DISCUSSION PAPER SERIES

No. 10921

INTER- AND INTRA-FIRM LINKAGES: EVIDENCE FROM MICROGEOGRAPHIC LOCATION PATTERNS

Kristian Behrens and Vera Sharunova

INTERNATIONAL TRADE AND REGIONAL ECONOMICS

ISSN 0265-8003

INTER- AND INTRA-FIRM LINKAGES: EVIDENCE FROM MICROGEOGRAPHIC LOCATION PATTERNS

Kristian Behrens and Vera Sharunova

Discussion Paper No. 10921
November 2015
Submitted 26 October 2015

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: (44 20) 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programme in **INTERNATIONAL TRADE AND REGIONAL ECONOMICS**. Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as an educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

INTER- AND INTRA-FIRM LINKAGES: EVIDENCE FROM MICROGEOGRAPHIC LOCATION PATTERNS

Abstract

Multiunit firms can draw on internal resources, thus their plants should depend less on external agglomeration benefits than comparable standalone plants. Because interacting at a distance is costly, multiunit firms should also be geographically 'compact'. We dissect the microgeographic location patterns of hundreds of thousands of Canadian establishments and find robust evidence for these predictions: multiunit firms are compact, and their plants locate in areas offering potentially less external agglomeration benefits. Within firms, plants with stronger vertical links are geographically more central. The latter effect is stronger for plants in high transport cost industries that produce durables and source a larger share of non-homogeneous inputs. These findings suggest that vertical supply chains are important in explaining firms' internal spatial organization.

JEL Classification: D22, L22, R12 and R32
Keywords: inter-firm linkages, intra-firm linkages, microgeographic location patterns, multiunit firms and spatial organization of firms

Kristian Behrens behrens.kristian@uqam.ca
Université du Québec à Montréal, Higher School of Economics, Moscow, CIRPÉE and CEPR

Vera Sharunova sharunov@bc.edu
Boston College and Higher School of Economics, Moscow

"Because a subsidiary may purchase its inputs or sell its outputs within the corporation, it may not be intimately involved with its neighbors." (Rosenthal and Strange, 2003, p.386)

"Cross-firm networking is less important for [multiunits (MUs) ...] and our interviews suggest MUs may rely more on intra-firm networks across their own establishments [...] estimation indicated that they do not benefit from networking with [singleunits (SUs)]." (Arzaghi and Henderson, 2008, p.1016).

1 Introduction

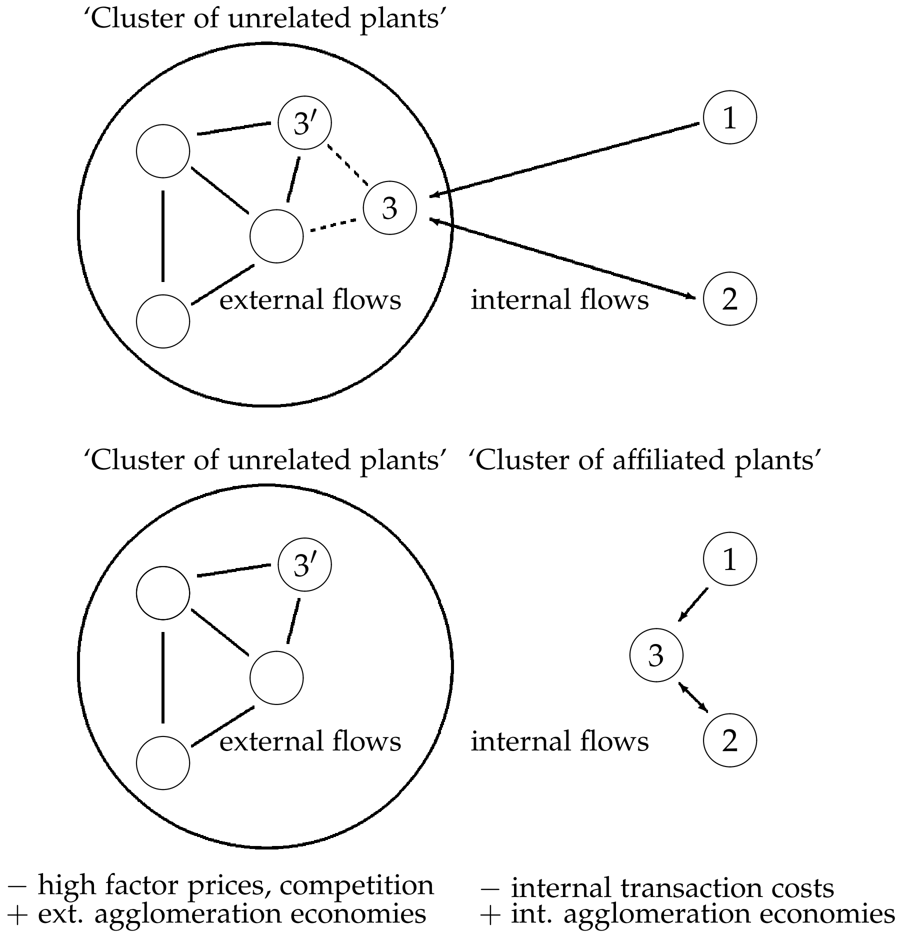
A large body of literature has substantiated the existence of agglomeration economies and their causal effect on workers' and firms' productivity (see Duranton and Puga, 2004, for the theory; and Combes and Gobillon, 2015, for recent empirical evidence). A key econometric challenge for models trying to identify the existence of agglomeration economies is that firms and plants do not choose locations randomly. Our aim is to precisely exploit the information contained in those non-random location choices: plants pick specific locations, and the association between locational characteristics and plant traits reveals how important the external environment is for certain plant types.[1] This approach, which looks at plants' revealed location choices, sheds light on the importance of external links between and internal links within firms in shaping their geographical structure and, more broadly, the location of economic activity.

Our empirical strategy is illustrated by Figure 1. Consider first the top panel. There are two plants, 3 and $3'$, that are identical save that plant 3 belongs to a multiunit firm, whereas plant $3'$ is standalone. Both plants interact with other plants in clusters of unrelated establishments ('external flows'). While these interactions are important for plant $3'$, they may be less important for plant 3. The reason is that plant 3 is affiliated with a multiunit firm and can draw on a larger pool of internal resources from establishments 1 and 2 of its firm ('internal flows'). Resource transfers within multiunit firms have been extensively documented, and we return to this point below. If the benefits from clustering to interact with unrelated plants are smaller for multiunit plants than for standalone plants, whereas the costs of clustering (e.g.,

[1] Arzaghi and Henderson (2008, pp.1015–1016) are among the few to document some associations between plant types and location choices. Focusing on the advertising agency industry they find that "MU [multiunit] and SU [single-unit] establishment stocks have different location patterns across zip codes in Manhattan in all years of our data." They link these differences in location patterns to differences in how firms 'network' among establishments. Turning to within-firm considerations, Alcácer and Delgado (2013, p.2) also "infer the existence of internal agglomerations from location choices." They use a conditional fixed-effects logit location choice model, where the key variable of interest is a firm's geographic footprint of activities in an economic area prior to the choice of location for the new establishment. They find for biopharmaceutical firms that within-firm Marshallian agglomeration forces are as important as between-firm agglomeration forces, and that "multi-unit firms are bound together through intra-firm linkages that impact performance." (*Ibid.*, p.3).

higher land rents; see Arzaghi and Henderson, 2008; Greenstone *et al.*, 2010) are the same for both types, multiunit plants should be 'less centrally' located in clusters but 'more centrally' in their firms, especially if internal flows are impeded by geographical distance.[2]

Figure 1: Possible geographical pattern of standalone and multiunit plants.


'Cluster of unrelated plants'
3′
3
1
2
external flows
internal flows
'Cluster of unrelated plants'
'Cluster of affiliated plants'
3′
1
3
2
external flows
internal flows
− high factor prices, competition
+ ext. agglomeration economies
− internal transaction costs
+ int. agglomeration economies

The bottom panel of Figure 1 depicts such a location pattern that reflects the differential degree to which heterogeneous types of plants depend on their external environment.[3] As depicted in the figure, multiunit firms should be geographically 'compact' by forming clusters of affiliated establishments to reduce internal transaction costs.

[2] There is a large literature documenting that geographical distance has a strong impact on various economic outcomes. See, e.g., Adams and Jaffé (1996) for intra-firm R&D transmission; Coval and Moskowitz (2001) for the geography of mutual fund performance; Petersen and Rajan (2002) for the geography of small business lending; Landier, Nair, and Wulf (2007) for plant divestments and labor-related aspects; Giroud (2013) for investments and productivity; Kalnins and Lafontaine (2014) for profitability and productivity; and Eichholtz, Holtermans, and Yönder (2015) for returns to commercial real estate investments.

[3] Differential benefits and identical costs lead to the sorting of heterogeneous agents across locations. See, e.g., Forslid and Okubo (2014) and Gaubert (2014) for empirical evidence compatible with spatial sorting of firms by productivity. Instead of productivity, we look at sorting by organizational traits like multiunit status.

We test these predictions by dissecting the microgeographic location patterns of hundreds of thousands of Canadian establishments. We pay particular attention to the following hypotheses: (i) input-output linkages and agglomeration economies are less important for multiunit plants, i.e., the characteristics of locations picked by multiunit plants differ in systematic ways from those picked by comparable standalone plants; (ii) internal transfers in multiunit firms are distance sensitive so that multiunit firms are geographically 'compact'; and (iii) plants more strongly involved in resource transfers are 'more centrally' located within multiunit firms.

Testing these predictions is difficult for two reasons. First, theory provides only limited guidance on how to measure a location's 'input potential', how to think about the internal geographical structure of multiunit firms, or how to assess which plants may be more strongly involved in internal transactions.[4] Second, the data requirements are fairly stringent. We need fine-grained spatial plant-level data that allows us to precisely track the geographical structure of multiunit firms. We also need to construct very granular spatial controls to purge the effects that other locational characteristics could have on the differential location behavior of standalone and multiunit plants.[5] Last, since internal transfers within firms are usually not directly observable from the data, we need to devise various proxies for them.

To deal with these issues, we develop a new approach for measuring inter- and intra-firm linkages using microgeographic data. Building on the fact that links between firms are more likely to exist when they are geographically close (Atalay, Hortaçsu, Roberts, and Syverson, 2011; Bernard, Moxnes, and Saito, 2015), our measure of a location's 'input potential' combines pairwise distance data between plants with highly disaggregated input-output tables. Recent work on coagglomeration has exploited the co-location of industry pairs (Duranton and Overman, 2005, 2008; Ellison, Glaeser, and Kerr, 2010; Faggio, Silva, and Strange, 2014), and our measures go further and aggregate location patterns across all industries. Our microgeographic data also allow us to track the internal structure of multiunit firms. We construct measures of 'within-firm centrality' and relate them to proxies for a plant's involvement in supply chains.

Previewing our key results, we find strong support for our hypotheses. Plants affiliated with multiunit firms are systematically located in areas that offer potentially worse access to external input suppliers. Their locations are also less specialized in their own industry than those chosen by comparable standalone plants. When taken together, the location patterns of multiunit plants strongly suggest that these plants do depend less on their external environment than comparable standalone plants. The effect is stronger for larger firms, as measured

[4] Input-output linkages, both between and within firms, are a staple of theoretical agglomeration models and models of firm's location choices. See Beckman and Thisse (1987) for a classical synthesis on the location of production activity. However, applied work remains limited. We also know little about the geographical structure of multiunit firms, except for some largely anecdotal evidence.

[5] Having access to spatially fine-grained data is crucial since agglomeration effects dissipate quickly with distance (between 1 mile and 5 mile, according to Rosenthal and Strange, 2010; after 750 meters for advertising agencies in Manhattan, according to Arzaghi and Henderson, 2008).

by either the number of plants or employment size. Our results are robust to a large range of plant-level and geographically fine-grained controls. They also hold when excluding older or larger plants that may be able to structure the economic environment surrounding them.

Turning to the internal geography of firms, we first substantiate evidence that multiunit firms are more compact than they 'should be' based on observable characteristics. Using stratified propensity score matching, we construct counterfactual multiunit firms from standalone plants, controlling finely for plant-level and geographical characteristics. The real multiunit firms are on average 50% more compact geographically than the counterfactual ones. This suggests that within-firm transfers are impeded by distance and important in shaping firms' geographical structure. We further find that plants that are more likely to be involved in vertical chains – plants that are in industries more strongly linked downstream to the industries of the other plants of the firm – are geographically more centrally located. This is especially true for establishments that operate in the firm's core segment of business, that are in industries that face high transportation costs, manufacture durable goods, and source a larger share of non-homogeneous ('complex') inputs. These findings suggest that vertical supply chains involving both the exchange of tangible inputs and the monitoring of complex transactions may be important in explaining the geographical structure of multiunit firms.

The remainder of the paper is structured as follows. Section 2 reviews the related literature. Section 3 provides information on our data sources and gives details on how we construct the key variables. It also provides a first set of descriptive facts on inter- and intra-firm linkages. Section 4 contains the baseline multivariate results and a large number of robustness checks. Finally, Section 5 concludes. We relegate a large number of data descriptions, technical issues, additional results, and supplemental material to an extensive set of appendices.

2 Related literature

Our work relates to three broad strands of literature. First and foremost, it is linked to the fairly sparse literature on the importance of input-output linkages for the spatial structure of industries and firms. While these linkages are a theoretical staple of agglomeration models (see Duranton and Puga, 2004, for a review), little is known empirically about their role in structuring economic activity. Even less is known about their role in structuring firms. Looking at industries, Holmes (1999) and Li and Lu (2009) document that firms located in more specialized areas have a larger intermediates-to-sales ratio, i.e., are vertically more disintegrated.[6] Duranton and Overman (2005, 2008) and Ellison *et al.* (2010) document the tendency

[6] Johnson and Noguera (2012) look at the geography of cross-border supply chains in international trade. They find a strong geographical component to vertical disintegration across national borders. Most of the rise in international supply chains involves exchanges over short geographical distances.

of industry pairs with strong input-output links – as measured by industry-level input-output coefficients – to coagglomerate (see Helsley and Strange, 2014, for the theory and some methodological caveats). The effect of these links is stronger for more coagglomerated industries, for small and young firms, and for low-tech and low-education industries in the UK (Faggio *et al.*, 2014). Overall, these results provide evidence for the strong spatial effects that input-output links have in explaining how the most coagglomerated industries relate to one another. Turning to firms, there is an emerging literature on between-firm networks. Arzaghi and Henderson (2008) document the existence of between-agency networks in the Manhattan advertising industry, while Bernard *et al.* (2015) use extensive Japanese firm-level data and find that lower costs of moving people lead firms to expand their geographical supplier networks by searching for potential match partners over longer distances. Atalay *et al.* (2011), Atalay, Hortasçsu, and Syverson (2014), and Raimondo, Rappoport, and Ruhl (2014) investigate vertical links within large firms. Their findings suggest that vertical links involving the exchange of tangible goods within firms play only a limited role in explaining why these firms exist. The importance of these links for the geographical structure of the firms – conditional on their existence – is however left virtually unexplored. Otazawa and van Ommeren (2015) connect the between-firm networks with the geographical location of Japanese firms in the city of Kure. They show that, consistent with the theoretical results of Helsley and Zenou (2014), firms that are more central in buyer-seller networks are also geographically more centrally located. Their analysis is, however, silent on multiunit firms and their internal structure. Duranton and Overman (2008) investigate the spatial structure of multiunit firms and find that they are more compact than randomness would predict. While important, their results do not tell us anything about why they are more compact, and how vertical links or access to inputs may help to explain this finding. The remaining literature on the geographical structure of firms has largely focused on headoffices. In particular, it has documented that headoffices are located 'geographically centrally' in firms (Baldwin and Brown, 2005; Aarland, Davis, Henderson, and Ono, 2007; and Henderson and Ono, 2008). Little else is known about the spatial structure of multiunit firms, the role of vertical links in shaping that structure, and the location or colocation of plants within firms (Alcácer, 2006, is an exception).

