from the effect of inter-industry relations on the location of new firms." *Journal of Urban Economics* 70**(2-3)**: 61–74.

[35] Kerr, William R. 2008. "Ethnic scientific communities and international technology diffusion." *Review of Economics and Statistics* 90**(3)**: 518–537.

[36] Kolko, Jed. 2010. "Urbanization, agglomeration, and the coagglomeration of service industries." In: Glaeser, Edward L. (ed.), *Agglomeration Economics*. NBER Books, University of Chicago Press, pp. 151–180.

[37] Madrian, Brigitte C., and Lars John Lefgren. 1999. "A note on longitudinally matching Current Population Survey (CPS) respondents." NBER Technical Working Paper #247, National Bureau of Economic Research, MA. Available online at `http://www.nber.org/papers/T0247`.

[38] Mori, Tomoya, Koji Nishikimi, and Tony E. Smith. 2008. "The Number-Average Size Rule: A New Empirical Relationship Between Industrial Location and City Size." *Journal of Regional Science* 48(1): 165–211.

[39] Rigby, David L., and W. Mark Brown. 2015. "Who benefits from agglomeration?" *Regional Studies* 49**(1)**: 28–43.

[40] Rosenthal, Stuart S., and William C. Strange. 2010. "Small establishments/big effects: Agglomeration, industrial organization and entrepreneurship." In: Edward L. Glaeser (ed.), *Agglomeration Economics* (NBER Books): University of Chicago Press, pp. 277–302.

[41] Lamont, Owen. 1997. "Cash flow and investment: Evidence from internal capital markets." *Journal of Finance* LII(1): 83–109.

[42] Maksimovic, Vojislav, and Gordon Phillips. 2002. "Do conglomerate firms allocate resources inefficiently across industries? Theory and evidence." *Journal of Finance* LVII(2), 721–767.

[43] Shearmur, Richard, and Mario Polèse. 2005. "Diversity and employment growth in Canada, 1971–2001: can diversification policies succeed?" *Canadian Geographer/Le Géographe canadien* 49**(3)**: 272-290.

[44] Rosenthal, Stuart S., and William C. Strange. 2003. "Geography, industrial organization, and agglomeration." *Review of Economics and Statistics* 85**(2)**: 377–393.

[45] Rosenthal, Stuart S., and William C. Strange. 2001. "The determinants of agglomeration." *Journal of Urban Economics* 50**(2)**: 191–229.

[46] Scherer, Frederic. 1984. "Using linked patent and R&D data to measure interindustry technology flows. R&D, patents, and productivity." University of Chicago Press: pp. 417–464.

[47] Scholl, Tobias, and Thomas Brenner. 2015. "Optimizing distance-based methods for large data sets." *Journal of Geographical Systems* 17**(4)**: 333–351.

[48] Steijn, Mathieu P.A., Hans R.A. Koster, and Frank G. Van Oort. 2022. "The dynamics of industry agglomeration: Evidence from 44 years of coagglomeration patterns." *Journal of Urban Economics* 120: 103456.

[49] Strange, William C., Walid Hejazi, and Jianmin Tang. 2006. "The uncertain city: Competitive instability, skills, innovation and the strategy of agglomeration." *Journal of Urban Economics* 59**(3)**: 331–351.

[50] Tate, Geoffrey, and Liu Yang. 2015. "The bright side of corporate diversification: Evidence from internal labor markets." *Review of Financial Studies* 28**(8)**: 2203–2249.

# Appendix

## A Data

### A.1 Plant-level data

Our plant-level data come from the *Scott's National All* database. This establishment-level database builds on the Business Register and contains information on plants operating in Canada. It contains about 47,000–50,000 manufacturing plants per year and extensively covers all manufacturing industries. Although the Scott's dataset is only a large sample and not the universe of manufacturing plants, it has a very wide (85–90%) and representative coverage. It contains almost all of the large plants and many small plants. Behrens and Bougna (2015, Appendix A) provide detailed information on the data quality and its representativeness—both in terms of provinces and industries—of the manufacturing portion of the database.

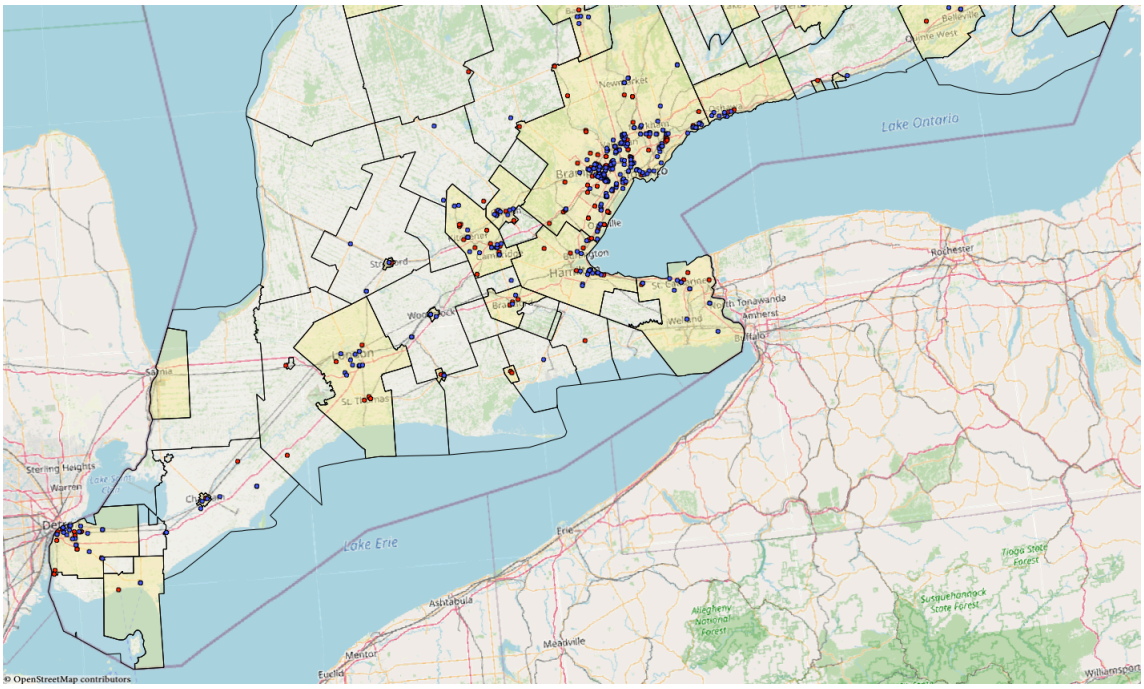
We use the Scott's data for the years 2001, 2003, and 2005. For every establishment, we have information on its primary 6-digit NAICS code, up to four secondary NAICS codes, its total employment, and its 6-digit postal code. To mitigate the role of outliers and erroneous information, we trim the top 0.5% of plants in terms of employment by industry. The *Scott's All* database unfortunately does not provide firm identifiers (only plant identifiers). Yet, it reports information on the legal name of the entity that owns the plant, and we use this information to group plants into firms. The shares of multiunit plants we obtain for the different industries are similar to those of the manufacturing portion of the Business Register or the Annual Survey of Manufacturers. We cross-checked our measure against aggregate measures from Statistics Canada—which we ordered as special tabulations from confidential data—and the correlations are high and in the range of 0.8 to 0.9.

### A.2 Geocoding plants based on postal codes

All plants provide address information. We geocode plants by latitude and longitude using their 6-digit postal code centroids obtained from *Statistics Canada*'s Postal Code Conversion Files (PCCF). We use the postal code data for the next year in order to take into account that there is a six months delay in the updating of postal codes. For example, the census geography of 1996 and the postal codes as of May 2002 (818,907 unique postal codes) were associated with the 2001 Scott's data, whereas we matched the 2003 and 2005 Scott's data with the 2001 census geography and the corresponding PCCF's. We also associate standard geographical identifiers of the postal code's census division (CD) and census metropolital area (CMA) with each plant: the 1996 census with the 2001 Scott's data, and the 2001 census with the 2003 and 2005 data. We then assign plants to census divisions based on their postal code centroid latitude and

longitude coordinates and compute the employment splits.

Figure A.1: Coagglomeration of NAICS 3361 and 3363 in south-western Ontario.




*Notes:* Locations of plants in NAICS 3361 (Motor vehicle manufacturing; red dots) and 3363 (Motor vehicle parts manufacturing; blue dots). The yellow shaded CDs are either census metropolitan areas or the urban parts of the CDs reported in the special tabulations.

Figure A.1 illustrates the granularity of our geographic data for the case of the colocation of plants in 'Motor vehicle manufacturing' (NAICS 3361; red dots) and 'Motor vehicle parts manufacturing' (NAICS 3363; blue dots) in south-western Ontario.

## A.3 Significance of the DO coagglomeration metric

To measure the statistical significance of localization, we follow Duranton and Overman (2005, 2008) and compare the $K$-density estimate to some appropriately defined confidence band. We construct (global) confidence bands $[\underline{K}_{ij}(d), \overline{K}_{ij}(d)]$ at each distance $d$, where $\underline{K}_{ij}(d)$ and $\overline{K}_{ij}(d)$ denote the lower- and upper bounds. The confidence band is obtained by simulating 200 counterfactual industry distributions. These are generated by randomly reshuffling the plants in industries $i$ and $j$ across all locations occupied by plants in either industry $i$ or industry $j$ and then computing the associated counterfactual kernel densities. We then take the band containing 90% of the counterfactual densities.[27] The statistical significance can be assessed using the excess of agglomeration or dispersion with respect to the confidence band:

$$\alpha_{ij}(d) = \max\left\{\widehat{K}_{ij}(d) - \overline{K}_{ij}(d), 0\right\} \quad \text{or} \quad \psi_{ij}(d) = \max\left\{0, \underline{K}_{ij}(d) - \widehat{K}_{ij}(d)\right\}. \tag{A.1}$$

[27] This procedure controls for the geographic concentration of the two industries and looks at how much closer pairs of plants in those two industries are conditional on the overall agglomeration patterns of the industries.

Using (A.1), a pair $ij$ is said to be significantly coagglomerated at distance $d$ if $\alpha_{ij}(d) > 0$, whereas it is said to be significantly codispersed at distance $d$ if $\psi_{ij}(d) > 0$. The pair is located 'as good as random' at distance $d$ if $\alpha_{ij}(d) = \psi_{ij}(d) = 0$.

## A.4 Flow- and similarity based Marshallian covariates

**Input-output linkages.** We use detailed input-output matrices from Statistics Canada for the years 1998, 2000, and 2002, which we pair with our coagglomeration measures in 2001, 2003, and 2005. These matrices are constructed using the finest public release of the Canadian input-output tables at the $L$-level (link level), which is between NAICS 3- and 4-digit. Following Behrens and Bougna (2015), we first disaggregate the input-output matrices to the $W$-level (NAICS 6-digit) using sales or employment data as sectoral weights, and then reaggregate them to the 4-digit level.[28] Let $\omega_{ij}^{\text{in}}$ denote the share of inputs sourced by industry $i$ from industry $j$. Conversely, let $\omega_{ij}^{\text{out}}$ denote the share of output sold by industry $i$ to industry $j$. These shares are computed taking into account all industries (including primary industries and services, but excluding private consumption and the different government aggregates and imports/exports).

To address potential endogeneity issues associated with input-ouput links, we follow Ellison *et al.* (2010) and construct instruments based on the US input-output benchmark tables from the Bureau of Economic Analysis (BEA). Using the detailed 6-digit BEA tables for 1997 and 2002, we construct the same input-output shares as explained above, using US data. We again work with the whole input-output tables, including services and primary industries and excluding private consumption, government aggregates, and imports/exports. We aggregate the data to the 4-digit level, which is perfectly comparable to the Canadian NAICS that we use.

**Worker flows.** As a flow-based measure of labor market pooling, we compute an index of labor mobility across manufacturing industries. To do so, we use the 2000–2005 annual public use files of the Current Population Survey (MORG, March supplement). Using the methodology detailed in Madrian and Lefgren (1999), we transform this into a panel from which we can trace year-by-year worker movements between manufacturing industries. We extract all moves from the database (12,269 moves between manufacturing industries), and we construct a matrix that contains the share of moves from industry $i$ to industry $j$, $\text{mov}_{ij}$. We consider that industries with a larger value of $\text{mov}_{ij}$ are more similar in terms of their labor requirements. Note that because of sample size limitations, we cannot compute a time-varying measure of labor movements. Hence, we use the same values of $\text{mov}_{ij}$ across the three years of our geographic data.

[28] Due to confidentiality reasons, we cannot directly use the $W$-level matrices that are internally available at *Statistics Canada*. However, tests we ran in Behrens *et al.* (2018) using those matrices yielded similar results to those using the matrices constructed by our methodology.

**Knowledge flows.** Last, we construct proxies for knowledge sharing following previous work by Kerr (2008) that uses the NBER Patent Citation database. Our proxy for knowledge flows is based on the 'use-based' measure, which is the maximum of the shares of patents that industries $i$ or $j$ use and which originate from the other industry. We also construct a 'make-based' measure, which we will use in robustness checks. Since this measure is based on US data we do not instrument it in our regressions.

**Input-output correlation.** We construct a measure of input-output similarity from the input-output tables explained before. To do so, we keep only the manufacturing portion and compute, for each industry $i$, the share of inputs it buys from industry $j$ and the share of its outputs it sells to industry $j$. We then correlate the vector of NAICS-4 industry shares of industries $i$ and $j$ to obtain our similarity measure.

**Occupational correlation.** We construct a measure of occupational employment similarity of the workforce in the different industries. To this end, we use Occupational Employment Survey (OES) data from the Bureau of Labor Statistics (BLS) for 2002, 2003, and, 2005 to compute the share of each of 554 occupations in each 4-digit NAICS industry. There are 808 occupations in total in the OES data. We only use occupations for which there is at least some employment in manufacturing (e.g., there are no 'Surgeons' in manufacturing industries, hence we exclude them completely from our data). We use 2002 data for the 2001 plant sample, and then data for each year $t$ for the plant sample in year $t$. Using 2002 as the starting year for the OES data allows us to avoid the difficult concordance from SITC to NAICS. Our measure of occupational employment similarity is finally computed as the correlation between the vectors of occupational shares of industries $i$ and $j$. Since this measure is based on US data we do not instrument it in our regressions.