Second, our paper is tied to the thin and scattered literature on industrial organization and agglomeration, as well as the heterogeneity of agglomeration benefits for firms. The existing evidence points towards the differential importance of agglomeration effects for small and large firms (a conjecture first made by Chinitz, 1961; see Shaver and Flyer, 2000; Glaeser and Kerr, 2009, and Rosenthal and Strange, 2005, 2010, who talk about a 'small establishment effect', which is largely due to customer-supplier relationships). While the main focus has been on firm size, the organizational dimensions of the firm have received limited attention. For example, the importance of multiunit versus standalone status of establishments has, to the best of our knowledge, received almost no attention until now. Rosenthal and Strange (2003) and Brown and Rigby (2015) are exceptions. The former look at the effects of affiliated versus unaffiliated plants on births and employment levels at new plants, whereas the latter examine how the economic environment benefits differentially multiunit and single-unit plants. Their results are relatively inconclusive and the internal spatial structure of firms is not analyzed. The strategic management literature has paid more attention to the organizational aspects of agglomeration, and how agglomeration and the spatial structure of firms interact (e.g., Audida, Pino, Sorensen, and Hage, 2000; Alcácer and Chung, 2014; Alcácer and Delgado, 2013). Most of that work deals, however, with specific industries or greenfield FDI, and organizational aspects are often subsumed by employment size variables. One of the main messages of that literature is that plants contribute to and benefit from agglomeration in a potentially asymmetric way. Plants that contribute a lot and benefit little face an 'appropriability risk' and therefore tend to shun places with large concentrations of firms in their own industry (locating in such places may make other firms likely to poach the labor force, to clutter the suppliers, or to steal ideas). In some sense, there is a geographical 'adverse selection' phenomenon.

Last, our paper relates to the growing literature on 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 (e.g., credit constraints, agency issues, or hold-up problems when intermediates are relationship-specific). Multiunit firms have internal capital markets (see Lamont, 1997; Maksimovic and Philips, 2002) and internal labor markets (see Silva 2013; Tate and Yang, 2015), which affect firm- and plant-level productivity and industry allocative efficiency (see Shoar, 2002; Ševčík, 2013).[7] The costs of interacting at a distance in those internal markets may explain why multiunit firms are geographically 'compact': locational constraints involving external suppliers being less stringent, those firms can geographically re-center on themselves to reduce costly exchanges in internal markets.

3 Data and measurement

We briefly present our data and explain how we construct the key variables – in particular the measures of inter- and intra-firm linkages. We relegate a more detailed description of the data and of the controls to Appendix A.

[7] Strange, Hejazi, and Tang (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.

3.1 Data overview

We use a large establishment-level dataset that contains information on manufacturing and non-manufacturing plants operating in Canada.[8] Our dataset extensively and representatively covers the manufacturing sector (see Behrens and Bougna, 2015, for additional details). It is much less exhaustive for non-manufacturing establishments, so we use the latter more parsimoniously for robustness checks only. Our data cover the years 2001 to 2013, in two-year intervals. For every plant, we have information on its primary 6-digit NAICS code and up to four secondary 6-digit NAICS codes; the year of establishment of the plant; its employment; whether or not it is an exporter; whether or not it is a headoffice; its 6-digit postal code; up to ten products produced by the plant; and the legal name of the entity to which it belongs. We make extensive use of the detailed geographical and industry-product nature of our dataset. Since the database has been developed and is maintained as a manufacturing business register, the location and the industry-product dimensions of the data are of high quality.

We geocode all plants by latitude and longitude using the postal code centroids of their address (see also Duranton and Overman, 2005; Behrens and Bougna, 2015).[9] We associate plants with firms using the name of the legal entity to which they belong. More details on that matching procedure – as well as its various advantages and drawbacks – are provided in Appendix A. After basic data cleaning – dropping plants with missing data or for which the postal code could not be matched with geographical coordinates – we are left with a sample of 321,589 manufacturing plants and 580,116 non-manufacturing establishments associated with 804,784 firms. Of those firms, 37,249 are multiunit and operate 134,170 plants in total.

3.2 Microgeographic inter-firm linkages

Input-output linkages are usually proxied in the literature by either industry- or plant-level data on intermediates-to-sales ratios (the 'purchased-inputs intensity' of the plant or the industry; Holmes, 1999; Rosenthal and Strange, 2001). Atalay *et al.* (2011) go a step further and exploit Compustat firm-level data on self-reported clients that account for more than 10% of firms' sales to construct the network structure of production linkages for a panel of large listed companies operating in the US. One finding of their analysis is that distance matters both for the *existence and formation of links* between firms: "Distance is also an important determinant of the probability that two vertices [firms] are linked to one another [...] Compared with firm-pairs for which the supplier and customer are 100–500 mi apart, two firms with headquarters less than 25 mi apart are 0.18% more likely to be connected." (Atalay *et al.*, 2011, Supporting Information, p.3). Bernard *et al.* (2015) also document that distance is a strong determinant

[8] We henceforth interchangeably use the terms 'plant' or 'establishment'.

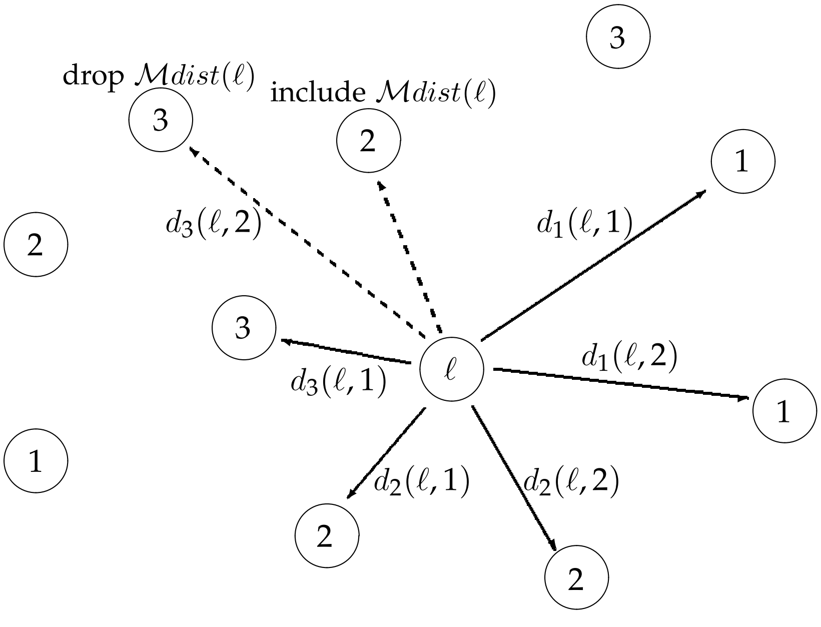
[9] Since Canadian postal codes are very fine grained – especially in more densely populated areas – using their centroids is, for all practical purposes, as good as geocoding the data.

of links: “Geographic proximity plays an important role in the matching of suppliers and customers. Most connections are local; the median distance to a supplier or customer is 30 kilometers.” (*Ibid.*, p.1). In what follows, we exploit the fact that geography matters for the existence of input-output linkages to construct a proxy for those linkages using the microgeographic nature of our data (see Behrens, Bougna, and Brown, 2015). In a nutshell, we develop a measure of the geographical access to potential input suppliers or clients, based on a plant's local environment, and then analyze whether there are systematic differences in that measure between multiunit and standalone plants.

3.2.1 Measurement

Figure 2 illustrates the logic underlying the construction of our measures for input- and output linkages between plants. We know exactly where plants are located. Since buyer-supplier links are statistically more likely to exist when plants are geographically close, we assume that plants source on average from close-by suppliers.

Figure 2: Construction of input-, output-, and minimum-distance measures.



To formalize this idea, fix a number N and compute industry-by-industry the average great-circle distance d of plant ℓ to the N closest plants in each *other* industry i.[10] Let $k_i(\ell,p)$ denote the p-th closest industry-i plant to plant ℓ. The average distance from plant ℓ to the N closest

[10] We exclude plant ℓ's own 6-digit industry, since the concentration of plants in the same industry is likely to reflect many factors that are unrelated to input-output linkages across industries. We also computed the linkages when excluding all plants in the same 4-digit industries. The correlations between the two series of linkages are 0.99 and 0.98 for inputs and for outputs, respectively. This shows that our measures do not just pick up clustering at a higher level of industrial aggregation.

plants in industry i is given by:

$$\overline{d}_i(\ell) \equiv \frac{1}{N}\sum_{p=1}^{N} d\big(\ell, k_i(\ell,p)\big).$$

We then compute the weighted average of these average distances, with weights given by national input-output shares at the 6-digit level. Let $\omega^{\text{in}}_{\Omega(\ell),i}$ and $\omega^{\text{out}}_{\Omega(\ell),i}$ denote the share of inputs sourced by industry $\Omega(\ell)$ of plant ℓ from industry i, and the share of outputs sold by industry $\Omega(\ell)$ of plant ℓ to industry i, respectively.[11] Our microgeographic measures of inter-firm linkages for plant ℓ are then constructed as follows:

$$\mathcal{I}\text{dist}_\ell = \sum_{i\in\Omega\backslash\Omega(\ell)} \omega^{\text{input}}_{\Omega(\ell),i} \times \overline{d}_i(\ell), \tag{1}$$

for inputs, or upstream interactions, and

$$\mathcal{O}\text{dist}_\ell = \sum_{i\in\Omega\backslash\Omega(\ell)} \omega^{\text{output}}_{\Omega(\ell),i} \times \overline{d}_i(\ell), \tag{2}$$

for outputs, or downstream interactions. Since the input and output shares sum to one, $\mathcal{I}\text{dist}_\ell$ is the minimum average distance (in kilometers) of plant ℓ to one dollar of inputs from its N closest potential suppliers. Analogously, $\mathcal{O}\text{dist}_\ell$ can be interpreted as the minimum average distance plant ℓ has to ship one dollar of its outputs to its N closest potential customers.[12] The larger are $\mathcal{I}\text{dist}_\ell$ or $\mathcal{O}\text{dist}_\ell$, the worse are a priori plant ℓ's external input or output linkages – it is on average farther away from a dollar of intermediate inputs or a dollar of demand emanating from the other industries. Note that our measure is an *inverse measure of input linkages*: large values of $\mathcal{I}\text{dist}_\ell$ mean that the plant is far away from potential input sources, whereas small values mean that the plant is close to potential input sources.

Several comments are in order. First, although we view our measures as proxies for input- and output linkages, they can more broadly capture various interactions between plants in vertically linked industries. These interactions need not involve the exchange of intermediate goods, but can capture access to information about technological change, or R&D expertise, or shifting market conditions in vertically linked industries. Although all these aspects may enter into consideration, the evidence that we substantiate later makes it hard to believe that the exchange of tangible goods plays no role at all. Second, our measures have two advantages compared to those of Atalay *et al.* (2011) or Bernard *et al.* (2015). The first advantage is that we

[11] The input-output shares are computed based either on 242 manufacturing industries alone (superscript 'mfg'), or based on all 864 industries comprising the whole economy (superscript 'all'). See Appendix A.2 for details.

[12] We have no information on the spatial distribution of final demand and thus cannot include it in our measures. We could construct a population-weighted market potential measure as a proxy. However, such a measure changes slowly through time so that its effect should be washed out by industry fixed effects. We also include a measure of density of economic activity in our estimations and this should capture most market potential effects.

work at the plant level, and not at the firm level. We believe that plants are the relevant unit of analysis when asking questions about agglomeration economies and the spatial structure of firms. The second advantage is that we exploit in a very fine way the microgeographic location patterns of plants for computing our linkages.[13] Third, we use a weighting scheme that derives from national input-output tables. Since geographical specialization is associated with more vertical disintegration (Holmes, 1999; Li and Lu, 2009), plant-level input-output linkages are potentially endogenous to location choices through input substitution. Using national averages, as we do in our measures, partly alleviates that problem. To further mitigate potential endogeneity concerns, we use three-year lagged values for the input-output tables for all years in our estimations (e.g., the 2011 plant-level measures are constructed using the weights from the 2008 input-output tables). Of course, location choices of 'large plants' may structure the regional economic environment by creating buyer-supplier networks around them (see, e.g., Klier and McMillen, 2008, who discuss the case of the US automotive industry). We deal with this potential issue in our estimations. Fourth, observe that our measures are standardized and independent of plant size. We control for plant size separately in our subsequent regressions.[14] Last, the choice of N is essentially arbitrary. In what follows, we compute our measures (1) and (2) for all years and for all plants using the $N = 3, 5, 7$, and 10 nearest plants in each industry. There is a trade-off between too small and too large values of N. For small values, the year-on-year volatility of the input and output distance measures increases significantly due to substantial plant turnover. For large values, the cross-plant variation decreases, which makes identification of any effects of these linkages more difficult. Also, as reported by Atalay *et al.* (2011), the average (and median) number of links between firms is fairly small, except for the very large firms. Our results are robust to the choice of $N = 3, 5, 7$ and 10. We use $N = 5$ as our baseline value, and report estimates using other values as robustness checks.

One final comment is in order. To control for the fact that (1) and (2) are mechanically smaller in denser areas, we also compute a 'minimum distance measure', i.e., the distance of plant ℓ from the $N \times (|\Omega| - 1)$ closest plants regardless of the industries they belong to.[15] Figure 2 shows, for example, that compared with the previous input-output computations, one plant of industry 3 would be dropped and replaced with a closer plant in industry 2. Including that measure into our regressions then controls for the overall density. Hence, our input-output linkage measures pick up the effect of being closer to a dollar of inputs or outputs conditional

[13] Bernard *et al.* (2015) investigate the links between Japanese firms. They have, however, only access to firm-level data. Thus, the address information they use corresponds to that of the firm's legal address. They cannot investigate the geographical structure of multiunit firms.

[14] We do not weight larger plants more heavily in the computations of our linkage measures because larger plants do not seem to interact more (Rosenthal and Strange, 2010; Alcácer and Chung, 2014).

[15] This also controls for the fact that plants in some industries, like 'Explosives manufacturing', may have long input- and output distances simply because they are generally quite far away from other plants, for reasons (e.g., zoning) that may be completely unrelated to input-output considerations per se.

on the overall density of the area. Let $M \equiv N \times (|\Omega| - 1)$ denote the number of closest plants for which we compute the minimum distance, which equals the number used to compute the corresponding input-output linkages. Formally, we compute

$$\mathcal{M}\text{dist}_\ell = \frac{1}{M} \sum_{n=1}^{M} d(\ell, k_{n\in\Omega\setminus\Omega(\ell)}(n, \ell)), \tag{3}$$

where $d(\ell, k_{n\in\Omega\setminus\Omega(\ell)}(n, \ell))$ denotes the distance to the nth closest plant in any industry but $\Omega(\ell)$. The measure (3) can be interpreted as the counterfactual input- or output-linkages plant ℓ would have if its input shares were the same in all industries and if inputs could be sourced from any industry or outputs shipped to any industry. We will either use the log of $\mathcal{I}\text{dist}_\ell$ in our regressions, controlling for the log of $\mathcal{M}\text{dist}_\ell$, or directly the log of the ratio $\mathcal{I}\text{dist}_\ell/\mathcal{M}\text{dist}_\ell$.

3.2.2 A first look at inter-firm linkages

We compute our input-, output-, and minimum-distance measures (1)–(3) for each plant and year. Table 1 summarizes the results. As one can see, the number of manufacturing plants has decreased substantially in Canada over our study period. At the same time, the average input distance across all manufacturing plants has increased from 168 kilometers in 2001 to 205 kilometers in 2013. Put differently, manufacturing plants in 2013 are on average 37 kilometers farther away from their potential input sources than plants in 2001, a 22% increase. The corresponding figure for output distances is 27 kilometers, a 15% increase. As can be further seen from Table 1, there is substantial variation in both measures across plants. As expected, there is also a lot of variation in the minimum distance measure.