**Technological relatedness.** Several measures of technological relatedness have been proposed and used in the literature (e.g., Audretsch and Feldman, 1996; Steijn et al., 2022). Our measure is based on detailed data from the Canadian Patent Office (CPO) and developed by the C.D. Howe Institute. For each registered manufacturing patent, the dataset provides a probabilistic concordance that assigns the patent to a 3-digit NAICS industry. Hence, we can compute, at the patent level, the correlation between the vectors of probabilities for industries $i$ and $j$. This captures the correlation in the likelihood that the industries use the same patents.

Table A.1 provides descriptive statistics for our Marshallian covariates. It also summarizes the data sources and shows what time variation is in our data. Table A.2 provides the correlations between our covariates.

Table A.1: Details and summary statistics for our Marshallian covariates.

| Variable | Type | Industry detail | Time variation | Data source | Mean | S.D. | Min. | Max. | $N$ |
|---|---|---|---|---|---|---|---|---|---|
| Input shares | Flow | NAICS 4-digit | Yes | Canada, Statcan | 0.011 | 0.029 | 0.000 | 0.630 | 10,965 |
| Output shares | Flow | NAICS 4-digit | Yes | Canada, Statcan | 0.009 | 0.029 | 0.000 | 0.806 | 10,965 |
| Worker flows | Flow | NAICS 3-digit | No | US, IPUMS MORG | 0.058 | 0.074 | 0.000 | 0.456 | 210 |
| Knowledge flows | Flow | NAICS 4-digit | No | US PO, Kerr (2008) | 0.018 | 0.047 | 0.000 | 0.798 | 3,655 |
| Input-output correlation | Similarity | NAICS 4-digit | Yes | Canada, Statcan | 0.165 | 0.312 | -0.201 | 1.000 | 10,965 |
| Occupational correlation | Similarity | NAICS 4-digit | Partial | US, BLS | 0.317 | 0.210 | 0.006 | 0.993 | 10,965 |
| Technological relatedness | Similarity | NAICS 3-digit | Yes | Canada, PO | 0.011 | 0.037 | 0.000 | 0.937 | 630 |

*Notes:* Descriptive statistics based on all industry pairs $ij$ for which we can compute our measures. Sample sizes vary because of years and industry detail, as indicated in the table.

Table A.2: Pairwise correlations between our Marshallian covariates.

| | Input-output linkages | Input-output corr. | Labor flows | Occupational corr. | Knowledge flows |
|---|---|---|---|---|---|
| Input-output correlation | 0.374 | | | | |
| Labor flows | 0.330 | 0.699 | | | |
| Occupational correlation | 0.212 | 0.546 | 0.450 | | |
| Knowledge flows | 0.034 | 0.073 | 0.049 | 0.074 | |
| Technological relatedness | 0.251 | 0.303 | 0.494 | 0.184 | 0.006 |

*Notes:* We report the correlation coefficients for the standardized covariates that we use in the regression analysis.

## A.5 Control variables

**Dissimilarity controls.** Following Faggio et al. (2017), we construct controls related to: (i) the pair's dissimilarity in terms of the input shares they buy from and the output shares they sell to primary industries; and (ii) the pair's dissimilarity in terms of the input shares they buy from and the output shares they sell to business services industries. For (i), we take all primary industries (NAICS 11–21) and construct the measure as one half of the sum of the absolute value of the differences in the industries' input (or output) shares with those primary industries. For (ii), we construct the same measure, but using all business service industries (NAICS 51–55).

**Share of multiunit firms.** We group establishments by the legal names of the firms they belong to. All legal names with more than one establishments are considered as being multiunit firms. We then compute, for each industry $i$, the share of establishments that belong to multiunit firms, $\text{multi}_i$. For the pair $ij$, our measure is then the maximum given by $\text{multiunit share}_{ij} = \max\{\text{multi}_i, \text{multi}_j\}$.

**Transport costs.** We use the NAICS 4-digit ad valorem transport costs estimtes from Behrens et al. (2018) and Behrens and Brown (2018). These authors have estimated trucking transport costs for all industries in Canada from confidential commodity flow survey micro data. Note that these measures vary across time.

**Within-plant coagglomeration.** The Scott's database reports one primary industry code for each plant, which corresponds to its 'main line of business'. Contrary to most other datasets, the Scott's data also report up to four secondary industry codes, which corresponds to additional activity of the plant that is not in the same industry than the main line of business. We can thus, for each industry, compute the share of plants that have secondary activities in industry $j$, $\text{secondary}_{ij}$. Our measure of within-plant coagglomeration, $\text{within}_{ij}$. is then given by $\text{within}_{ij} = \max\{\text{secondary}_{ij}, \text{secondary}_{ji}\}$.

## A.6 Occupational types

The special census tabulations report six aggregate functional employment types, which are based on the 1991 Standard Occupational Classification (SOC): 'Managers, directors, and related occupations' (type 1); 'Natural sciences, engineering, mathematics, social sciences' (type 2); 'Religion, education, health care, arts, recreation' (type 3); 'Administration and related activities' (type 4); 'Retail and services' (type 5); and 'Primary Industry, Processing, Manufacturing and Utilities, Trades and Transport' (type 6).

Table A.3: Functional employment categories.

| Occupation title, special tabulations | Our classification | 1991 SOC categories |
|---|---|---|
| All occupations | Total employment | All |
| Managers, directors and related occupations | Management and R&D | A, B0, B1, B3 |
| Natural sciences, engineering, mathematics, social sciences | Management and R&D | C, E0, E211, E212, E213 |
| Primary Industry, Processing, Manufacturing and Utilities, Trades and Transport | Production | H, I, J |
| Religion, education, health care, arts, recreation | (excluded) | D, E1, E214, E215, E216, F |
| Administration and related activities | (excluded) | B2, B4, B5 |
| Retail and services | (excluded) | G |