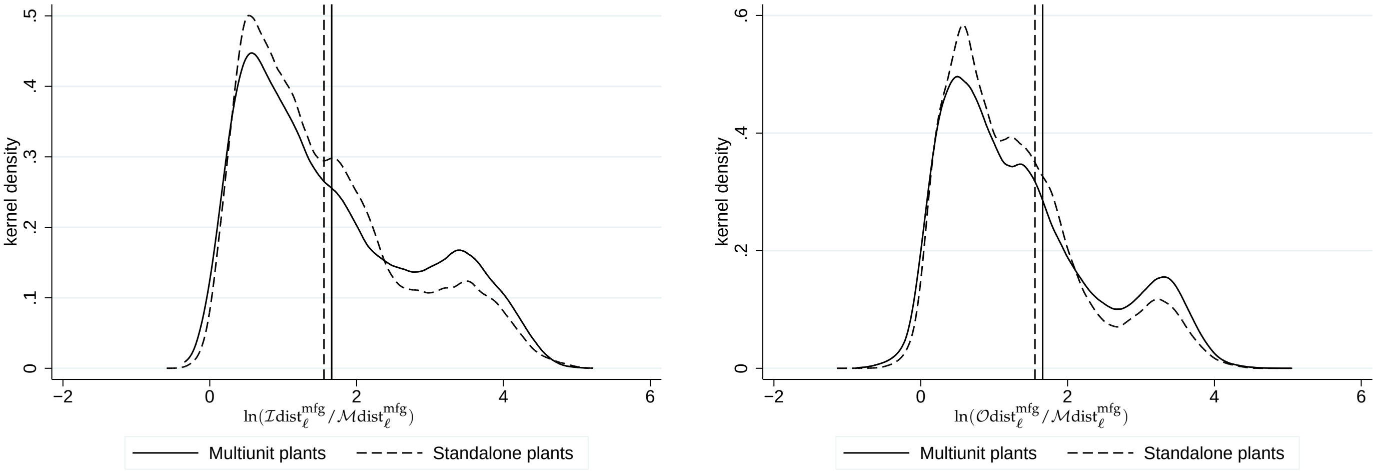
Table 1: Descriptive statistics for input-, output-, and minimum distance measures.

Input-output-minimum distances, all manufacturing industries							
		$\mathcal{I}dist_\ell^{\text{mfg}}$		$\mathcal{O}dist_\ell^{\text{mfg}}$		$\mathcal{M}dist_\ell^{\text{mfg}}$	
Year	Number of plants	Mean	Std dev.	Mean	Std dev.	Mean	Std dev.
2001	52,031	167.91	180.64	126.88	137.11	51.03	84.33
2003	51,876	172.73	174.83	134.47	142.40	51.15	84.82
2005	49,223	166.31	175.95	130.71	144.33	53.50	92.72
2007	46,243	175.73	181.16	134.77	148.46	55.54	97.38
2009	44,681	173.64	178.05	134.01	144.57	56.71	98.30
2011	42,210	195.09	201.02	141.23	152.60	58.46	99.80
2013	35,325	204.79	208.60	156.35	163.37	66.28	106.10

Notes: Descriptive statistics for 242 (concorded) manufacturing industries. We report results for $N = 5$. Results are unweighted industry averages by year. Additional details by industry for 2005 are provided in Table 12 in Appendix D.1, which summarizes the top-ten industries with the shortest input distances and the bottom-ten industries with the longest input distances in 2005.

Figure 3 depicts the plant-level distribution and the average (vertical lines) of our relative input and output distance measures – $\ln(\mathcal{I}\text{dist}_\ell/\mathcal{M}\text{dist}_\ell)$ and $\ln(\mathcal{O}\text{dist}_\ell/\mathcal{M}\text{dist}_\ell)$ – for all manufacturing plants and all years pooled. We separately depict the distributions for plants affili-

Figure 3: Distribution of relative input- and output distances, multiunit vs standalone plants.



ated with multiunit firms and for standalone plants. As can be seen from Figure 3, conditional on overall density, multiunit plants have on average worse input and output linkages than standalone plants. Multiunit plants are slightly overrepresented at very short distances, underrepresented at intermediate distances, and substantially overrepresented at long distances. Hence, there is more variation in access to inputs for multiunit plants (standard deviation of 1.17 versus 1.09 for standalone plants). Of course, many reasons may drive this pattern. We show later that even conditional on a large number of plant-level, firm-level, and geographical controls, this pattern survives, thereby suggesting that multiunit plants are less dependent on local external suppliers, possibly because they can draw more on internal resources.[16]

Before proceeding further, let us point out that inter-firm linkages – as measured by access to inputs or outputs – provide only one measure of the potential interactions between plants. Another one is given by more traditional geographical 'specialization' measures, i.e., the share of plants or employment within the same industry in some radius around the plant. We detail in Appendix B.1 how we construct those measures – which we use as controls later – and we show in Appendix B.2 that multiunit plants also systematically locate in less specialized areas. This again suggests that multiunit plants are less dependent on external interactions and agglomeration economies than standalone plants.

3.3 Microgeographic intra-firm linkages

The literature on the spatial structure of multiunit firms is thin at best. Most of that literature considers either descriptive case studies or has looked at the distance between firms' headof-

[16] A similar pattern, though less pronounced, holds for multiunit plants versus standalone plants when we measure input linkages using the full input-output tables with all 864 industries (and not just the 242 manufacturing industries). Plants affiliated with multiunit firms are slightly farther away from all of their inputs. However, they are slightly closer to all of their potential customers. Again, these are unconditional results.

fices (HOs) and production plants to investigate the impact of that distance on various outcomes and performance measures (see Baldwin and Brown, 2005; Aarland *et al.*, 2007; Landier, Nair, and Wulf, 2007; Henderson and Ono, 2008; Giroud, 2013; Kalnins and Lafontaine, 2014). To our knowledge, there is no agreement on how to think about the spatial structure of multiunit firms. Aarland *et al.* (2007) look at the colocation patterns of HOs and production units using US counties as their spatial unit of reference. They find that firms with plants located across multiple different counties are more likely to have centralized HOs, "possibly to locate the [HOs] in a central place to enhance coordination as well as monitoring." (*Ibid.*, p.493). They compute the employment-weighted distance between the county centroids where the firm's plants are established. This provides a measure of the 'average spacing' of the establishments of the firm. Baldwin and Brown (2005) follow a similar strategy and compute the average straight-line distance of the headquarter to the different production plants of the firms to capture its spatial structure. Giroud (2013) and Kalnins and Lafontaine (2014) look at pairwise interactions between HOs and plants but disregard the overall spatial structure of the multiunit firm.

3.3.1 Measurement

We follow Baldwin and Brown (2005) and exploit the microgeographic nature of our data to define the *spatial extent of a multiunit firm* using the bilateral distances between its plants. We compute unweighted and employment weighted measures for each plant individually, and for all plants within the same firm. Let $j(f)$ and $k(f)$ denote two plants, j and k, belonging to the same firm f, which has $n(f)$ plants in total. Let $d_{j(f),k(f)}$ stand for the distance – measured using the great circle distance in kilometers – between plants j and k of firm f. We define the average intra-firm distance of plant j – both unweighted and employment weighted – as follows:

$$\overline{d}_{j(f)} = \frac{1}{n(f)-1}\sum_{k\neq j} d_{k(f),j(f)} \quad \text{and} \quad \overline{d}^{w}_{j(f)} = \frac{1}{\sum_{k\neq j} e(k(f))}\sum_{k\neq j} e(k(f)) d_{k(f),j(f)}, \tag{4}$$

where $e(k(f))$ denotes the employment in plant k of firm f. Hence, at the plant level, we compute the distance of the plant to all other plants in the same firm. This measure tells us how far or close a single plant is located with respect to the other components of the multiunit firm. We can then average these measures across plants within the firm to obtain a firm-level measure of its geographical extent as follows:

$$\overline{d}_f = \frac{1}{n(f)}\sum_j \overline{d}_{j(f)} = \frac{2}{n(f)[n(f)-1]}\sum_k \sum_{j>k} d_{k(f),j(f)}, \tag{5}$$

where the second equality stems from the symmetry of the distances. In words, the unweighted spatial extent $\overline{d}_f$ of firm f is the average bilateral distance between all of its plants. Because the unweighted measure may assign too much weight to small plants, and because large plants may behave differently, we also compute an *employment-weighted spatial extent* as follows:

$$\overline{d}^w_f = \frac{1}{\sum_j e(j(f))} \sum_j e(j(f)) \overline{d}^w_{j(f)} = \frac{1}{\sum_k \sum_{j>k} e(k(f)) e(j(f))} \sum_k \sum_{j>k} e(k(f)) e(j(f)) d_{k(f),j(f)}. \quad (6)$$

Last, we compute a purely non-geographical measure of the strength of vertical relationships for each plant and for the firm as a whole as follows:

$$\overline{IO}_{j(f)} = \frac{1}{n(f)-1} \sum_{k \neq j} \omega^{IO}_{\Omega(k(f)),\Omega(j(f))} \quad \text{and} \quad \overline{IO}_f = \frac{2}{n(f)[n(f)-1]} \sum_k \sum_{j>k} \omega^{IO}_{\Omega(k(f)),\Omega(j(f))}, \quad (7)$$

where $\omega^{IO}_{\Omega(k(f)),\Omega(j(f))}$ is either the share of inputs, the share of outputs, or the simple average between the two. As stated before, the shares are observed at the industry level and they are computed either for manufacturing only or for all industries. We show later that plants with larger measure (7), i.e., plants that are in industries that are vertically strongly linked to the industries in which the other plants of the firm operate, are also geographically more centrally located within the firm.

3.3.2 A first look at intra-firm linkages

We now take a first look inside the geographical structure of multiunit firms. Table 2 summarizes the ten industries with the most and the ten industries with the least spatially compact multiunit firms, pooled across all years. We measure the spatial extent using the weighted average distance (6). We also report the average number of plants of the multiunit firms in those industries.[17] The last column of Table 2 reports the average internal input-output shares between plants in the firm, as given by (7).

Comparing the bottom ten and the top ten industries, the average within-firm input-output shares between plants are a meager 0.29% for the bottom ten industries. Yet, for the top ten industries, that share is about 1.33%, more than four times as large. Although this result is just an unconditional comparision of means, it already suggests that multiunit firms that operate plants in vertically more strongly linked industries are geographically more compact than multiunit firms operating plants in vertically less strongly linked industries. We show below that this result is robust to a large number of controls in a multivariate setting.

Another way to see that firms with more strongly linked plants are geographically more compact is to plot the spatial extent (5) against an 'input-output weighted' spatial extent.[18] Figure 4 depicts that relationship for all multiunit firms in our dataset. As can be seen, the two

[17] Firms with more plants may be mechanically more dispersed. We control for this effect later on and show that it does not drive our results.

[18] To this end, we compute the following measure:

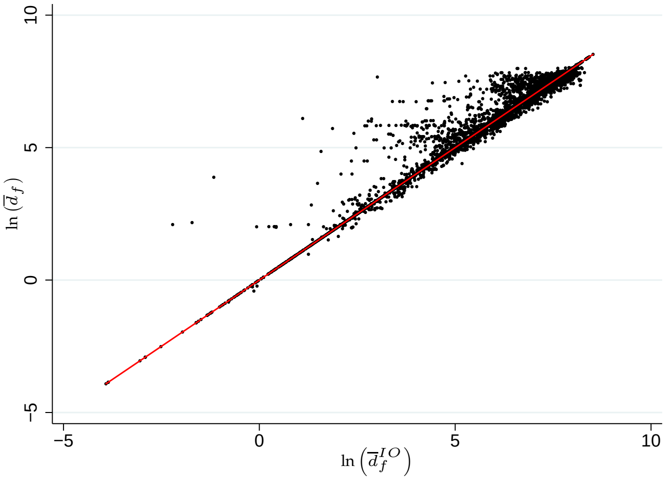
$$\overline{d}^{IO}_f = \frac{1}{\sum_k \sum_{j \neq k} \omega^{IO}_{\Omega(k(f)),\Omega(j(f))}} \sum_k \sum_{j \neq k} \omega^{IO}_{\Omega(k(f)),\Omega(j(f))} d_{k(f),j(f)}, \quad (8)$$

Table 2: Manufacturing industries with the most and the least compact multiunit firms.

NAICS	Industry name	# multiunit firms	$\overline{d}_f^w$	average # plants	$\overline{IO}_f$
	Industries with the most compact multiunit firms				
311230	Breakfast cereal manufacturing	22	153.03	2.00	0.20%
333511	Industrial mould manufacturing	27	162.71	2.22	0.43%
315110	Hosiery and sock mills	18	168.22	2.00	0.45%
315220	Men's and boys' cut and sew clothing manufacturing	50	176.74	2.10	2.47%
324121	Asphalt paving mixture and block manufacturing	100	189.69	3.78	0.42%
327320	Ready-mix concrete manufacturing	330	197.54	6.04	4.58%
325620	Toilet preparation manufacturing	37	205.45	2.57	0.47%
313210	Broad-woven fabric mills	25	210.59	2.40	1.29%
311811	Retail bakeries	31	219.97	2.03	1.61%
325314	Mixed fertilizer manufacturing	70	224.77	3.51	1.43%
	Average internal input-output share				**1.33%**
	Industries with the least compact multiunit firms				
327215	Glass product manufacturing from purchased glass	43	1,588.39	2.70	0.58%
337121	Upholstered household furniture manufacturing	31	1,647.37	2.23	0.38%
337910	Mattress manufacturing	30	1,657.92	2.87	0.25%
334511	Navigational and guidance instruments manufacturing	18	1,718.28	3.22	0.61%
332720	Turned product and screw, nut and bolt manufacturing	41	1,743.38	2.61	0.07%
334290	Other communications equipment manufacturing	39	1,824.67	2.41	0.21%
333210	Sawmill and Woodworking Machinery Manufacturing	20	1,926.28	2.00	0.09%
327420	Gypsum product manufacturing	19	2,308.56	4.11	0.46%
335990	All other electrical equipment and component manufacturing	24	2,421.00	2.50	0.17%
326191	Plastic plumbing fixture manufacturing	18	2,531.45	2.06	0.08%
	Average internal input-output share				**0.29%**

Notes: Results are pooled across all years. We restrict the table to industries with at least 10 observations across the different years. Multiunit firms are allocated to industries based on the most frequent primary NAICS 6-digit code of their constitutent plants. Results using employment are similar.

Figure 4: Unweighted and input-output weighted spatial extent.



measures are highly correlated by construction. Yet, there are a lot of deviations 'to the left' suggesting that the input-output-weighted spatial extent of firms is generally smaller than their unweighted spatial extent. This pattern suggests again that the plants characterized by strong vertical linkages in firms are closer to one another than the plants characterized by weaker vertical linkages, so that vertically more integrated firms are also spatially more compact.

3.3.3 Are multiunit firms geographically 'compact'?

The foregoing descriptive evidence suggests that some multiunit firms are 'compact'. We now investigate whether they are more compact than they 'should be'. Our approach is similar to that of Duranton and Overman (2008, p.227), who test "whether two establishments that belong to the same firm are closer to each other than to any random pair of establishments in the industry." Using UK data, they find that "152 industries (or 71 percent) exhibit localization of establishments that belong to the same firm while 23 (or 11 percent) exhibit dispersion [...] our results strongly suggest that economizing on interaction costs dominates the forces that push toward dispersion in a large majority of industries." Although Duranton and Overman (2008) control for the industries the plants belong to, they cannot control for plant- and firm-level characteristics.[19] Hence, it is not clear whether it is really multiunit status of plants that drives the observed location patterns or some other unobserved characteristics that vary systematically across multiunit plants and standalone plants.

We use a different methodology to show that multiunit firms are *substantially more compact* than randomness would predict. We proceed as follows. First, we construct counterfactual multiunit firms. To this end, we use propensity score matching (PSM). The multiunit plants are the treated units, while the standalone plants are the untreated units. We use one-to-one nearest neighbor matching on a large number of plant-level and geographical controls, stratified at the 3-digit NAICS level (see Appendix C for additional details and for the balance of observables).[20] The underlying logic is to select plants that are comparable both with regards to their own characteristics, and with regards to the characteristics of the locations that they have chosen. Those plants thus are in locations that a priori could have been chosen by the plants of the multiunit firms. Once we have assigned to each multiunit plant a 'comparable' standalone plant, we group those comparable plants into the corresponding counterfactual

where $\omega^{IO}_{\Omega(k(f)),\Omega(j(f))}$ can be either the share of inputs, the share of outputs, or some (weighted) average between the two. Whereas (5) and (6) are symmetric with respect to j and k, because the distances and the weights are symmetric, this is no longer the case for (8) because the input-output shares are not symmetric across industries. In Figure 4, we use the simple average between input and output shares.