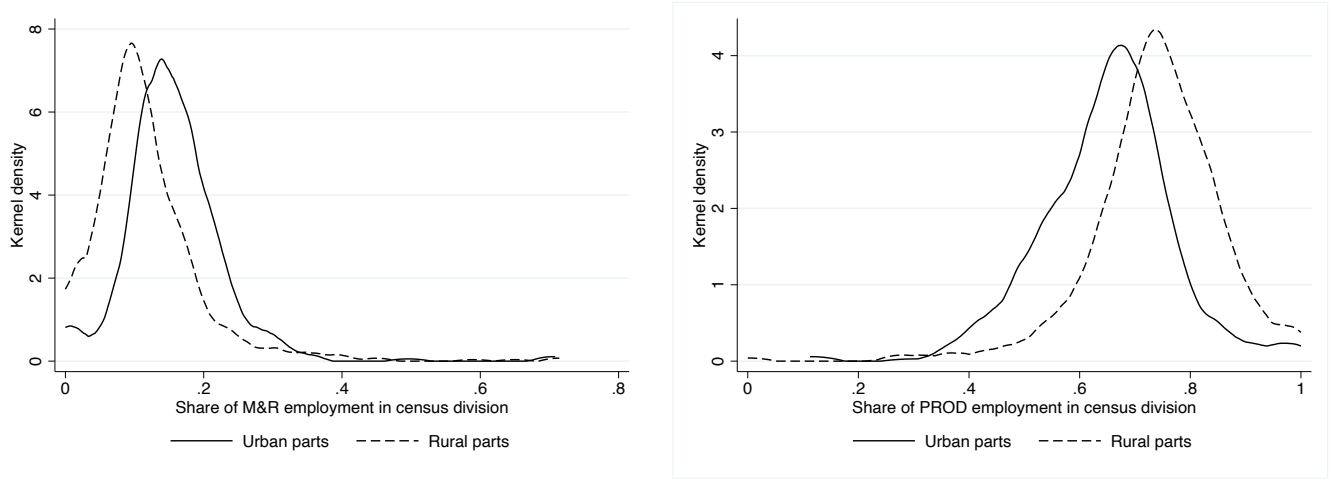
*Notes:* Relationship between the 1991 SOC classification and our functional employment categories. See Polèse and Shearmur (2005) for additional details on the data. We exclude the grayed categories from our analysis. (1) What we call 'Production' is: 1991 SOC H, I and J (Trades, Transport and Equipment Operators; Occupations Unique to Primary Industry, Processing, Manufacturing, and Utilities). (2) What we call MRD is: 1991 SOC A, B0, B1, B3, C, E0, E211, E212, E213 (Managers, directors and related occupations; Natural sciences, engineering, mathematics, social sciences).

We retain three broad functional employment types: 'Management and research' (MRD; sum of types 1 and 2); 'Production' (PROD; type 6); and 'total employment'(type 0), which is the sum of all types. Table A.3 summarizes the categories and shows their relations with the SOC classification. Each type is reported by industry and by census division. Concerning industries, the special tabulations split employment by functions at an intermediate level between the 3- and the 4-digit NAICS. We create a concordance that associates each 4-digit NAICS code with an industry code from the special tabulations. We focus on the 86 manufacturing industries only, because we have detailed microgeographic data for those industries (see below). Data are for the 1996 and the 2001 censuses. Geographic units are time consistent.

## A.7 Rural-urban split

One of the key aspects of the special census tabulations is that they split census divisions into rural and urban parts.[29] Census divisions are indeed administrative constructs that are not clearly linked to any urban or rural divide. This poses problems when using such administrative units to work on functional specialization because urban and rural areas, as well as cities of different sizes, differ substantially in the functions they perform. Figure A.2 illustrates the functional specialization patterns across the urban and rural parts of census divisions. As is the case for other countries—e.g., the US or France—there is substantial functional specialization in Canada, with urban areas having more management and research and rural areas more production.

Figure A.2: Rural urban distributions of functional employment shares.




*Notes:* We show the share of employment in MRD and in PROD aggregated across our plants for the different census divisions in Canada. We separate the urban from the rural parts using the special census tabulations. The figures pool the years 2001, 2003, and 2005, but year-by-year figures look similar.

## A.8 Allocating shares to plants

We compute the shares as follows. First, we allocate to each postal code in each special tabulation census division the division-wide share of employment type $o$ in industry $i$. Then, we put a 25 kilometers buffer around each establishment. We choose 25 kilometers since 90% of Canadians commute less than 25 kilometers to work (Statistics Canada, 2006; catalogue number `97-561-XCB2006010`). Within each buffer, we compute the mean share of employment of

[29] There are 232 census divisions in our data, 114 of them being fully rural. The remaining 118 census divisions are split into their census metropolitan and rural remainder parts. For 12 census divisions that are considered 'rural' by *Statistics Canada*, the 'urban core' parts are reported separately (e.g., Bracebridge, Ontario).

type $o$ in industry $i$ across all postal codes. Observe that our share measures vary along two dimensions. First, since our special tabulations split census divisions into rural and urban parts, two plants with the same industry-year-census division can have different employment breakdowns if one of them is located in the rural part and the other is located in the urban part of the census division. Second, even within the same census division, plants will have different shares depending on the nearby census divisions that intersect with the 25 kilometers buffers around the plants.

# B Additional tables and results

This appendix provides additional results. It contains the following tables:

- Table B.4 shows robustness results for the within-plant metric that complement the results in Table 1.
- Tables B.5 and B.6 show robustness results that complement the results for the coagglomeration using total employment in Table 2.
- Tables B.7 and B.8 show results of our basic coagglomeration regressions when using separately MRD or production employment.
- Table B.9 shows results from a battery of robustness checks for the determinants of the relative coagglomeration of MRD vs production.

Table B.4: Within-establishment coagglomeration (robustness).

| | (1) (EGK) | (2) (EGK) | (3) (EGK) | (4) (EGK) | (5) (EGK) | (6) (Flow) | (7) (EGK) | (8) (Flow) | (9) (Simi) | (10) (EGK) | (11) (Flow) | (12) (All) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Input-output linkages | 0.342$^a$ | 0.257$^a$ | 0.313$^a$ | 0.348$^a$ | | | 0.337$^a$ | 0.201$^a$ | | 0.535$^a$ | 0.412$^a$ | 0.163$^a$ |
| | (0.021) | (0.019) | (0.033) | (0.040) | | | (0.021) | (0.017) | | (0.050) | (0.043) | (0.018) |
| Input linkages | | | | | 0.261$^a$ | 0.181$^a$ | | | | | | |
| | | | | | (0.027) | (0.023) | | | | | | |
| Output linkages | | | | | 0.143$^a$ | 0.102$^a$ | | | | | | |
| | | | | | (0.026) | (0.014) | | | | | | |
| Input-output correlation | | | | | | | | | 0.428$^a$ | | | 0.122$^a$ |
| | | | | | | | | | (0.017) | | | (0.016) |
| Labor flows | | | | | | 0.279$^a$ | | 0.292$^a$ | | | 0.241$^a$ | 0.167$^a$ |
| | | | | | | (0.015) | | (0.015) | | | (0.017) | (0.018) |
| Occupational correlation | 0.343$^a$ | 0.246$^a$ | 0.330$^a$ | 0.357$^a$ | 0.346$^a$ | | 0.353$^a$ | | 0.200$^a$ | 0.300$^a$ | | 0.132$^a$ |
| | (0.013) | (0.011) | (0.022) | (0.024) | (0.014) | | (0.014) | | (0.014) | (0.017) | | (0.011) |
| Knowledge flows | 0.066$^a$ | 0.042$^a$ | 0.066$^a$ | 0.063$^a$ | 0.059$^a$ | 0.043$^a$ | | | | 0.053$^a$ | 0.043$^a$ | 0.035$^a$ |
| | (0.010) | (0.007) | (0.017) | (0.018) | (0.010) | (0.007) | | | | (0.010) | (0.007) | (0.006) |
| Knowledge flows (make) | | | | | | | 0.085$^a$ | 0.064$^a$ | | | | |
| | | | | | | | (0.012) | (0.009) | | | | |
| US technological relatedness | | | | | | | | | 0.171$^a$ | | | |
| | | | | | | | | | (0.030) | | | |
| Technological relatedness | | | | | | | | | | | | 0.040$^a$ |
| | | | | | | | | | | | | (0.009) |
| Ad valorem transport costs | 0.003 | -0.022$^a$ | 0.012 | -0.023 | -0.009 | 0.008 | -0.005 | 0.012$^c$ | -0.024$^b$ | 0.007 | 0.013$^c$ | -0.019$^b$ |
| | (0.011) | (0.008) | (0.020) | (0.020) | (0.011) | (0.007) | (0.012) | (0.007) | (0.011) | (0.012) | (0.007) | (0.008) |
| Multiplant share | -0.011 | -0.009 | -0.023 | 0.020 | -0.001 | -0.041$^a$ | 0.001 | -0.039$^a$ | 0.020$^c$ | -0.021$^c$ | -0.050$^a$ | -0.010 |
| | (0.009) | (0.008) | (0.017) | (0.017) | (0.010) | (0.008) | (0.010) | (0.008) | (0.010) | (0.011) | (0.008) | (0.008) |
| Observations | 10,710 | 9,506 | 3,403 | 3,321 | 10,045 | 9,506 | 10,045 | 9,506 | 10,045 | 10,045 | 9,506 | 9,506 |
| R-squared | 0.371 | 0.262 | 0.369 | 0.385 | 0.393 | 0.311 | 0.377 | 0.299 | 0.399 | 0.340 | 0.255 | 0.335 |