[19] Duranton and Overman (2008) also make use of size constraints (size bins for establishments) in another application, but the 'case-control' nature of their methodology unfortunately does not allow to easily control for multidimensional characteristics of plants and firms.

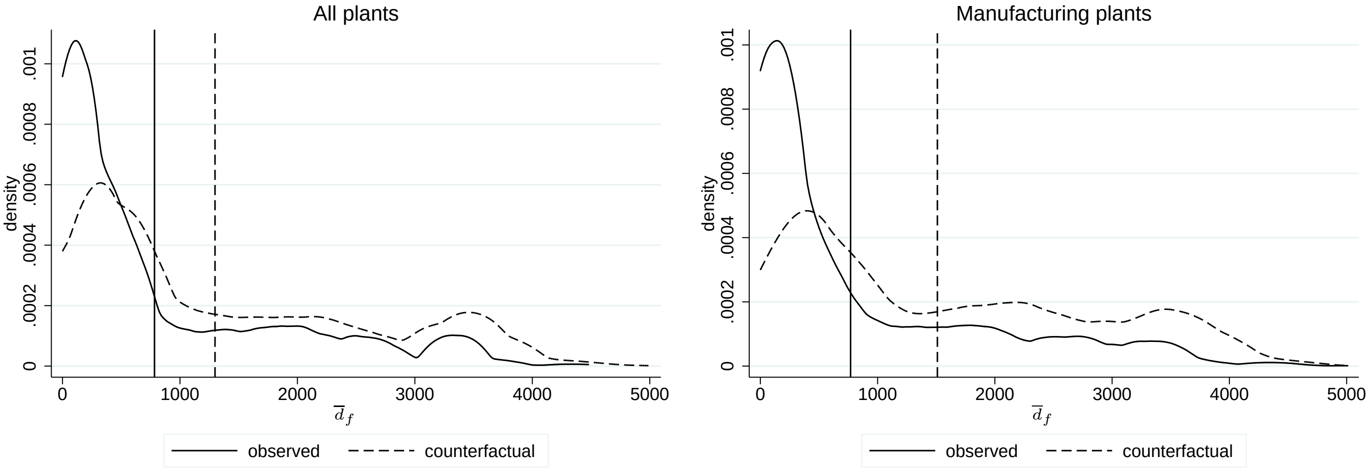
[20] Ideally, we would like to stratify at the 6-digit NAICS level, but this leaves us with a lot of treated plants for which we cannot find adequate controls. This is a classical problem that plagues PSM techniques.

Table 3: Propensity score matching (PSM) results, all plants and all years.

Year	Sample	# multiunit firms	Average $\overline{d}_f$	Std dev	Average PSM $\overline{d}_f$	Std dev PSM	Difference
2001	All industries	4,732	783.21	14.53	1,298.49	17.72	515.27
2003	All industries	4,832	748.75	14.18	1,296.55	17.08	547.80
2005	All industries	4,887	708.95	13.65	1,195.68	16.47	486.73
2007	All industries	5,148	856.35	14.69	1,710.75	17.60	854.40
2009	All industries	5,368	869.00	14.47	1,694.19	17.09	825.19
2011	All industries	5,397	868.03	14.07	1,697.04	16.28	829.00
2013	All industries	5,607	858.04	13.59	1,719.93	16.03	861.88
2001	Manufacturing	1,643	769.57	25.32	1,509.85	31.19	740.28

Notes: Samples are stratified by 3-digit NAICS industries. The balance of observables is provided in Appendix C. We drop industry-year pairs with less than 30 plants and less than 10 multiunit plants. We also drop industries in which multiunit plants account for more than 50% of all plants. We finally drop the 3-digit NAICS industries 488 ('Support activities for transportation') and 491 ('Postal service') in 2007 since the computations do not converge when including our set of controls. 'Difference' stands for the difference between the counterfactual and the observed mean values. All differences are significant at the 1% level.

Figure 5: Distribution of $\overline{d}_f$, observed vs counterfactual (PSM) in 2001.



multiunit firms – which thus consist of the same number of plants in the same industries (at the 3-digit level) than the plants of the 'real' multiunit firms – and recompute the internal distance measures for those counterfactual multiunit firms. The plants making up those firms are similar in terms of their plant-level characteristics and the characteristics of the locations they choose, but differ by their organizational status. If the counterfactual multiunit firms are much less compact than the observed ones – despite their similarity in terms of plant-level and locational characteristics – we may tentatively conclude that some distance-sensitive firm-specific factors (control, knowledge sharing, 'soft' information, vertical links in supply chains, etc.) may drive the clustering of plants within multiunit firms.

Figure 5 depicts the observed and counterfactual distributions of average internal distances for plants in all industries (left panel) and for manufacturing plants only (right panel). As can clearly be seen from that figure, conditional on plant-level and locational characteristics, the observed multiunit firms are much more spatially compact than predicted. This pattern is even stronger for manufacturing multiunit firms, as can be seen from the right panel. Table 3 summarizes the distributions for all years and reports difference-in-means tests. Clearly, observed multiunit firms are much more compact (about 50%) than comparable counterfactual multiunit firms assembled from standalone plants.

4 Empirical methodology and results

The descriptive evidence in the previous section suggests that multiunit plants and standalone plants display different location patterns. It also suggests that multiunit firms have systematic components to their geographical structure. We now dissect this evidence further using multivariate analysis with a large number of controls and different identification strategies.

We run two types of regressions. First, we run *inter-firm linkage* regressions that explore differences in the characteristics of the geographical environment in which multiunit versus standalone plants locate. These regressions allow us to test the prediction that multiunit establishments behave differently than standalone units, in particular with regard to their need to be close to external suppliers.[21] Second, we run *intra-firm linkage* regressions to look into multiunit firms to investigate what types of establishments are more centrally located within firms. In doing so, we test the prediction that stronger involvement in within-firm vertical supply chains should be reflected in the establishments' geographical position in the firm. We find empirical evidence consistent with those different hypotheses.

[21] The scattered empirical evidence on the impact of input-output links on plant-level outcomes has essentially focused on productivity. Whereas Brown and Rigby (2015) find that multiunit plants, older plants, and larger plants benefit more from the local 'upstream supplier density', Greenstone *et al.* (2010) find the opposite result. Focusing on the openings of new large plants, they fail to find evidence that incumbents in industries more strongly linked by input-output relationships benefit from larger productivity gains. Yet, 'labor similarity' and 'technological similarity' of plants is associated with such gains. One possible interpretation of these results is that the latter two effects are not under the control of the firm. While a firm can hardly prevent its workers from switching workplace or information to leak, input-output relationships are one of the variables that are fully under the firm's control. If larger firms do not interact much with their local environment, this should affect mostly the input-output relationships that the firm does (or does not) establish. Alcácer and Chung (2014, p.1757) also find that "small firms locate where supplier agglomeration economies exist, but larger firms do not, a finding at odds with our expectation of lower competitor appropriation risk [for supplier linkages]." This result is, however, consistent with the idea that larger firms have more internal resources.

4.1 Inter-firm linkage regressions

4.1.1 Baseline results

We first run various regressions of the following form:

$$\begin{aligned} \ln(\mathcal{I}dist^{l}_{j(i,f),t}) &= \beta_0 + \gamma_1 \,\text{multiunit}_{j(i,f),t} + \gamma_2 \ln(\text{spec}_{j(i,f),t}) \\ &\quad + \beta_1 \ln(\mathcal{M}dist^{l}_{j(i,f),t}) + \mathbf{X}_{j(i,f),t}\beta_2 + \mathbf{G}_{j(i,f),t}\beta_3 + \zeta_{(i,t)} + \varepsilon_{j(i,f),t} \end{aligned} \tag{9}$$

where $j(i,f)$ designates plant j in industry i and belonging to firm f; where t is the year; and where $l \in \{\text{mgf}, \text{all}\}$ indicates whether the *linkage* and distance variables are constructed using manufacturing industries only or all industries. Unless stated otherwise, we use in what follows count-based specialization measures, $\text{spec}_{j(i,f),t}$, with 5 kilometer distance rings and strict 6-digit industry definitions (see Appendix B.1 for details).

Turning to the variables in equation (9), our primary coefficient of interest is γ_1. We use three different measures for 'multiunit': (i) a multiunit dummy, with value 1 if the establishment belongs to a multiunit firm and 0 otherwise; (ii) the log of the number of establishments $n(f)$ of the firm; and (iii) the log of one plus the employment in the other establishments of firm f, excluding establishment j. We thus use both continuous and discrete information on multiunit status and firm size. Results are robust to that choice. Our second coefficient of interest is γ_2, which captures the link between the geographical specialization of a location and the access to inputs it offers. On top of controlling for various agglomeration effects, this coefficient provides evidence on whether more specialized areas are more vertically disintegrated, as documented by Holmes (1999) and Li and Lu (2009).

Turning to controls, we include the overall establishment density of a location using the minimum distance variable, as explained in Section 3.2. We also include numerous plant- and firm-level controls, $\mathbf{X}_{j(i,f),t}$, and geographical controls, $\mathbf{G}_{j(i,f),t}$, which all enter the model in a linear way unless specified otherwise. The former include the plant's employment, an exporter dummy, a headoffice dummy, a core segment dummy, a measure of the firm's industrial diversification, and plant-level measures of the average frequency of its products and the range of products produced. The latter include the share of highly educated workers within a 15 kilometer radius around the plant, the share of workers in management and business occupations within a 15 kilometer radius, a full set of 'urban type' dummies for the plant's census division, the plant's distance to the nearest US land border crossing, and an employment-weighted measure of occupational employment similarity with the other plants within a 15 kilometer radius, oes^{w}. Appendix A provides more details on the different variables and the way we construct them.

We estimate (9) as a pooled cross section and include industry-year fixed effects $\zeta_{(i,t)}$ unless indicated otherwise.[22] Standard errors are clustered at the firm level across years. There are

[22] Cross-sectional regressions yield qualitatively identical results for all years and are summarized in Supple-

four reasons for that choice. First, plants within the same firm are affected by shocks that hit other parts of the firm. There is ample empirical evidence that internal markets transmit shocks across units of the same firm.[23] Hence, we cannot assume that the errors across establishments within the same firm are uncorrelated. Also, shocks to firms are persistent through time. Hence, there is temporal correlation in shocks, which will be captured by clustering at the firm level in the pooled cross section. Second, some establishment-level data may be misreported at the firm level (e.g., what the firm does, instead of what the establishment does; or the firm employment instead of the establishment employment), which creates correlated reporting errors across establishments of the same firm. Third, and more technically, clustering at the industry level provides too few clusters with too many observations (see, e.g., Wooldridge, 2010). Last, our key variables of interest – the multiunit dummy, the number of establishments, or the firm employment size – are all recorded at the firm level, which thus provides the natural level for clustering in our estimations.

Table 4 reports our benchmark estimates of equation (9). It reveals two key findings.

First, plants affiliated with multiunit firms are systematically farther away from their potential input suppliers. This holds for all three measures of 'multiunit': the dummy, the number of establishments, and the employment at other establishments. The latter two coefficients show that the larger the multiunit firm, the less close its establishments are located to potential input suppliers. This may be because the establishments can source in-house, or because it is easier for them to overcome distance frictions such as transport costs or informational constraints. However, conditional on multiunit status, larger establishments are located closer to their potential input suppliers. In what follows, we will mainly use the multiunit dummy as our variable of choice, but results are similar (albeit slightly less precise) using the other measures. Columns (4)–(6) show that the multiunit coefficients decrease by about half when the full set of geographical controls is included. This suggests that the distance of multiunit establishments to their potential input sources is partly explained by the fact that these establishments value locational characteristics that are themselves correlated with less proximity to input sources. Yet, even conditional on that, the multiunit coefficients remain highly significant. Note also that establishments that belong to the firm's core segment and establishments that produce less ubiquitious products are located on average closer to their potential input sources. While

mental Appendix D, Table 16, for brevity. Following Maksimovic and Phillips (2002), we also computed the simple time average of the cross-sectional coefficients for the multiunit variables across years. This yields values that are very close to the benchmark coefficient estimates reported in column (5) of Table 4.

[23] Lamont (1997) documents, for example, the interconnections of a firm's segments via internal capital markets. He shows that investment expenditures of nonoil subsidiaries of oil companies were negatively affected by the fall in oil prices in 1986, though the profitability of their investment opportunities did not change a priori. Maksimovic and Philips (2002, p.723) document that "demand shocks faced by a segment of a conglomerate firm affect the growth rates of other segments [...] even in the absence of agency costs and financial market imperfections." Shocks are thus correlated across segments and establishments of multiunit firms.

Table 4: Baseline results for the inter-firm linkages regressions.

	Plant traits only			**Plant and location traits**			**Placebo tests**	
							PSM	**Random**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
multiunit dummy	0.189^{a}			0.111^{a}			0.002	0.000*
	(0.019)			(0.014)			(0.005)	[-0.006; 0.008]*
ln(number of plants)		0.064^{a}			0.033^{b}			
		(0.023)			(0.014)			
ln(firm employment)			0.029^{a}			0.013^{a}		
			(0.005)			(0.004)		
ln(plant employment)	-0.058^{a}	-0.054^{a}	-0.057^{a}	-0.019^{a}	-0.017^{a}	-0.018^{a}	-0.015^{a}	
	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	
exporter dummy	-0.064^{a}	-0.065^{a}	-0.065^{a}	0.014^{b}	0.013^{b}	0.013^{b}	0.012^{c}	
	(0.008)	(0.008)	(0.008)	(0.006)	(0.006)	(0.006)	(0.006)	
headoffice dummy	-0.079^{a}	-0.071^{a}	-0.071^{a}	-0.018^{a}	-0.013^{b}	-0.014^{b}	-0.014^{b}	
	(0.009)	(0.009)	(0.009)	(0.007)	(0.007)	(0.007)	(0.007)	
core segment dummy	-0.033	-0.100^{a}	-0.062^{b}	-0.015	-0.058^{a}	-0.044^{b}	-0.088^{a}	
	(0.024)	(0.031)	(0.027)	(0.018)	(0.022)	(0.020)	(0.017)	
ln(firm diversification)	0.018^{c}	0.028^{a}	0.027^{b}	0.001	0.010	0.011	0.025^{a}	
	(0.011)	(0.010)	(0.011)	(0.008)	(0.007)	(0.008)	(0.007)	
ln(plant diversification)	0.027^{a}	0.022^{a}	0.024^{a}	0.005	0.001	0.001	-0.007	
	(0.008)	(0.007)	(0.008)	(0.006)	(0.005)	(0.006)	(0.005)	
ln(average product frequency)	0.024^{a}	0.024^{a}	0.024^{a}	0.006^{b}	0.005^{b}	0.005^{b}	0.005^{c}	
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	
ln(specialization count)				-0.023^{a}	-0.023^{a}	-0.023^{a}	-0.023^{a}	
				(0.003)	(0.003)	(0.003)	(0.003)	
ln($\mathcal{M}dist$)				0.594^{a}	0.595^{a}	0.595^{a}	0.594^{a}	
				(0.003)	(0.003)	(0.003)	(0.003)	
ln(oes^{w})				0.153^{a}	0.152^{a}	0.152^{a}	0.151^{a}	
				(0.018)	(0.018)	(0.018)	(0.018)	
ln(share of highly educated)				0.231^{a}	0.231^{a}	0.231^{a}	0.231^{a}	
				(0.010)	(0.010)	(0.010)	(0.010)	
ln(share of 'BM' occupations)				0.353^{a}	0.354^{a}	0.354^{a}	0.355^{a}	
				(0.020)	(0.020)	(0.020)	(0.020)	
ln(minimum distance to US)				-0.032^{a}	-0.032^{a}	-0.032^{a}	-0.032^{a}	
				(0.004)	(0.004)	(0.004)	(0.004)	
'urban strong' dummy				-0.217^{a}	-0.217^{a}	-0.217^{a}	-0.218^{a}	
				(0.010)	(0.010)	(0.010)	(0.010)	
'urban moderate' dummy				-0.020^{b}	-0.021^{b}	-0.021^{b}	-0.021^{b}	
				(0.010)	(0.010)	(0.010)	(0.010)	
'urban weak' dummy				0.184^{a}	0.184^{a}	0.184^{a}	0.183^{a}	
				(0.011)	(0.011)	(0.011)	(0.011)	
'rural' dummy				0.066^{a}	0.066^{a}	0.065^{a}	0.065^{a}	
				(0.010)	(0.010)	(0.010)	(0.010)	
Observations	321,528	321,528	321,528	319,319	319,319	319,319	319,092	321,528
R-squared	0.217	0.216	0.217	0.551	0.551	0.551	0.551	

Notes: Pooled cross section results for manufacturing plants for the years 2001, 2003, 2005, 2007, 2009, 2011, and 2013. The dependent variable in all regressions is ln($\mathcal{I}dist$), and it is constructed using the $N = 5$ nearest neighbors in each industry. The specialization variable is constructed at the 6-digit level and for a radius of 5 kilometers, using only the plants' primary NAICS codes. The education and occupation variables are the share of population with higher education ('some college') and the share of population in 'business and management' occupations within a 15 kilometer radius around the plant (distance is measured to the centroid of the census dissemination areas). Distance to the US is the great circle distance to the nearest land border crossing. The occupational employment similarity measure, oes^{w}, is weighted by plants' employment. The excluded category for the 'urban type' dummies is 'urban CMA'. All regressions include a full set of industry-year fixed effects. Robust standard errors, clustered by firm across years, in parentheses. *Specifications: **(7)** Placebo regressions where multiunit status is assigned to standalone plants identified using propensity score matching; **(8)** Placebo regressions where multiunit status is assigned randomly, 100 replications (we report the mean and 95% confidence bands for the distribution of the 'multiplant dummy' coefficient). Significance levels: ^{a}: $p < 0.01$, ^{b}: $p < 0.05$, ^{c}: $p < 0.1$.

we do not want to read too much out of those coefficients, they suggest that more central operations in the firm and rarer products (which may be more complex) may require closer geographical connections with potential input suppliers.