*Notes:* The dependent variables is the share of secondary NAICS codes in industry $j$ reported by plants in industry $i$. Results are for all unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All regressions include cohort fixed effects and the industry-pair $ij$ controls. Column (1) includes all small industries (less than 30 plants). Column (2) excludes NAICS 4-digit pairs in the same NAICS 3-digit industry. Columns (3) and (4) show cross-sectional results in 2001 and 2003. Columns (5) and (6) use separate input and output linkages. Column (7) and (8) use make-based measures of patent citation flows. Column (9) uses US technological relatedness from Steijn et al. (2022). Columns (10) and (11) instrument the input-output variable with its US counterpart. Last, column (12) includes all six Marshallian covariates (all flow- and similarity based measures). Huber-White robust standard errors in parentheses. $^a p < 0.01$, $^b p < 0.05$, $^c p < 0.1$.

Table B.5: Coagglomeration regressions for total employment, DO measure (robustness).

| | (1) (EGK) | (2) (EGK) | (3) (EGK) | (4) (EGK) | (5) (EGK) | (6) (Flow) | (7) (EGK) | (8) (Flow) | (9) (Similarity) | (10) (EGK) | (11) (Flow) | (12) (EGK) | (13) (All) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Input-output linkages | 0.031$^a$ | 0.052$^a$ | 0.013 | 0.043$^b$ | | | 0.026$^b$ | 0.059$^a$ | | 0.048$^b$ | 0.090$^a$ | 0.049$^a$ | 0.067$^a$ |
| | (0.011) | (0.014) | (0.018) | (0.019) | | | (0.011) | (0.015) | | (0.020) | (0.035) | (0.010) | (0.014) |
| Input linkages | | | | | -0.035$^a$ | -0.049$^a$ | | | | | | | |
| | | | | | (0.013) | (0.017) | | | | | | | |
| Output linkages | | | | | 0.062$^a$ | 0.115$^a$ | | | | | | | |
| | | | | | (0.014) | (0.015) | | | | | | | |
| Input-output correlation | | | | | | | | | 0.002 | | | | 0.024 |
| | | | | | | | | | (0.014) | | | | (0.019) |
| Labor flows | | | | | | 0.003 | | -0.006 | | | -0.011 | | 0.010 |
| | | | | | | (0.011) | | (0.011) | | | (0.012) | | (0.015) |
| Occupational correlation | 0.052$^a$ | 0.053$^a$ | 0.053$^a$ | 0.045$^b$ | 0.049$^a$ | | 0.047$^a$ | | 0.049$^a$ | 0.047$^a$ | | 0.073$^a$ | 0.051$^a$ |
| | (0.011) | (0.012) | (0.019) | (0.020) | (0.011) | | (0.011) | | (0.013) | (0.011) | | (0.010) | (0.013) |
| Knowledge flows | -0.007 | -0.000 | -0.004 | -0.011 | -0.009 | 0.004 | | | | -0.011 | 0.003 | -0.006 | -0.002 |
| | (0.010) | (0.012) | (0.019) | (0.018) | (0.010) | (0.012) | | | | (0.010) | (0.012) | (0.010) | (0.012) |
| Knowledge flows (make) | | | | | | | 0.022$^b$ | 0.030$^b$ | | | | | |
| | | | | | | | (0.011) | (0.012) | | | | | |
| Technological relatedness | | | | | | | | | | | | | -0.086$^a$ |
| | | | | | | | | | | | | | (0.011) |
| US technological relatedness | | | | | | | | | 0.024$^b$ | | | | |
| | | | | | | | | | (0.012) | | | | |
| Within-plant share | 0.074$^a$ | 0.107$^a$ | 0.078$^a$ | 0.069$^a$ | 0.069$^a$ | 0.133$^a$ | 0.063$^a$ | 0.127$^a$ | 0.069$^a$ | 0.058$^a$ | 0.124$^a$ | | 0.119$^a$ |
| | (0.011) | (0.014) | (0.020) | (0.019) | (0.011) | (0.015) | (0.011) | (0.015) | (0.011) | (0.012) | (0.016) | | (0.015) |
| Ad valorem transport costs | -0.158$^a$ | -0.160$^a$ | -0.169$^a$ | -0.152$^a$ | -0.157$^a$ | -0.151$^a$ | -0.157$^a$ | -0.152$^a$ | -0.160$^a$ | -0.159$^a$ | -0.153$^a$ | -0.160$^a$ | -0.148$^a$ |
| | (0.008) | (0.009) | (0.015) | (0.015) | (0.009) | (0.009) | (0.009) | (0.009) | (0.008) | (0.009) | (0.009) | (0.008) | (0.009) |
| Multiplant share | -0.262$^a$ | -0.274$^a$ | -0.245$^a$ | -0.260$^a$ | -0.274$^a$ | -0.287$^a$ | -0.276$^a$ | -0.288$^a$ | -0.273$^a$ | -0.277$^a$ | -0.289$^a$ | -0.274$^a$ | -0.268$^a$ |
| | (0.009) | (0.009) | (0.018) | (0.016) | (0.009) | (0.009) | (0.009) | (0.009) | (0.009) | (0.009) | (0.009) | (0.009) | (0.009) |
| Observations | 10,710 | 9,506 | 3,403 | 3,321 | 10,045 | 9,506 | 10,045 | 9,506 | 10,045 | 10,045 | 9,506 | 10,045 | 9,506 |
| $R^2$ | 0.195 | 0.211 | 0.210 | 0.205 | 0.208 | 0.212 | 0.206 | 0.209 | 0.206 | 0.206 | 0.208 | 0.203 | 0.216 |

*Notes:* The dependent variable is the DO $K$-density of total employment at 15 kilometers distance. Results are for all unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. We exclude pairs where at least one industry has less than 30 plants. Sample sizes vary across specifications because of missing covariates. All variables are standardized. All regressions include cohort fixed effects and industry-pair controls $ij$ (dissimilarity in industry-level input- and output shares with primary and business services industries). Column (1) includes all small industries (less than 30 plants). Column (2) excludes NAICS 4-digit pairs in the same NAICS 3-digit industry. Columns (3) and (4) show cross-sectional results in 2001 and 2003 (results for 2005 are not shown to save space). Columns (5) and (6) use separate input and output linkages. Column (7) and (8) use make-based measures of patent citation flows. Column (9) uses US technological relatedness from Steijn et al. (2022). Columns (10) and (11) instrument the input-output linkages with US data. Column (12) excludes the within-plant control. Last, column (13) includes all six Marshallian covariates (all flow- and similarity based measures). Huber-White robust standard errors in parentheses. $^a p < 0.01$, $^b p < 0.05$, $^c p < 0.1$.