Second, as can be seen from Table 4, plants located in more specialized areas (as measured by the share of plants in the same industry) are also plants that benefit from better inter-firm linkages. As the latter are constructed by excluding the industry of the plants itself, our measure does not just pick up own-sector concentration. This result is in line with findings by Holmes (1999) and Li and Lu (2009), who document for the case of the US and of China, respectively, that plants located in more specialized areas are more vertically disintegrated.[24] We confirm their findings using more fine-grained spatial data, a large set of plant-level and geographical controls, and a quite different microgeographic definition of input linkages. Doubling the measure of specialization around a plant is associated with a 2.3% reduction in our input-distance measure. The mean for the log of our count-based specialization measure is 0.057 (looking at manufacturing plants only), with standard deviation of 0.159. Hence, a one standard deviation increase in the specialization measure from the sample mean is associated with a 8.7% reduction in the distance to a dollar of inputs. At the sample mean of 178 kilometers, this equals a reduction of about 15.5 kilometers in the distance to a dollar of inputs.[25]

Finally, the last two columns of Table 4 report the results of two placebo tests. First, we run a 'standard' placebo in column (8), where we randomly reallocate the multiunit dummy across plants, keeping all other characteristics unchanged. As one can see, the counterfactual distribution of the multiunit dummy (using 100 replications) is tightly centered on zero. The other coefficients are not substantially affected (not reported here). While this test is informative, it has the drawback of (i) also reshuffling the multiunit dummy also across multiunit plants, and (ii) to not control for any other characteristics. We hence report in column (7) a stronger placebo test, where we associate the multiunit dummy with the standalone plants identified by the propensity score matching procedure of Section 3.3.3.[26] As can be seen, even using plants that are observationally similar, the multiunit dummy is not significant whereas the other coefficients remain fairly stable. Hence, it is unlikley that our results are spurious and driven by

[24] Purchased inputs as a percent of the value of outputs – the 'purchased input intensity' – is higher for plants in US manufacturing industries that are geographically more concentrated (Holmes, 1999). This finding suggests that input-output linkages may drive industry localization. However, reverse causality cannot be ruled out, i.e., plants in industries that concentrate geographically for unobserved reasons may vertically disintegrate more because of that concentration. Furthermore, the purchased input intensity does not exclude purchases from the own industry, thereby making reverse causality problems more acute.

[25] Table 17 in Supplemental Appendix D.3 reports estimates for all the specialization measures that we construct: 4- and 6-digit NAICS industries, with 5 or 15 kilometer radius, for strict and extended definitions of industries and for both establishment counts and employment numbers. The coefficient on multiunit status is robust to all the specialization measures that we use.

[26] The slightly smaller number of observations is due to the fact that the same control can serve several times for a treated unit, thus reducing the sample size.

unobserved plant- or geographical characteristics.

4.1.2 Robustness checks

Table 5 reports a large number of robustness checks. Columns (1)–(3) show that the results are largely unchanged if we compute the input- and output-linkages using the seven nearest plants in each industry. Column (4) uses $\ln(\mathcal{I}dist/\mathcal{M}dist)$ as the dependent variable, i.e., we impose a unitary coefficient on the minimum distance variable. Columns (5)–(7) report results where the input distance variable is constructed using road distance estimates instead of great circle distances (see Appendix A.3 for details on how we map great circle distances onto road distances). Given the overall high correlation of road distances and great circle distances, the results barely change. Columns (8)–(10) report results where all industries in the same 4-digit industry of the plant are excluded when computing the input-distance measures. In those regressions, we also include geographical specialization at the same 4-digit level for consistency. As one can see, the specialization variables lose significance, whereas the multiunit variables remain highly significant. This suggests that better access to inputs in more specialized areas is essentially driven by the clustering of related industries within the same 4-digit classification. Those industries also tend to buy a substantial share of inputs from each other. Columns (11)–(13) show that our results go through when we compute the input linkages using the full input-output tables with 864 industries instead of only the manufacturing sub-tables with 242 industries. In columns (14)–(16), we use the full sample of plants, manufacturing and non-manufacturing. To highlight potential differences in location patterns, we interact a manufacturing dummy with selected key variables. As can be seen from the top part of Table 5, the key results for the multiunit variables hold also in the extended sample, with little difference between manufacturing and non-manufacturing establishments. Last, column (17) reports results using a version of the input linkage measure that is computed by excluding plants *within the same firm* as potential suppliers. One may indeed worry that, since multiunit firms are compact, our access measure picks up partly a within-firm effect. As can be seen, the results are virtually unchanged, thus showing that the input distance measure is not driven by within-firm location patterns.

Table 18 in Appendix D.4 reports a variety of additional robustness checks with respect to clustering, sample selection, and specifications. Our key results are robust to the trimming of the dependent and independent variables, the inclusion or exclusion of zeros, and the estimation in levels versus the estimation in logs: establishments affiliated with multiunit firms are farther away from their potential input sources or output destinations. Furthermore, more specialized areas (in terms of counts) provide generally better access to potential input sources, but this effect tends to be driven by the clustering of related 4-digit industries (see Table 5) and weakens when specialization is measured using employment instead of plant counts (see

Table 5: Robustness checks for the inter-firm linkages regressions.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
	$N = 7$	$N = 7$	$N = 7$	Unitary	Road	Road	Road	Excl4	Excl4	Excl4	All links	All links	All links	All estab	All estab	All estab	Excl Firm
multiunit dummy	0.073^a			0.058^a	0.109^a			0.112^a			0.084^a			0.097^a			0.112^a
	(0.012)			(0.016)	(0.014)			(0.015)			(0.012)			(0.010)			(0.014)
mfg × multiunit dummy														-0.022			
														(0.016)			
ln(number of plants)		0.017				0.032^b			0.035^b			0.021^c			0.025^a		
		(0.012)				(0.013)			(0.014)			(0.012)			(0.009)		
mfg × ln(number of plants)															-0.007		
															(0.013)		
ln(firm employment)			0.007^b				0.013^a			0.014^a			0.010^a			0.018^a	
			(0.003)				(0.004)			(0.004)			(0.003)			(0.003)	
mfg × ln(firm employment)																-0.010^b	
																(0.004)	
core segment dummy	-0.006	-0.039^b	-0.031^c	0.006	-0.014	-0.056^a	-0.043^b	-0.025	-0.067^a	-0.053^a	-0.005	-0.041^b	-0.028	-0.057^a	-0.096^a	-0.061^a	-0.016
	(0.016)	(0.019)	(0.018)	(0.021)	(0.018)	(0.021)	(0.019)	(0.018)	(0.022)	(0.020)	(0.016)	(0.019)	(0.018)	(0.014)	(0.014)	(0.015)	(0.018)
ln(average product frequency)	0.004	0.003	0.003	-0.002	0.005^b	0.005^b	0.005^b	0.007^b	0.006^b	0.006^b	0.004^b	0.004^c	0.004^c	-0.006^a	-0.006^a	-0.006^a	0.006^a
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
ln(specialization count)	-0.043^a	-0.043^a	-0.043^a	-0.138^a	-0.022^a	-0.022^a	-0.022^a				-0.044^a	-0.044^a	-0.044^a	0.024^a	0.024^a	0.024^a	-0.023^a
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)				(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
ln(specialization count, NAICS4)								-0.001	-0.001	-0.001							
								(0.003)	(0.003)	(0.003)							
mfg × core segment dummy														0.043^b	0.048^b	0.017	
														(0.020)	(0.023)	(0.023)	
mfg × ln(average product frequency)														0.010^a	0.010^a	0.011^a	
														(0.003)	(0.003)	(0.003)	
mfg × ln(specialization count)														-0.047^a	-0.047^a	-0.047^a	
														(0.002)	(0.002)	(0.002)	
Observations	319,319	319,319	319,319	319,319	319,319	319,319	319,319	319,319	319,319	319,319	319,319	319,319	319,319	901,324	901,324	901,324	319,319
R-squared	0.633	0.633	0.633	0.459	0.559	0.559	0.559	0.550	0.549	0.549	0.563	0.563	0.563	0.587	0.586	0.586	0.551

Notes: Pooled cross section results for manufacturing and non-manufacturing plants for the years 2001, 2003, 2005, 2007, 2009, 2011, and 2013. All regressions include a full set of industry-year fixed effects. We report selected variables only. All specifications include the following plant- and firm-level controls: log of plant employment, exporter dummy, headoffice dummy, plant-level diversity measure, firm-level diversity measure. We also include a full set of geographical controls: minimum distance to the US, employment-weighted occupational employment similarity, oes^w, share of highly educated ('some college') within a 15 kilometer distance, share of workers in 'management or business' occupations within a 15 kilometer distance, and a full set of 'urban type' dummies. Robust standard errors, clustered by firm across years, in parentheses. **Specifications:** **(1)** to **(3)**: $\mathcal{I}dist$ is computed using the $N = 7$ nearst plants in each industry; **(4)** dependent variables is $\ln(\mathcal{I}dist/\mathcal{M}dist)$, i.e., we impose a unitary coefficient on $\mathcal{M}dist$; **(5)** to **(7)**: distances are computed using road distance estimates obtained from fitting a polynomial using Google API map data (see Appendix A for details); **(8)** to **(10)**: $\mathcal{I}dist$ and $\mathcal{O}dist$ are computed excluding plants in the establishment's own 4-digit industry; **(11)** to **(13)**: $\mathcal{I}dist$ and $\mathcal{O}dist$ are computed using all 864 NAICS industries (not only the 242 manufacturing industries); **(14)** to **(16)**: regressions include all plants (not only manufacturing); **(17)** $\mathcal{I}dist$ is computed excluding nearest plants that belong to the same multiunit firm. Significance levels: ^{a}: $p < 0.01$, ^{b}: $p < 0.05$, ^{c}: $p < 0.1$.

Table 17 in Appendix D.2).

4.1.3 Mitigating possible identification issues

Although the use of detailed plant-level data and the numerous disaggregated controls we include should address many potential identification issues, we now provide a number of additional checks. The most severe problem we face is that our input- or output-distance measures may depend on the presence of either old and/or 'locally large plants' that have been present for a long enough time and/or that have sufficient economic weight to structure regional or local supply chains around them.[27] The idea is that large 'million dollar plants' may actually be *drivers* of the structure of industries and supply chains around them, rather than pick locations based on a preexisting structure. This in turn may drive the clustering of the industry around those focal points. Thus, the causality would run the other way round: large plants pick specific locations for unobserved reasons, and since they are large those locations develop the support industries. Dropping old and/or large establishments to focus on younger and/or smaller establishments reduces the potential for this reverse causality. Observe that the presence of this 'local supply chain structuring' phenomenon would, however, downward bias our estimate of the multiunit coefficient. The reason is that (manufacturing) multiunit establishments are on average larger (45.1 employees versus 24.9 employees) and older (45 years versus 27.8 years) than standalone establishments in our sample. Hence, our results should be even stronger if older and/or locally large plants have the anticipated structuring effect on local economic activity.

To deal with this potential problem, we estimate the model by: (i) focusing on young plants only; and (ii) by dropping all locally large or isolated plants. We define young plants as establishments for which the difference between the current year and the 'year established' variable in our data is less than either 2, 3, or 4 years. We define locally large plants as establishments that are in the top 5% of the distribution of the size of the establishment relative to the size of its own industry in either a 5 or a 15 kilometer radius. For example, a plant that counts 100 employees and is located in a region where there are only 5 other employees in the same industry (excluding the plant itself) would probably be 'locally large'. Last, we define isolated plants in a very intuitive way as plants that are the only employers in their industry in either a 5 or a 15 kilometer radius. By definition, isolated plants are obviously an extreme

[27] Greenstone *et al*. (2010) document that large 'million dollar plants' increase the TFP of incumbent firms in regions where they locate. Though they do not provide evidence for how these plants geographically structure supply chains, it is likely that such plants have an impact on those chains. Klier and McMillen (2008, p.261) find that the changes in the US automotive industry during the 1990s, especially the advent of just-in-time logistics, implied that a number of "supplier functions must be performed in very close proximity to the assembly location. In a number of cases, this tendency has led to the construction of a supplier park immediately adjactent to an assembly plant."

Table 6: Inter-firm linkage regressions for young establishments and when dropping 'locally large plants'.

	Young plants only						Dropping locally large and isolated plants									
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**	**(13)**	**(14)**	**(15)**	**(16)**
	All	All	All	All	All	All	Mfg	Mfg	Mfg	Mfg	Mfg	Mfg	All	All	All	All
	≤ 2	≤ 3	≤ 4	≤ 2	≤ 3	≤ 4	LLP 5	LLP 15	LLP 5	LLP 15	isolated5	isolated15	LLP 5	LLP 15	LLP 5	LLP 15
multiunit dummy	0.247^b	0.091	0.062	0.219^b	0.122^b	0.087^b	0.124^a	0.139^a	0.125^a	0.139^a	0.124^a	0.141^a	0.091^a	0.097^a	0.092^a	0.098^a
	(0.113)	(0.065)	(0.044)	(0.089)	(0.054)	(0.037)	(0.018)	(0.017)	(0.018)	(0.017)	(0.018)	(0.017)	(0.011)	(0.011)	(0.011)	(0.011)
mfg × multiunit dummy	-0.118	0.025	-0.093	0.132	0.025	-0.095							0.003	0.008	0.009	0.011
	(0.310)	(0.152)	(0.088)	(0.214)	(0.119)	(0.070)							(0.018)	(0.017)	(0.018)	(0.017)
core segment dummy	0.233	0.005	-0.001	-0.002	-0.071	-0.038	-0.053^b	-0.028	-0.053^b	-0.027	-0.057^b	-0.030	-0.078^a	-0.072^a	-0.077^a	-0.072^a
	(0.183)	(0.109)	(0.075)	(0.133)	(0.086)	(0.061)	(0.025)	(0.022)	(0.025)	(0.022)	(0.025)	(0.023)	(0.015)	(0.014)	(0.015)	(0.014)
mfg × core segment dummy	-0.133	-0.042	0.025	0.413	0.082	0.048							0.032	0.042^c	0.038	0.047^b
	(0.475)	(0.266)	(0.148)	(0.309)	(0.201)	(0.130)							(0.025)	(0.024)	(0.025)	(0.024)
ln(average product frequency)	0.029	0.011	0.010	0.021	0.016	0.019^b	0.008^b	0.006^b	0.008^b	0.006^b	0.008^b	0.007^b	-0.006^a	-0.008^a	-0.005^b	-0.007^a
	(0.023)	(0.014)	(0.009)	(0.017)	(0.011)	(0.008)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
mfg × ln(average product frequency)	0.034	0.024	0.013	0.012	0.041^b	0.014							0.010^a	0.011^a	0.008^b	0.009^a
	(0.053)	(0.024)	(0.015)	(0.033)	(0.019)	(0.012)							(0.003)	(0.003)	(0.003)	(0.003)
ln(specialization count, 5km)	0.048	0.039^b	0.034^a	0.021	0.030^b	0.030^a	-0.031^a				-0.032^a	-0.028^a	0.049^a			
	(0.031)	(0.017)	(0.011)	(0.023)	(0.013)	(0.009)	(0.004)				(0.004)	(0.003)	(0.003)			
ln(specialization employment, 5km)									-0.012^a	0.003^b					0.052^a	0.044^a
									(0.003)	(0.002)					(0.002)	(0.001)
ln(specialization count, 15km)								0.031^a						0.102^a		
								(0.005)						(0.003)		
mfg × ln(specialization count/emp)	-0.113^c	-0.045^c	-0.025	-0.045	-0.039^b	-0.050^a							-0.104^a	-0.118^a	-0.088^a	-0.057^a
	(0.059)	(0.026)	(0.015)	(0.038)	(0.018)	(0.012)							(0.004)	(0.004)	(0.003)	(0.002)
Industry-year fixed effects	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clustering	Firm	Firm	Firm	No	No	No	Firm	Firm	Firm	Firm	Firm	Firm	Firm	Firm	Firm	Firm
Observations	1,997	5,129	10,319	1,997	5,129	10,319	191,188	238,150	191,188	238,150	185,415	231,071	632,614	729,317	632,614	729,317
R-squared	0.820	0.738	0.700	0.641	0.586	0.584	0.545	0.523	0.544	0.523	0.546	0.523	0.591	0.579	0.592	0.578

Notes: Pooled cross section results for manufacturing and non-manufacturing plants for the years 2001, 2003, 2005, 2007, 2009, 2011, and 2013. We report selected variables only. All specifications include the following establishment- and firm-level controls: log of plant employment, exporter dummy, headoffice dummy, plant-level diversity measure, firm-level diversity measure. We also include a full set of geographical controls: minimum distance to the US, employment-weighted occupational employment similarity, oes^w, share of highly educated within a15 kilometer distance, share of workers in management or business occupations within 5 kilometer distance, and a full set of 'urban type' dummies. **Specifications: (1)** to **(3)**: Results for young plants, with industry-year fixed effects and firm-level clustered standard errors; **(4)** to **(6)**: separate industry and year fixed effects, and heteroskedasticity-robust standard errors; **(7)** to **(16)**: dropping locally large plants (LLP at either 5 or 15 kilometer distance, or dropping isolated plants at either 5 or 15 kilometer distance. Significance levels: ^{a}: $p < 0.01$, ^{b}: $p < 0.05$, ^{c}: $p < 0.1$.

form of locally large plants.

Table 6 summarizes our results. Columns (1)–(3) provide results for young plants. Because of sample size considerations, we report results pooled across manufacturing and non-manufacturing establishments.[28] As can be seen, there is a significant and positive coefficient on the multiunit variable for establishments of less than two years of age. The coefficients for 3 or 4 years of age are positive, yet insignificant. When using separate industry and year fixed effects, these coefficients turn positive and significant, as can be seen from columns (4)–(6).[29] Turning to locally large and isolated plants, the results are reported in columns (7)–(16). In all specifications, the qualitative results are the same as in our baseline regressions in Table 4, irrespective of the geographical scope (5 or 15 kilometers) or the stringency (locally large or isolated) with which we define the sample. It is worth pointing out that the coefficients become all marginally larger, which would be in line with the downward bias due to the supply chain structuring around older and/or larger establishments.[30] To summarize, establishments affiliated with multiunit firms are located in areas that provide significantly worse access to external input suppliers. Our estimates for 'young plants' also suggest that multiunit firms can expand into areas that offer worse conditions, possibly because their establishments are less dependent on these conditions due to the presence of internal resources and markets.[31]

[28] There are relatively few new manufacturing establishments in our database. One of the reasons may be the strong de-industrialization trend in Canada since the early 2000s. Indeed, the number of manufacturing establishments decreased from about 52,000 in 2001 to about 35,000 in 2013 in our data. There are even less new manufacturing establishments affiliated with multiunit firms. If anything, these firms have been closing and relocating establishments more easily abroad than other firms (see, e.g., Bernard and Jensen, 2007).

[29] Brown and Rigby (2015) find that green entrants experience less productivity benefits than entrants to incumbent firms with respect to the 'local density of upstream suppliers', especially for plants born in the 80s. Both their dependent variable and their measure of input-linkages are, however, very different from ours.

[30] Since multiunit plants are larger and older, they could get easily stuck at a particular location if moving or switching costs are higher for bigger and older plants. This could lead to an upward bias in our coefficients. However, Bernard and Jensen (1999) find that, conditional on individual characteristics, multiunit plants are more – not less – likely to exit than comparable standalone plants. One of the reasons may be the existence of internal markets that allow for a better redeployment of resources from closing plants.

[31] When firms add establishments in 'new' industries, they may locate them in places where there is better access to inputs and outputs or agglomeration economies. On the opposite, firms that add establishments in their 'core' industry may be able to draw on more internal resources and locate them in more remote places (see, e.g., Chinitz, 1961). As can be seen from Table 6, we do not find significant effects in our estimations for the importance of access to inputs for new plants belonging to core segments. The 'core segment dummy' and its interactions are insignificant in all estimations.

4.2 Intra-firm linkage regressions

4.2.1 Baseline results

We now look inside multiunit firms to analyze whether some establishments – especially those more likely to be involved in vertical production chains – are geographically more centrally located *within* the firm. To this end, we run various regressions of the following form:

$$\begin{aligned}\ln(\overline{d}_{j(i,f),t}) &= \alpha_0 + \gamma_1 \ln(\overline{IO}_{j(i,f),t}) + \gamma_2 \ln\big(\text{spec}_{j(i,f),t}\big) \\ &\quad + \mathbf{X}_{j(i,f),t}\alpha_2 + \mathbf{G}_{j(i,f),t}\alpha_2 + \zeta_{(i,t)} + \varepsilon_{j(i,f),t}\end{aligned} \tag{10}$$

where the notation is the same as in (9). We use both unweighted and employment-weighted internal distance measures in our regressions. Since there is no concept of 'internal distance' for single-plant firms, the estimation is restricted to plants affiliated with multiunit firms.

Building on the findings from Table 2 and Figure 4, our key variable of interest is the strength of internal vertical linkages, $\ln(\overline{IO}_{j(i,f),t})$, between plant j and the other plants of the firm. If vertical linkages within firms are important and distance sensitive, we should expect that plants that are more strongly vertically linked are closer to the other plants of the firm, ceteris paribus. We control for a large number of plant- and firm-level characteristics, subsumed by **X**. These include the number of plants of the firm (more plants mechanically increase the spatial extent of the firm); a headoffice dummy (see Aarland *et al.*, 2007); the log of plant employment; an exporter dummy; a core segment dummy; our firm diversity measure; the plant-level measure of diversity; and the plant-level measure of product ubiquity. Since the diagonal elements of the input-output matrix are large – and since those elements are hard to interpret in terms of vertical versus horizontal linkages (see, e.g., Raimondo *et al.*, 2014) – we include a 'monoindustry dummy' that takes value 1 if all plants of the firm are active in the same primary NAICS sector. In some specifications, we also include the same geographical controls, **G**, as before: the weighted occupational employment similarity measure, oes^w; the share of highly educated within a 15 kilometer radius; the share of workers in management or business occupations within a 15 kilometer radius; the minimum distance to the US; and a full set of 'urban type' dummies. As before, all controls enter the model in a linear way unless specified otherwise. The regressions also include a full set of industry-year fixed effects, $\zeta_{(i,t)}$, and standard errors are clustered by firm across years.

Table 7 summarizes the results for our baseline estimation of equation (10). As one can see across all specifications, firms with more establishments are on average less compact geographically. This may just reflect 'spacing out', or the fact that firms with more plants need less proximity since they can coordinate more easily their activities. Evidence for this is provided in Aarland *et al.* (2007) and Henderson and Ono (2008), who show that larger firms are much more likely to establish separate 'central administrative offices' (CAO) that serve to coordinate a number of functions within the firm. Since establishments delegate a number

Table 7: Baseline results for intra-firm linkage regressions, and relation between intra- and inter-firm linkages.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**	**(13)**	**(14)**	**(15)**	**(16)**
		w		**w**		**w**		**w**		**w**		**w**		**w**		**w**
ln(number of plants)	0.735^{a}	0.731^{a}	0.743^{a}	0.740^{a}	0.453^{a}	0.466^{a}	0.744^{a}	0.740^{a}	0.741^{a}	0.738^{a}	0.746^{a}	0.742^{a}	0.743^{a}	0.740^{a}	0.740^{a}	0.737^{a}
	(0.063)	(0.056)	(0.062)	(0.056)	(0.030)	(0.030)	(0.058)	(0.052)	(0.058)	(0.052)	(0.058)	(0.052)	(0.058)	(0.052)	(0.057)	(0.051)
headoffice dummy	-0.272^{a}	-0.271^{a}	-0.258^{a}	-0.256^{a}	-0.269^{a}	-0.280^{a}	-0.220^{a}	-0.216^{a}	-0.216^{a}	-0.211^{a}	-0.219^{a}	-0.215^{a}	-0.218^{a}	-0.214^{a}	-0.218^{a}	-0.214^{a}
	(0.041)	(0.041)	(0.042)	(0.042)	(0.023)	(0.023)	(0.040)	(0.040)	(0.040)	(0.041)	(0.040)	(0.041)	(0.040)	(0.041)	(0.040)	(0.041)
ln(specialization count)	-0.051^{a}	-0.046^{b}	-0.066^{a}	-0.065^{a}	-0.021	0.022	-0.149^{a}	-0.149^{a}	-0.149^{a}	-0.149^{a}	-0.148^{a}	-0.149^{a}	-0.151^{a}	-0.151^{a}	-0.149^{a}	-0.150^{a}
	(0.018)	(0.018)	(0.022)	(0.023)	(0.013)	(0.015)	(0.017)	(0.018)	(0.017)	(0.018)	(0.017)	(0.018)	(0.017)	(0.018)	(0.017)	(0.018)
ln($\overline{IO}_{j(i,f),t}$)	-0.081^{a}	-0.093^{a}	-0.080^{a}	-0.092^{a}	0.053	0.052	-0.077^{a}	-0.088^{a}	-0.079^{a}	-0.090^{a}	-0.078^{a}	-0.089^{a}	-0.077^{a}	-0.088^{a}	-0.077^{a}	-0.088^{a}
	(0.026)	(0.026)	(0.026)	(0.026)	(0.033)	(0.033)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
mfg × ln($\overline{IO}_{j(i,f),t}$)					-0.125^{a}	-0.126^{a}										
					(0.039)	(0.038)										
mfg × log(number of plants)					0.304^{a}	0.295^{a}										
					(0.049)	(0.050)										
mfg × log(specialization count)					-0.031	-0.028										
					(0.021)	(0.021)										
ln($\mathcal{I}dist/\mathcal{M}dist$)						0.108^{a}										
						(0.012)										
ln($\mathcal{I}dist_x^{\mathrm{mfg}}$), excl. same-firm plants							0.616^{a}	0.675^{a}	0.544^{a}	0.576^{a}	0.487^{a}	0.531^{a}	0.628^{a}	0.666^{a}	0.604^{a}	0.680^{a}
							(0.044)	(0.045)	(0.030)	(0.031)	(0.073)	(0.074)	(0.044)	(0.044)	(0.080)	(0.083)
core segment dummy × ln($\mathcal{I}dist_x^{\mathrm{mfg}}$)							-0.106^{b}	-0.143^{a}							-0.112^{a}	-0.150^{a}
							(0.042)	(0.043)							(0.042)	(0.043)
ad valor transport costs × ln($\mathcal{I}dist_x^{\mathrm{mfg}}$)									-0.140	-0.143					-0.276	-0.273
									(0.177)	(0.176)					(0.174)	(0.175)
share of non-homog. inputs × ln($\mathcal{I}dist_x^{\mathrm{mfg}}$)											-0.015	-0.012			-0.035	-0.031
											(0.019)	(0.019)			(0.021)	(0.021)
mfg durables dummy × ln($\mathcal{I}dist_x^{\mathrm{mfg}}$)													-0.136^{b}	-0.146^{a}	-0.169^{a}	-0.178^{a}
													(0.055)	(0.056)	(0.060)	(0.061)
Geographical controls	No	No	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No
Observations	31,887	31,887	31,629	31,629	131,559	131,559	31,887	31,887	31,776	31,776	31,887	31,887	31,887	31,887	31,776	31,776
R-squared	0.265	0.253	0.269	0.257	0.373	0.376	0.314	0.308	0.313	0.306	0.314	0.307	0.315	0.308	0.314	0.308

Notes: Pooled cross section results for manufacturing plants for the years 2001, 2003, 2005, 2007, 2009, 2011, and 2013. The dependent variables in the regressions are either $\ln(\overline{d}_{j(i,f),t})$ or $\ln(\overline{d}^{w}_{j(i,f),t})$ (weighted, **w**). The specialization variable is constructed at the 6-digit level and for a radius of 5 kilometers, using only the plants' primary NAICS codes. The education and occupation variables are the share of population with higher education (some college) and the share of population in business and management occupations within a 15 kilometer radius around the plant (where distance is measured to the centroid of the census dissemination areas). Distance to the US is the great circle distance to the nearest land border crossing. The occupational employment similarity measure, oes^{w}, is weighted by plants' employment. All regressions include a full set of industry-year fixed effects. We report selected variables only. All specifications include the following controls: log of plant employment, exporter dummy, core segment dummy, firm diversity measure, plant-level number of products reported, plant-level product frequency measure, monoindustry dummy. **Specifications: (1)** and **(2)**: baseline specifications, with the average input-output measure; **(3)** and **(4)**: replicate (1) and (2) with a full set of geographical controls (weighted occupational employment similarity, share of highly educated within a 15 kilometer distance, share of workers in management or business occupations within a15 kilometer distance, minimum distance to US land border crossing, and urban type dummies; the excluded category for the 'urban type' dummies is 'urban CMA'); **(5)** and **(6)**: results for all plants, including interactions with a manufacturing dummy ('mfg'); **(7)–(16)**: includes the input distance, computed excluding same-firm plants, interacted with core segment dummy, ad valorem transport costs, Nunn's (2007) measure of product differentiation, and durables dummy. Robust standard errors, clustered by firm across years, in parentheses. Significance levels: ^{a}: $p < 0.01$, ^{b}: $p < 0.05$, ^{c}: $p < 0.1$.

of those functions to the CAO, they are also de facto more footloose. Turning to headoffices, we see that they are geographically more centrally located within the firm, as expected. This confirms previous findings by Aarland *et al.* (2007) and Henderson and Ono (2008), who have shown, using US manufacturing data, that headoffices are 'centrally located' with respect to production plants.[32] The specialization variable remains consistently negative for manufacturing establishments, thus showing that plants located in more specialized areas are also more geographically central in the firm. This finding is consistent with the idea that manufacturing firms structure themselves – and especially their core activities – around specialized areas.

Turning to our key variable of interest, all coefficients for manufacturing plants associated with the average internal input-output strength, $\ln(\overline{IO}_{j(i,f),t})$, are negative and highly significant. In words, plants in industries that are more strongly linked to the other industries of the plants in the firm are geographically more centrally located.[33] This is consistent with the case in which downstream plants interact with several upstream plants, but where upstream plants do not interact much with each other. Observe that while the internal input-output strength is highly significant for manufacturing establishments, it vanishes for non-manufacturing establishments (see columns (5) and (6) of Table 7). This suggests that proximity of plants that are strongly linked in terms of inputs and outputs is important for the geographical structure of manufacturing firms, but not for the geographical structure of non-manufacturing firms. Our result is reminiscent of that by Kolko (2010), who finds that the effect size of input-output links for service industry pairs (he looks at coagglomeration) is much smaller than that for manufacturing. Hence, input-output effects, and their impact on the geography of the firm, may be driven by the exchange of tangible goods in supply chains. If the effect was solely one of transfer of intangible resources (as argued, e.g., by Atalay *et al.*, 2014) or the ease of coordination of related activities, we would expect this effect to not be solely limited to manufacturing but to cut broadly across industries.