Table B.6: Coagglomeration regressions for total employment, EGK measure (robustness).

| | (1) (EGK) | (2) (EGK) | (3) (EGK) | (4) (EGK) | (5) (EGK) | (6) (Flow) | (7) (EGK) | (8) (Flow) | (9) (Similarity) | (10) (EGK) | (11) (Flow) | (12) (EGK) | (13) (All) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Input-output linkages | 0.048$^{a}$ | 0.062$^{a}$ | 0.051$^{b}$ | 0.047$^{b}$ | | | 0.045$^{a}$ | 0.069$^{a}$ | | 0.071$^{a}$ | 0.072$^{b}$ | 0.065$^{a}$ | 0.073$^{a}$ |
| | (0.013) | (0.017) | (0.024) | (0.024) | | | (0.013) | (0.018) | | (0.021) | (0.028) | (0.013) | (0.018) |
| Input linkages | | | | | -0.025$^{c}$ | -0.010 | | | | | | | |
| | | | | | (0.015) | (0.017) | | | | | | | |
| Output linkages | | | | | 0.075$^{a}$ | 0.079$^{a}$ | | | | | | | |
| | | | | | (0.018) | (0.020) | | | | | | | |
| Input-output correlation | | | | | | | | | 0.004 | | | | -0.010 |
| | | | | | | | | | (0.016) | | | | (0.021) |
| Labor flows | | | | | | 0.003 | | -0.003 | | | -0.004 | | -0.001 |
| | | | | | | (0.011) | | (0.011) | | | (0.012) | | (0.015) |
| Occupational correlation | 0.041$^{a}$ | 0.055$^{a}$ | 0.079$^{a}$ | 0.043$^{b}$ | 0.060$^{a}$ | | 0.060$^{a}$ | | 0.063$^{a}$ | 0.057$^{a}$ | | 0.082$^{a}$ | 0.062$^{a}$ |
| | (0.012) | (0.013) | (0.023) | (0.020) | (0.012) | | (0.012) | | (0.014) | (0.013) | | (0.012) | (0.015) |
| Knowledge flows | 0.030$^{b}$ | 0.026$^{b}$ | 0.039$^{c}$ | 0.031 | 0.036$^{a}$ | 0.030$^{b}$ | | | | 0.034$^{a}$ | 0.029$^{b}$ | 0.039$^{a}$ | 0.026$^{b}$ |
| | (0.013) | (0.013) | (0.022) | (0.022) | (0.012) | (0.013) | | | | (0.012) | (0.013) | (0.012) | (0.013) |
| Knowledge flows (make) | | | | | | | 0.028$^{b}$ | 0.021$^{c}$ | | | | | |
| | | | | | | | (0.012) | (0.012) | | | | | |
| Technological relatedness | | | | | | | | | | | | | -0.032$^{b}$ |
| | | | | | | | | | | | | | (0.012) |
| US technological relatedness | | | | | | | | | 0.050$^{a}$ | | | | |
| | | | | | | | | | (0.009) | | | | |
| Within-plant share | 0.065$^{a}$ | 0.075$^{a}$ | 0.064$^{a}$ | 0.060$^{a}$ | 0.062$^{a}$ | 0.101$^{a}$ | 0.061$^{a}$ | 0.099$^{a}$ | 0.068$^{a}$ | 0.050$^{a}$ | 0.098$^{a}$ | | 0.086$^{a}$ |
| | (0.011) | (0.013) | (0.022) | (0.019) | (0.011) | (0.015) | (0.011) | (0.015) | (0.012) | (0.013) | (0.016) | | (0.015) |
| Ad valorem transport costs | -0.009 | -0.004 | 0.001 | -0.006 | -0.002 | 0.004 | -0.005 | 0.003 | -0.009 | -0.003 | 0.003 | -0.005 | 0.001 |
| | (0.009) | (0.010) | (0.019) | (0.016) | (0.009) | (0.010) | (0.009) | (0.010) | (0.009) | (0.010) | (0.010) | (0.009) | (0.010) |
| Multiplant share | -0.031$^{a}$ | -0.030$^{a}$ | -0.041$^{c}$ | -0.021 | -0.032$^{a}$ | -0.043$^{a}$ | -0.033$^{a}$ | -0.043$^{a}$ | -0.028$^{a}$ | -0.036$^{a}$ | -0.044$^{a}$ | -0.033$^{a}$ | -0.028$^{b}$ |
| | (0.010) | (0.011) | (0.021) | (0.019) | (0.011) | (0.011) | (0.011) | (0.011) | (0.011) | (0.011) | (0.011) | (0.011) | (0.011) |
| Observations | 10,710 | 9,506 | 3,403 | 3,321 | 10,045 | 9,506 | 10,045 | 9,506 | 10,045 | 10,045 | 9,506 | 10,045 | 9,506 |
| $R^2$ | 0.020 | 0.017 | 0.025 | 0.023 | 0.026 | 0.016 | 0.023 | 0.015 | 0.023 | 0.023 | 0.015 | 0.021 | 0.018 |

*Notes:* The dependent variable is the EGK coagglomeration measure of total employment at the census division level. Results are for all unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. We exclude pairs where at least one industry has less than 30 plants. Sample sizes vary across specifications because of missing covariates. All variables are standardized. All regressions include cohort fixed effects and industry-pair controls $ij$ (dissimilarity in industry-level input- and output shares with primary and business services industries). Column (1) includes all small industries (less than 30 plants). Column (2) excludes NAICS 4-digit pairs in the same NAICS 3-digit industry. Columns (3) and (4) show cross-sectional results in 2001 and 2003 (results for 2005 are not shown to save space). Columns (5) and (6) use separate input and output linkages. Column (7) and (8) use make-based measures of patent citation flows. Column (9) uses US technological relatedness from Steijn et al. (2022). Columns (10) and (11) instrument the input-output linkages with US data. Column (12) excludes the within-plant control. Last, column (13) includes all six Marshallian covariates (all flow- and similarity based measures). Huber-White robust standard errors in parentheses. $^{a}p < 0.01$, $^{b}p < 0.05$, $^{c}p < 0.1$.

Table B.7: Coagglomeration regressions (MRD employment).

| | (a) DO coagglomeration measure | | | | | | | (b) EGK coagglomeration measure | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) (Univar) | (2) (EGK) | (3) (Flow) | (4) (Similarity) | (5) (EGK) | (6) (Flow) | (7) (Similarity) | (8) (Univar) | (9) (EGK) | (10) (Flow) | (11) (Similarity) | (12) (EGK) | (13) (Flow) | (14) (Similarity) |
| Input-output linkages | 0.084$^a$ | 0.027$^b$ | 0.045$^a$ | | 0.025$^b$ | 0.052$^a$ | | 0.093$^a$ | 0.054$^a$ | 0.072$^a$ | | 0.039$^a$ | 0.059$^a$ | |
| | (0.012) | (0.011) | (0.015) | | (0.010) | (0.014) | | (0.012) | (0.012) | (0.016) | | (0.013) | (0.016) | |
| Input-output correlation | 0.152$^a$ | | | 0.110$^a$ | | | 0.042$^a$ | 0.115$^a$ | | | 0.045$^a$ | | | 0.001 |
| | (0.010) | | | (0.015) | | | (0.015) | (0.009) | | | (0.016) | | | (0.016) |
| Labor flows | 0.143$^a$ | | 0.104$^a$ | | | -0.001 | | 0.073$^a$ | | 0.047$^a$ | | | 0.002 | |
| | (0.010) | | (0.011) | | | (0.011) | | (0.008) | | (0.010) | | | (0.011) | |
| Occupational correlation | 0.161$^a$ | 0.132$^a$ | | 0.136$^a$ | 0.056$^a$ | | 0.058$^a$ | 0.131$^a$ | 0.106$^a$ | | 0.099$^a$ | 0.072$^a$ | | 0.074$^a$ |
| | (0.010) | (0.011) | | (0.013) | (0.011) | | (0.013) | (0.010) | (0.011) | | (0.014) | (0.012) | | (0.015) |
| Knowledge flows | 0.023$^c$ | 0.005 | 0.026$^c$ | | -0.010 | 0.004 | | 0.051$^a$ | 0.034$^a$ | 0.032$^a$ | | 0.027$^b$ | 0.023$^b$ | |
| | (0.012) | (0.012) | (0.014) | | (0.010) | (0.012) | | (0.011) | (0.011) | (0.011) | | (0.011) | (0.012) | |
| Technological relatedness | -0.073$^a$ | | | -0.151$^a$ | | | -0.092$^a$ | -0.005 | | | -0.038$^a$ | | | -0.030$^b$ |
| | (0.012) | | | (0.012) | | | (0.011) | (0.011) | | | (0.012) | | | (0.012) |
| Within-plant share | | | | | 0.061$^a$ | 0.125$^a$ | 0.132$^a$ | | | | | 0.060$^a$ | 0.102$^a$ | 0.104$^a$ |
| | | | | | (0.010) | (0.015) | (0.015) | | | | | (0.011) | (0.014) | (0.014) |
| Ad valorem transport costs | | | | | -0.159$^a$ | -0.150$^a$ | -0.147$^a$ | | | | | -0.022$^b$ | -0.010 | -0.017$^c$ |
| | | | | | (0.008) | (0.009) | (0.009) | | | | | (0.009) | (0.009) | (0.009) |
| Multiplant share | | | | | -0.298$^a$ | -0.310$^a$ | -0.286$^a$ | | | | | -0.075$^a$ | -0.087$^a$ | -0.065$^a$ |
| | | | | | (0.009) | (0.009) | (0.009) | | | | | (0.011) | (0.011) | (0.011) |
| Controls $ij$ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Cohort FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 10,292 | 10,292 | 9,729 | 9,729 | 10,045 | 9,506 | 9,506 | 10,292 | 10,292 | 9,729 | 9,729 | 10,045 | 9,506 | 9,506 |
| $R^2$ | | 0.067 | 0.059 | 0.085 | 0.211 | 0.211 | 0.220 | | 0.026 | 0.014 | 0.018 | 0.035 | 0.026 | 0.028 |

*Notes:* The dependent variable in columns (1)–(7) is the DO $K$-density at 15 kilometers distance. The dependent variable in columns (8)–(14) is the EGK coagglomeration measure computed at the densus division level. Both variables are computed using plants' management and R&D employment shares. Results are for all unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. Sample sizes vary across specifications because of missing covariates. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All regressions include cohort fixed effects and our industry-pair controls $ij$ (dissimilarity in industry-level input- and output shares with primary and business services industries), except in the univariate regressions in columns (1) and (8). Huber-White robust standard errors in parentheses. $^a p < 0.01$, $^b p < 0.05$, $^c p < 0.1$.

Table B.8: Coagglomeration regressions (production employment).

| | (a) DO coagglomeration measure | | | | | | | (b) EGK coagglomeration measure | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) (Univar) | (2) (EGK) | (3) (Flow) | (4) (Similarity) | (5) (EGK) | (6) (Flow) | (7) (Similarity) | (8) (Univar) | (9) (EGK) | (10) (Flow) | (11) (Similarity) | (12) (EGK) | (13) (Flow) | (14) (Similarity) |
| Input-output linkages | 0.067$^a$ | 0.023$^b$ | 0.035$^b$ | | 0.015 | 0.041$^a$ | | 0.079$^a$ | 0.061$^a$ | 0.081$^a$ | | 0.043$^a$ | 0.065$^a$ | |
| | (0.011) | (0.011) | (0.014) | | (0.010) | (0.013) | | (0.012) | (0.012) | (0.017) | | (0.012) | (0.017) | |
| Input-output correlation | 0.126$^a$ | | | 0.103$^a$ | | | 0.036$^b$ | 0.075$^a$ | | | 0.042$^a$ | | | 0.016 |
| | (0.009) | | | (0.016) | | | (0.016) | (0.009) | | | (0.016) | | | (0.018) |
| Labor flows | 0.117$^a$ | | 0.099$^a$ | | | -0.007 | | 0.033$^a$ | | 0.016$^c$ | | | -0.005 | |
| | (0.009) | | (0.011) | | | (0.011) | | (0.007) | | (0.009) | | | (0.010) | |
| Occupational correlation | 0.126$^a$ | 0.114$^a$ | | 0.113$^a$ | 0.038$^a$ | | 0.040$^a$ | 0.075$^a$ | 0.058$^a$ | | 0.045$^a$ | 0.034$^a$ | | 0.033$^b$ |
| | (0.010) | (0.011) | | (0.013) | (0.011) | | (0.013) | (0.010) | (0.011) | | (0.014) | (0.012) | | (0.014) |
| Knowledge flows | 0.031$^b$ | 0.014 | 0.028$^c$ | | -0.002 | 0.007 | | 0.052$^a$ | 0.040$^a$ | 0.037$^a$ | | 0.037$^a$ | 0.033$^b$ | |
| | (0.013) | (0.012) | (0.014) | | (0.011) | (0.012) | | (0.013) | (0.013) | (0.014) | | (0.013) | (0.014) | |
| Technological relatedness | -0.043$^a$ | | | -0.123$^a$ | | | -0.067$^a$ | 0.018$^b$ | | | 0.001 | | | -0.007 |
| | (0.010) | | | (0.011) | | | (0.010) | (0.009) | | | (0.009) | | | (0.010) |
| Within-plant share | | | | | 0.076$^a$ | 0.127$^a$ | 0.129$^a$ | | | | | 0.056$^a$ | 0.078$^a$ | 0.082$^a$ |
| | | | | | (0.011) | (0.015) | (0.015) | | | | | (0.011) | (0.014) | (0.014) |
| Ad valorem transport costs | | | | | -0.162$^a$ | -0.157$^a$ | -0.155$^a$ | | | | | 0.027$^a$ | 0.029$^a$ | 0.022$^b$ |
| | | | | | (0.009) | (0.009) | (0.009) | | | | | (0.010) | (0.010) | (0.010) |
| Multiplant share | | | | | -0.271$^a$ | -0.283$^a$ | -0.266$^a$ | | | | | -0.009 | -0.012 | 0.001 |
| | | | | | (0.010) | (0.009) | (0.010) | | | | | (0.012) | (0.012) | (0.012) |
| Controls $ij$ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Cohort FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 10,292 | 10,292 | 9,729 | 9,729 | 10,045 | 9,506 | 9,506 | 10,292 | 10,292 | 9,729 | 9,729 | 10,045 | 9,506 | 9,506 |
| $R^2$ | | 0.074 | 0.072 | 0.088 | 0.205 | 0.209 | 0.213 | | 0.014 | 0.007 | 0.005 | 0.017 | 0.010 | 0.008 |