[32] Our headoffices are not necessarily physically separated from manufacturing plants, i.e., they can be embedded. Focusing on separate headoffices and the lodging industry, Kalnins and Lafontaine (2014) document that establishment survival probability and duration, as well as profitability, are all decreasing functions of distance to the headquarter, except for the largest establishment owners in their dataset.

[33] This result holds also when we replace $\ln(\overline{IO}_{j(i,f),t})$ with $\ln(\overline{I}_{j(i,f),t})$, using only input linkages. However, the coefficients on output linkages $\ln(\overline{O}_{j(i,f),t})$ are not significant. Hence, it seems that downstream plants tend to be located more closely to the other plants in manufacturing firms, whereas upstream plants are less centrally located. We have furthermore verified that the results are not driven by monoindustry firms that have larger input-output values because most transactions are within industries. In addition to controlling with a dummy, we have also excluded all monoindustry firms from the estimation. The coefficient on input-output links is -0.075 and significant at 1%. Last, we also ran the regression including a full set of economic region fixed effects. The coefficient on the internal input-output strength variable is -0.078 and significant at 1%.

4.2.2 Plant types and within-firm location choices

Theory does not say much on the correlation between access to external suppliers and access to internal supplies. The former may act as a centripetal force, whereas the latter may act as a centrifugal force for the firm (Alcácer and Delgado, 2013; see also the model in Supplemental Appendix E). Multiplant firms may want to be compact to reduce coordination costs, which drives them away from input sources or specialized locations; or they may want to be less compact to be closer to inputs or some other locational advantage. Specifications (7)–(16) in Table 7 suggest that the correlation is positive: plants that are remote with respect to potential external suppliers are also remote from other plants within the firm. In other words, there is a positive link between external and internal distance, which runs a priori counter to the 'substitution hypothesis'. To gain additional insights, we now further differentiate plants along a number of dimensions that make them potentially more sensitive to distance. The idea is to interact the $\ln(\mathcal{I}dist)$ variable with a number of variables that are related to either industry- or product characteristics that might be especially distance sensitive. We use four such characteristics below: (i) the core segment dummy, thereby testing whether plants that are farther away from external suppliers are especially more centrally located if they belong to the firms' primary activity; (ii) an industry-year specific inverse measure of ad valorem transportation costs (see Appendix A.5 for additional details); (iii) a measure of the share of intermediates of the establishment's industry that is classified as non-homogeneous (either reference priced or differentiated; see Nunn, 2007, and Appendix A.5); and (iv) a dummy that indicates whether the industry produces predominantly 'durables' or 'non-durables'.

As can be seen from Table 7, plants in the firms' core segments and plants that manufacture durables are more centrally located in the firm as their distance to external suppliers increases. These plants may involve large production scales and the costly exchange of intermediates. Note, however, that the effect of transportation costs and the measure of 'complexity' (as proxied using the Nunn measure) are insignificant.

To provide additional evidence on the potential importance of vertical links for structuring the internal geography of firms, Table 8 summarizes the results for a large number of variations on equation (10). As in columns (7)–(16) of Table 7, we interact our key variable of interest – here the strength of internal input-output links – with the same variables: a core segment dummy; the ad valorem transport cost measure; the Nunn measure; and the manufacturing durables dummy. Table 8 reports the detailed estimation results for these interaction terms, both for unweighted and employment-weighted internal distances, using the internal input-output strength measures.[34]

As can be seen from columns (1)–(2) of Table 8, establishments with stronger input-output links are more centrally located if they belong to the firm's core segment.[35] Columns (3)–(4)

[34] Using the internal input strength yields very similar results (not reported but available upon request).

[35] Maksimovic and Phillips (2002, pp.722–724) find that "plants in the larger segments of conglomerate firms are

Table 8: Intra-firm linkage regressions with industry and product characteristics.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**
		w		**w**		**w**		**w**		**w**
log(number of establishments)	0.751^a	0.748^a	0.738^a	0.735^a	0.742^a	0.739^a	0.739^a	0.736^a	0.743^a	0.741^a
	(0.063)	(0.056)	(0.062)	(0.056)	(0.062)	(0.056)	(0.062)	(0.056)	(0.062)	(0.056)
headoffice dummy	-0.263^a	-0.261^a	-0.256^a	-0.254^a	-0.257^a	-0.255^a	-0.258^a	-0.255^a	-0.260^a	-0.258^a
	(0.042)	(0.042)	(0.042)	(0.042)	(0.041)	(0.042)	(0.041)	(0.042)	(0.042)	(0.042)
log(specialization count)	-0.065^a	-0.065^a	-0.066^a	-0.065^a	-0.066^a	-0.065^a	-0.066^a	-0.065^a	-0.066^a	-0.066^a
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.023)	(0.022)	(0.023)	(0.022)	(0.022)
$\ln(\overline{IO}_{j(i,f),t})$	-0.044	-0.051	-0.085^a	-0.096^a	-0.115^a	-0.131^a	-0.041	-0.049	-0.052	-0.061
	(0.031)	(0.031)	(0.026)	(0.026)	(0.029)	(0.029)	(0.031)	(0.030)	(0.042)	(0.042)
core segment dummy × $\ln(\overline{IO}_{j(i,f),t})$	-0.059^b	-0.066^b							-0.061^b	-0.068^b
	(0.029)	(0.029)							(0.029)	(0.029)
ad valor transport costs × $\ln(\overline{IO}_{j(i,f),t})$			0.242^c	0.202					0.313^b	0.281^b
			(0.136)	(0.136)					(0.136)	(0.137)
share of non-homog. inputs × $\ln(\overline{IO}_{j(i,f),t})$					-0.399^a	-0.452^a			-0.348^b	-0.395^a
					(0.128)	(0.128)			(0.139)	(0.139)
mfg durables dummy × $\ln(\overline{IO}_{j(i,f),t})$							-0.085^b	-0.091^b	-0.058	-0.061
							(0.041)	(0.041)	(0.044)	(0.045)
Observations	31,629	31,629	31,518	31,518	31,629	31,629	31,629	31,629	31,518	31,518
R-squared	0.270	0.258	0.268	0.256	0.270	0.259	0.270	0.258	0.270	0.258

Notes: Pooled cross section results for manufacturing plants for the years 2001, 2003, 2005, 2007, 2009, 2011, and 2013. The dependent variables in the regressions are either $\ln(\overline{d}_{j(i,f),t})$ or $\ln(\overline{d}^{w}_{j(i,f),t})$. All regressions include a full set of industry-year fixed effects. We report selected variables only. All specifications include the following establishment- and firm-level controls: log of plant employment, exporter dummy, plant-level diversity measure, firm-level diversity measure, monoindustry dummy. We also include a full set of geographical controls: minimum distance to the US, employment-weighted occupational employment similarity, oes^w, share of highly educated within a 15 kilometer distance, share of workers in management or business occupations within a 15 kilometer distance, and a full set of 'urban type' dummies. **Specifications: (1)** to **(2)**: interactions with core segment dummy; **(3)** to **(4)**: interactions with (inverse) ad valorem transport cost measure; **(5)** to **(6)**: interaction with Nunn's (2007) measure of non-homogeneous inputs share; **(7)** to **(8)**: interactions with manufacturing durables dummy; **(9)** to **(10)**: interactions with all four variables together. Robust standard errors, clustered at the firm level, reported in parentheses. Significance levels: ^{a}: $p < 0.01$, ^{b}: $p < 0.05$, ^{c}: $p < 0.1$.

reveal that stronger input-output links lead to more geographical centrality of a plant especially in industries where transportation costs are high (the ad valorem transport cost measure is smaller in industries with high transport costs, hence the positive coefficient). Columns (5)–(6) show that stronger input-output links have a larger effect on centrality in industries where the share of intermediates is relatively 'specialized' (i.e., not classified as being homogeneous). Columns (7)–(8) show that the effects of internal input-output links on centrality are especially strong for plants involved in the manufacturing of durable goods. Arguably, the latter involve more specialized intermediates, as well as possibly higher transportation costs. Last, colums (9)–(10) show that when including all interaction terms into one specification, the durables interaction loses its significance. This is likely due to the high correlation between the manu-

more efficient than plants in smaller segments. The plants in the largest segments of conglomerates with a large number of segments are particularly efficient [...] larger segments have higher average plant productivity than smaller segments." Ševčík (2013) obtains similar results using Canadian data. Our results suggest that there may be a geographical component to the observed productivity advantage: core plants are 'more centrally' located in the firms' geographical structure, which may contribute to their measured productivity advantage, either through easier access to intermediates or headquarter services.

facturing of durables dummy and Nunn's (2007) measure of 'complexity', which may pick up the same effect.

When taken together, the results summarized in Tables 7 and 8 suggest that internal input-output links are important in the sense that more strongly linked plants within firms are also geographically more central within firms. This effect is stronger for plants that operate in the firm's core segment, that produce durables, that operate in industries with higher transport costs, and that source a larger share of intermediates that are not classified as homogeneous. These plants also tend to locate closer to other plants within the same firm as the distance to external suppliers increases. Given these findings, it seems fairly unlikely that considerations related to the costly exchange of tangible goods have only little bearing on how firms structure themselves across space, as suggested by Atalay *et al.* (2014) and Raimondos *et al.* (2014). One may argue that the patterns we are picking up are inherited from the past: plants once located close to other plants because trading goods was costly, and they are still close to those plants although trading goods is much less costly nowadays. We do not think that this explanation is plausible. Indeed, the cross-sectional estimates and interactions of our key variables with time trends show that the effects we pick up tend to get stronger with time (see the Supplemental Appendix D.2). Furthermore, if location patterns are inherited from the past, young plants should be less constrained by geographical location, yet this is not the case (see Table 6.)

5 Conclusions

We have developped a new approach for measuring inter- and intra-firm linkages, and we have used it to dissect the microgeographic location patterns of hundreds of thousands of Canadian establishments. The evidence suggests that input-output linkages and external agglomeration economies are less important for multiunit plants, i.e., the characteristics of locations picked by multiunit plants differ in systematic ways from those picked by comparable standalone plants. One explanation for these results is that multiunit plants can draw on internal resources, which makes them less sensitive to their external environment when it comes to location choices. Multiunit firms are also geographically about 50% more 'compact' than expected, which suggests that internal transfers are important and fairly distance sensitive. Last, we have shown that plants that are potentially more strongly involved in resource transfers are also more centrally located within multiunit firms. This is especially true for establishments that operate in the firm's core segment of business, that are in industries that face high transportation costs, that manufacture durable goods, and that source a larger share of non-homogeneous ('complex') inputs. These findings suggest that, despite the declining costs of trading goods across space, vertical supply chains involving both the exchange of tangible inputs and the monitoring of complex transactions are important in explaining the geographical structure of multiunit firms.

Our findings contribute to the literature that recognizes that agglomeration effects affect heterogeneous plants in different ways. This has potentially many important implications for public policies that seek to develop underperforming regions by attracting plants or by providing incentives for their creation. Some plants, while providing direct benefits to the regions that host them, generate much smaller indirect benefits since they interact less with the local economic environment. While this point has been repeatedly made in a fairly informal way in policy circles, recent evidence – including Rosenthal and Strange (2010), Greenstone *et al.* (2010), Brown and Rigby (2015), and this paper – shows that plant traits and local interactions are systematically linked. This should be kept in mind when designing public policies that aim to attract certain types of investments and leveraging 'agglomeration multipliers' for regional development. In the words of Alcácer and Chung (2014, p.1761): "If new entrants are substantially less attracted to locations where industry employment is concentrated in fewer large firms, landing a single large corporation might not be preferable to attracting a group of smaller firms." This statement echoes old warnings already voiced by Chinitz (1961), which seem to have been largely forgotten due to the lack of hard empirical evidence. We hope that our results provide such hard evidence.

Acknowledgements. We thank our discussant, Tomoya Mori, as well as Mark Brown, Taiji Fujiwara, Michael D. Grubb, Catherine Haeck, Tom Holmes, Bill Kerr, Julien Martin, Florian Mayneris, Yasusada Murata, Dennis Novy, Jos van Ommeren, Pavel Ševčík, Jens Suedekum, and Matt Turner for helpful comments and suggestions. We further thank conference participants at the 2014 NARSC Meetings in Washington DC, the SCSE Meetings in Montreal, and the IOSE Workshop in Saint Petersburg; and seminar participants in UQÀM, Boston College, and the Tinbergen Institute. Special thanks go to Will Strange who suggested many useful references, and to Théophile Bougna who provided outstanding research assistance in preparing the disaggregated input-output tables. Behrens gratefully acknowledges financial support from the CRC Program of the Social Sciences and Humanities Research Council (SSHRC) of Canada for the funding of the *Canada Research Chair in Regional Impacts of Globalization*. Behrens and Sharunova gratefully acknowledge financial support from the Russian Government grant No 11.G34.31.0059. The views expressed in this paper and all remaining errors are ours.

References

[1] Aarland, Kristin, James C. Davis, J. Vernon Henderson, and Yukako Ono. 2007. "Spatial organization of firms: the decision to split production and administration." *Rand Journal of Economics* 38**(2)**: 480–494.

[2] Adams, James D., and Adam B. Jaffe. 1996. "Bounding the effects of R&D: An investigation using matched establishment-firm data." *RAND Journal of Economics* 27**(4)**: 700–721.

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

[4] Alcácer, Juan, and Wilbur Chung. 2014. "Location strategies and agglomeration economies." *Strategic Management Journal* 35**(12)**: 1749–1761.

[5] Alcácer, Juan, and Mercedes Delgado. 2013. "Spatial organization of firms and location choices through the value chain." Harvard Business School Working Paper #13-025, Harvard University.

[6] Arzaghi, Mohammad, and J. Vernon Henderson. 2008. "Networking off Madison Avenue." *Review of Economic Studies* 75**(4)**: 1011–1038.

[7] Atalay, Enghin, Ali Hortaçsu, James Roberts, and Chad Syverson. 2011. "Network structure of production." *Proceedings of the National Academy of Sciences* 108**(13)**: 5199–5202. [Supporting information available online at `www.pnas.org/lookup/suppl/doi:10.1073/pnas.1015564108/-/DCSupplemental`].

[8] Atalay, Enghin, Ali Hortaçsu, and Chad Syverson. 2014. "Vertical integration and input flows." *American Economic Review* 104**(4)**: 1120–1148.

[9] Audia, Pino G., Olav Sorenson, and Jerald Hage. 2001. "Tradeoffs in the organization of production: Multiunit firms, geographic dispersion and organizational learning." *Multiunit Organization and Multimarket Strategy* 18: 75–105.

[10] Baldwin, John R., and W. Mark Brown. 2005. "Foreign multinationals and head office employment in Canadian manufacturing firms." Economic Analysis (EA) Research Paper Series #11F0027MIE No.034, Statistics Canada.

[11] Beckmann, Martin J., and Jacques-François Thisse. 1987. "The location of production activities." In: Nijkamp, Peter (ed.), *Handbook of Regional and Urban Economics, vol. 1*. North-Holland: Elsevier B.V., pp. 21–95.

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

[13] Behrens, Kristian, Théophile Bougna, and W. Mark Brown. 2015. "The world is not yet flat: Transport costs matter!" CEPR Discussion Paper #10356, Centre for Economic Policy Research, London, UK.

[14] Bernard, Andrew B., and J. Bradford Jensen. 2007. "Firm structure, multinationals, and manufacturing plant deaths." *Review of Economics and Statistics* 89**(2)**: 193–204.

[15] Bernard, Andrew B., Andreas Moxnes, and Yukiko Saito. 2015. "Production networks, geography, and firm performance." CEPR Discussion Paper #10551, Centre for Economic Policy Research, London, UK.

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

[17] Chinitz, Benjamin. 1961. "Contrasts in agglomeration: New York and Pittsburgh." *American Economic Review Papers and Proceedings* 51**(2)**: 279–289.