*Notes:* The dependent variable in columns (1)–(7) is the DO $K$-density at 15 kilometers distance. The dependent variable in columns (8)–(14) is the EGK coagglomeration measure computed at the densus division level. Both variables are computed using plants' production employment shares. Results are for all unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. Sample sizes vary across specifications because of missing covariates. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All regressions include cohort fixed effects and our industry-pair controls $ij$ (dissimilarity in industry-level input- and output shares with primary and business services industries), except in the univariate regressions in columns (1) and (8). Huber-White robust standard errors in parentheses. $^a p < 0.01$, $^b p < 0.05$, $^c p < 0.1$.

Table B.9: Differenced regressions (MRD vs production employment; robustness).

| | (1) (EGK) | (2) (EGK) | (3) (EGK) | (4) (EGK) | (5) (EGK) | (6) (Flow) | (7) (EGK) | (8) (Flow) | (9) (Similarity) | (10) (EGK) | (11) (Flow) | (12) (EGK) | (13) (All) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Input-output linkages | -0.029$^a$ | -0.034$^a$ | -0.026$^c$ | -0.032$^b$ | | | -0.027$^a$ | -0.035$^a$ | | -0.046$^a$ | -0.058$^a$ | -0.021$^a$ | -0.031$^a$ |
| | (0.009) | (0.009) | (0.014) | (0.014) | | | (0.007) | (0.009) | | (0.017) | (0.020) | (0.007) | (0.009) |
| Input linkages | | | | | -0.003 | 0.005 | | | | | | | |
| | | | | | (0.011) | (0.011) | | | | | | | |
| Output linkages | | | | | -0.026$^a$ | -0.037$^a$ | | | | | | | |
| | | | | | (0.008) | (0.010) | | | | | | | |
| Input-output correlation | | | | | | | | | -0.041$^a$ | | | | -0.058$^a$ |
| | | | | | | | | | (0.011) | | | | (0.015) |
| Labor flows | | | | | | -0.031$^a$ | | -0.027$^a$ | | | -0.023$^b$ | | -0.002 |
| | | | | | | (0.009) | | (0.009) | | | (0.010) | | (0.013) |
| Occupational correlation | -0.084$^a$ | -0.077$^a$ | -0.078$^a$ | -0.064$^a$ | -0.075$^a$ | | -0.074$^a$ | | -0.057$^a$ | -0.073$^a$ | | -0.067$^a$ | -0.060$^a$ |
| | (0.010) | (0.010) | (0.017) | (0.017) | (0.010) | | (0.010) | | (0.011) | (0.010) | | (0.009) | (0.011) |
| Knowledge flows | 0.026$^b$ | 0.035$^a$ | 0.048$^b$ | 0.051$^a$ | 0.041$^a$ | 0.032$^a$ | | | | 0.042$^a$ | 0.032$^a$ | 0.042$^a$ | 0.037$^a$ |
| | (0.011) | (0.012) | (0.021) | (0.019) | (0.011) | (0.011) | | | | (0.011) | (0.011) | (0.011) | (0.012) |
| Knowledge flows (make) | | | | | | | 0.028$^a$ | 0.015 | | | | | |
| | | | | | | | (0.010) | (0.010) | | | | | |
| Technological relatedness | | | | | | | | | | | | | 0.038$^a$ |
| | | | | | | | | | | | | | (0.012) |
| US technological relatedness | | | | | | | | | -0.017$^b$ | | | | |
| | | | | | | | | | (0.008) | | | | |
| Within-plant share | -0.001 | -0.009 | 0.023 | 0.022 | 0.020$^b$ | -0.029$^b$ | 0.021$^b$ | -0.026$^b$ | 0.031$^a$ | 0.028$^a$ | -0.021$^c$ | | -0.004 |
| | (0.010) | (0.011) | (0.017) | (0.017) | (0.010) | (0.012) | (0.009) | (0.011) | (0.010) | (0.011) | (0.012) | | (0.012) |
| Ad valorem transport costs | -0.122$^a$ | -0.106$^a$ | -0.141$^a$ | -0.058$^a$ | -0.106$^a$ | -0.116$^a$ | -0.107$^a$ | -0.117$^a$ | -0.108$^a$ | -0.107$^a$ | -0.116$^a$ | -0.106$^a$ | -0.111$^a$ |
| | (0.010) | (0.010) | (0.017) | (0.018) | (0.010) | (0.010) | (0.010) | (0.010) | (0.010) | (0.010) | (0.010) | (0.010) | (0.010) |
| Multiplant share | 0.155$^a$ | 0.160$^a$ | 0.146$^a$ | 0.172$^a$ | 0.172$^a$ | 0.175$^a$ | 0.173$^a$ | 0.176$^a$ | 0.173$^a$ | 0.174$^a$ | 0.177$^a$ | 0.172$^a$ | 0.158$^a$ |
| | (0.014) | (0.015) | (0.027) | (0.029) | (0.015) | (0.015) | (0.015) | (0.015) | (0.015) | (0.015) | (0.015) | (0.015) | (0.015) |
| Observations | 10,710 | 9,506 | 3,403 | 3,321 | 10,045 | 9,506 | 10,045 | 9,506 | 10,045 | 10,045 | 9,506 | 10,045 | 9,506 |
| $R^2$ | 0.070 | 0.086 | 0.060 | 0.102 | 0.083 | 0.082 | 0.082 | 0.081 | 0.082 | 0.083 | 0.081 | 0.083 | 0.089 |

*Notes:* The dependent variable the DO $K$-density of MRD relative to that of production employment at 15 kilometers distance. Results are for all unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All regressions include cohort fixed effects and the industry-pair $ij$ controls. Column (1) includes all small industries (less than 30 plants). Column (2) excludes NAICS 4-digit pairs in the same NAICS 3-digit industry. Columns (3) and (4) show cross-sectional results in 2001 and 2003. Columns (5) and (6) use separate input and output linkages. Column (7) and (8) use make-based measures of patent citation flows. Column (9) uses US technological relatedness from Steijn et al. (2022). Columns (10) and (11) instrument the input-output variable with its US counterpart. Column (12) does not include the within-plant control. Last, column (13) includes all six Marshallian covariates (all flow- and similarity based measures). Huber-White robust standard errors in parentheses. $^a p < 0.01$, $^b p < 0.05$, $^c p < 0.1$.