[18] 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.

[19] Coval, Joshua D., and Tobias J. Moskowitz. 2001. "The geography of investment: Informed trading and asset prices." *Journal of Political Economy* 109**(4)**: 811–841.

[20] Duranton, Gilles, and Henry G. Overman. 2005. "Testing for localisation using micro-geographic data." *Review of Economic Studies* 72**(4)**: 1077–1106.

[21] 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.

[22] 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*. North-Holland: Elsevier B.V., pp. 2063–2117.

[23] Eichholtz, Piet, Rogier Holtermans, and Erkan Yönder. 2015. "The economic effects of owner distance and local property management in U.S. office markets." Forthcoming, *Journal of Economic Geography*.

[24] 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.

[25] Faggio, Giulia, Olmo Silva, and William C. Strange. 2014. "Heterogeneous agglomeration." SERC Discussion Paper #152, Spatial Economic Research Center, *London School of Economics*, UK.

[26] Forslid, Rickard, and Toshihiro Okubo. 2014. "Spatial relocation with heterogeneous firms and heterogeneous sectors." *Regional Science and Urban Economics* 46(2): 42–56.

[27] Gaubert, Cécile. 2014. “Firm sorting and agglomeration.” *Processed*, Princeton University.

[28] Giroud, Xavier. 2013. “Proximity and investment: Evidence from plant-level data.” *Quarterly Journal of Economics* 128**(2)**: 861–915.

[29] Glaeser, Edward L., and William R. Kerr. 2009. “Local industrial conditions and entrepreneurship: how much of the spatial distribution can we explain?” *Journal of Economics & Management Strategy* 18**(3)**: 623–663.

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

[31] Hayes, Rachel M., and Russell Lundholm. 1996. “Segment reporting to the capital market in the presence of a competitor.” *Journal of Accounting Research* 34**(2)**: 261–279.

[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] Helsley, Robert W., and Yves Zenou. 2014. “Social networks and interactions in cities.” *Journal of Economic Theory* 150: 426–466.

[34] Henderson, J. Vernon, and Yukako Ono. 2008. “Where do manufacturing firms locate their headquarters?” *Journal of Urban Economics* 63**(2)**: 431–450.

[35] Holmes, Thomas J. 1999. “Localization of industry and vertical disintegration.” *Review of Economics and Statistics* 81**(2)**: 314–325.

[36] Holmes, Thomas J., and John J. Stevens. 2014. “An alternative theory of the plant size distribution, with geography and intra- and international trade.” *Journal of Political Economy* 122**(2)**: 369–421.

[37] Hyland, David C., and David J. Diltz. 2002. “Why firms diversify: An empirical examination.” *Financial Management* 31**(1)**: 51–81.

[38] Johnson, Robert C., and Guillermo Noguera. 2012. “Proximity and production fragmentation.” *American Economic Review, Papers and Proceedings* 102**(3)**: 407–411.

[39] Kalnins, Arturs, and Francine Lafontaine. 2013. “Too far away? The effect of distance to headquarters on business establishment performance.” *American Economic Journal: Micro* 5**(3)**: 157–179.

[40] Klier, Thomas, and Daniel P. McMillen. 2008. “Evolving agglomeration in the U.S. auto supplier industry.” *Journal of Regional Science* 48**(1)**: 245–267.

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

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

[43] Landier, Augustin, Vinay B. Nair, and Julie Wulf. 2007. "Trade-offs in staying close: Corporate decision making and geographic dispersion." *Review of Financial Studies* 22**(3)**: 1119–1148.

[44] Li, Ben, and Yi Lu. 2009. "Geographic concentration and vertical disintegration: Evidence from China." *Journal of Urban Economics* 65**(3)**, 294–304.

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

[46] Otazawa, Toshimori, and Jos van Ommeren. 2015. "Inter-firm transaction networks and location in a city." Processed, *Kobe University* and *Vrije Universitaet Amsterdam*.

[47] Pacter, Paul. 1993. "Reporting disaggregated information". Technical report 123-A. Stamford, CT: Financial Accounting Standards Board.

[48] Petersen, Mitchell A., and Raghuram G. Rajan. 2002. "Does distance still matter? The information revolution in small business lending." *Journal of Finance* LVII**(6)**: 2533–2570.

[49] Ramondo, Natalia, Veronica Rappoport, and Kim J. Ruhl. 2015. "Intrafirm trade and vertical fragmentation in U.S. multinational corporations." Forthcoming, *Journal of International Economics*.

[50] Rosenthal, Stuart S., and William C. Strange. 2005. "The Geography of Entrepreneurship in the New York Metropolitan Area." FRBNY *Economic Policy Review*, December 2005: 29–53.

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

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

[53] 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.

[54] Ševčík, Pavel. 2013. "Plant size, productivity, and the efficiency of corporate diversification." Processed, Université du Québec à Montréal.

[55] Shaver, J. Myles, and Frederick Flyer. 2000. "Agglomeration economies, firm heterogeneity, and foreign direct investment in the United States." *Strategic Management Journal* 21: 1175–1193.

[56] Silva, Rui. 2013. "Internal labor markets and investment in conglomerates." US Census Bureau Center for Economic Studies, Paper #CES-WP-13-26.

[57] Shoar, Antoinette. 2002. "Effects of corporate diversification on productivity." *Journal of Finanace* LVII**(6)**: 2379–2403.

[58] 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.

[59] 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.

[60] Wooldrigde, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Appendix

This set of appendices is structured as follows. Appendix A contains detailed information on the data sources and the way we construct our main variables: plant- and firm-level data (Appendix A.1); disaggregated input-output matrices (Appendix A.2); distance measures and geography (Appendix A.3); geographical controls and census data (Appendix A.4); and additional industry and product-level measures (Appendix A.5). Appendix B provides details on the construction of the geographical specialization measures and summarizes additional results on the links between plant types, local specialization, and vertical disintegration. Last, Appendix C contains the propensity score matching procedure and the balance of controls.

Appendix A: Data and definition of variables

We first document our data sources and provide information on data treatment and the way we construct our key variables. Table 11 in Supplemental Appendix D summarizes the descriptive statistics for all the variables that we use.

A.1. Plant- and firm-level data

(i) Core dataset. Our analysis is based on the *Scott's National All Business Directories* database. This establishment-level database contains information on plants operating in Canada, with an extensive coverage of the manufacturing sector. Our data span the years 2001 to 2013, in two-year intervals. For every establishment, we have information on its primary 6-digit NAICS code and up to four secondary 6-digit NAICS codes; the opening year of the establishment; its employment; whether or not it is an exporter; whether or not it is a headoffice; its 6-digit postal code; up to ten products produced by the establishment; and the legal name of the entity to which it belongs. The latter is used to group plants into firms (see paragraph (ii) below). We geocode all plants by latitude and longitude using their 6-digit postal code centroids obtained from Statistics Canada's Postal Code Conversion Files (PCCF). See Appendix A.3 for details on the geocoding of the database.

The Scott's database constitutes probably the best alternative to Statistics Canada's proprietary *Annual Survey of Manufacturers Longitudinal Microdata File* or the micro-level *Canadian Business Patterns*. Although the dataset is only a large sample and not the universe of manufacturing plants, it has a very wide (85–90%) and similar coverage. It contains most of the large plants and many small plants.[36] 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 part of the Scott's database. Contrary to the manufacturing part, the non-manufacturing part is a smaller selective sample which covers mostly the three large metropolitan areas of Toronto, Montreal, and Vancouver. We mainly use the manufacturing part and use the non-manufacturing part only for various robustness checks.

To summarize, our manufacturing data are very similar to those of the *Annual Survey of Manufacturers* or the *Canadian Business Patterns* in terms of coverage and both province- and industry-level breakdown of plants and, therefore, provide a fairly accurate and representative picture of the overall manufacturing structure in Canada. The non-manufacturing data is of lesser quality and scope, but it is one of the only options at this level of geographical disaggregation.

Figure 6 depicts the spatial structure of a large multiunit firm to illustrate various aspects of our core dataset: *Air Liquide Canada Incorporated*. In 2013, it reported 53 establishments that employed about 1,100 people in 9 out of 10 provinces, with 38 establishments in Ontario and Québec. The average distance between the 53 establishments was 1,730 kilometers; and they

[36] There is no 'sampling frame' strictly speaking (though Scott's uses the *Canadian Business Patterns* – which contains the universe of entities – to contact the different establishments in a systematic way to include them into their database). There are some selection and updating biases, since firms are contacted to sign up but are of course free to not do so. Also, small/new establishments may appear in the database with a lag only (and establishments may exit with a lag only), but this is not a big issue for our purpose since we do not exploit in detail the time-series variation of the database.

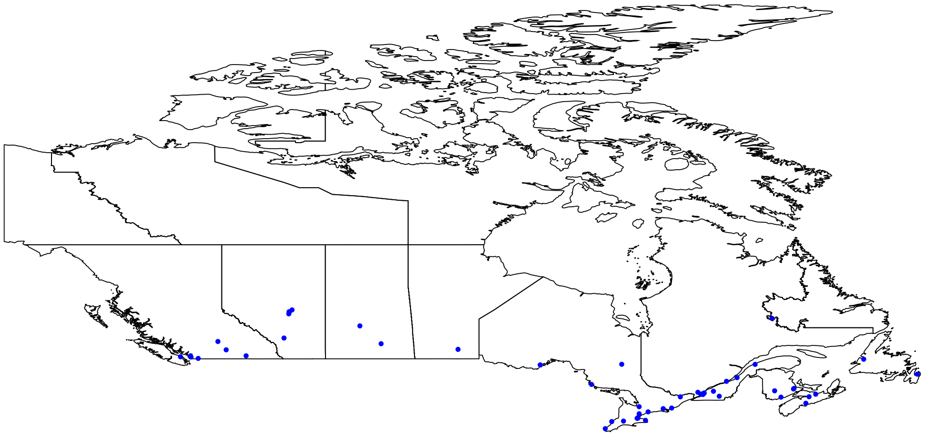


Figure 6: Spatial structure of Air Liquide Canada Incorporated in 2013.

reported 17 different NAICS codes, ranging from 213118 ('Services to oil and gas extraction') and 325120 ('Industrial gas manufacturing'), to 418410 ('Chemical (except agricultural) and allied product wholesaler-distributors'). Overall, 10 plants were classified as manufacturers (with an average employment of 54) and the remaining 43 establishments were classified as wholesalers, distributors, and support activities (with an average employment of 13). Air Liquide's manufacturing core activity was NAICS 325120 ('Industrial gas manufacturing'), with 6 out of 10 plants; whereas its non-manufacturing core activity was 418410 ('Chemical (except agricultural) and allied product wholesaler-distributors'), with 27 out of 43 establishments. The 53 establishments also reported 47 different product categories, including: various industrial, medical, cryogenic, and specialty gases (oxygen, nitrogen, hydrogen, carbon dioxide, propane, argon); electrodes; cylinders; valves and pressure instruments; industrial supplies; torches; and welding equipment and products. In total, 8 establishments reported exporting goods abroad, and the average product ubiquity in 2013 was 87% of the overall manufacturing average – meaning that Air Liquide manufactured 'more specialized' products that appear less frequently than the average in our database.

(ii) Creating firm identifiers for multiunit firms. The Scott's database provides plant-level data without the corresponding grouping of establishments into firms. We therefore exploit relevant information in the database to associate the establishments with the firms they belong to. The affiliation with a firm can only be backed out in two ways: (i) by cross-comparison of the establishments' legal names; and (ii) by cross-comparison of the unique plant identifiers which are stable across time. Although the procedure of creating the firm identifiers is fairly straightforward, it is subject to the problems that typically arise while working with string variables and which lead to some measurement error.

The idea underlying the assignment procedure is very simple: if two establishments have identical legal names they must belong to the same firm (legal entity). To this end, we loop over the sorted legal names of the establishments, where the running variable is the firm identifier. Since the usual 'string problems' arise, we pre-clean the data to allow for more accurate results. In particular, we trim the plant names to get rid of extra spaces and unify general naming patters (e.g., replacing the rare cases of "&" instead of "and", "mngmt" instead of "mgmt", etc.). We also eliminate differences in legal names stemming from the fact that Canada is a bilingual country, i.e., although 'Enterprise Rent-A-Car' and 'Enterprise Location d'Autos' belong to the same firm, the loop will split them in two different firms depending on the primary language of the province of operation.

The comparison of time-invariant plant identifiers allows us to associate a plant in year t with itself in year $t+1$. This provides a refinement of the assignment of the firm identifier in case the establishment's name has changed in a way that the preliminary data treatment could not accommodate. However, use of this 'tool' is limited to the 2003-2013 sample due to a structural change in the plant identifier design implemented by Scott's.

Besides having information on the establishments' legal names, we occasionally have information on the 'conglomerate's' legal name included into the establishment's legal name, for example "Gescan - Div. Of Sonepar" and "Lumen Inc. - Div. De Sonepar". This example raises an important methodological question of what is the relevant decision level when it comes to analyzing multiunit firms' structure and operations. It is worth noting that the Scott's database states the information on conglomerate names based on both industry affiliation and firms' divisional structure. Maksimovic and Phillips (2002) point out that industry-based divisions suffer from at least one considerable drawback. The implied industry-based divisions may not always reflect the true operational structure of conglomerates, because often times conglomerates can use their discretion to conceal the operational structure not to aid their competitors as shown by Hayes and Lundholm (1996), or do so because of changes in reporting standards. Refer to Pacter (1993) and Hyland and Diltz (2002) for an elaborate discussion on how changes in reporting requirements and relative flexibility of those undermines a significant share of the Compustat sample of firms-'diversifiers'. Another argument in favor of the firm-level instead of the conglomerate-level analysis is that establishments only report their headoffice status if they have a certain degree of discretion in regulating their activities. The usage of the headoffice dummy is thus ambiguous in case of two firm-level divisions having different headoffices. Since the coefficient on the headoffice dummy is higly significant and robust across specifications, we infer that the firm-level analysis is a viable one, at least when it comes to location choices and firms' spatial structure.

In case we made a poor choice of the level of the analysis, we still provide a lower bound for the magnitudes of the coefficients of interest. We may incorrectly classify some plants as standalone though they are multiunit. However, this would reinforce our results since we underestimate the differential effect of being a multiplant compared to being standalone.

(iii) Manufacturing sample and industry concordances. We consider that a plant is a manufacturer if it reports a manufacturing industry (NAICS 31–33) as its primary sector of activity.[37] Scott's assigns primary NAICS codes based on the main line of business of the establishment. All establishments that do not report a primary NAICS code in the range 31–33 are considered to be part of the non-manufacturing sample. Note that multiunit firms can contain both manufacturing and non-manufacturing establishments, and they typically do. When we work with the 'manufacturing' sample, we restrict the multiunit firms then to include only their manufacturing plants, disregarding all other (non-manufacturing) plants like retail outlets, wholesalers, and other support activities.

Our data span four different industrial classifications: NAICS1997, NAICS2002, NAICS2007, and NAICS2012. We have concorded those classifications to a stable set of 242 manufacturing industries and 864 industries in total. The manufacturing classification remained fairly stable over time. The largest changes have been in the construction sector (between 1997 and 2002) and in the internet-related publishing and database management (after 2002). We report results using our stable concordance, which allows for consistent use of industry fixed effects. It also allows to consistently assign the variables that are coded at the industry level.

A.2. Disaggregated input-output matrices

We use three-year lagged input-output matrices (1998, 2000, 2002, 2004, 2006, 2008, and 2010), which we concorded to the stable set of 242 manufacturing industries and 864 industries comprising the whole economy. Since the finest public release of the input-output matrices is at the L-level (link level), which is between NAICS 3 and 4, we disaggregated those matrices further to the W-level (NAICS 6) using either sales or employment data as sectoral weights.[38]

(i) Manufacturing industries only. We use the L-level national input-output tables from Statistics Canada at buyers' prices. For each manufacturing industry, i, we allocate inputs purchased or outputs sold in the L-level matrix (at the 3- or 4-digit level) to the corresponding NAICS 6-digit subsectors. To do so, we allocate the total sales of each sector to all subsectors in proportion to those sectors' sales in the total sales to obtain a 242×242 matrix of NAICS 6-digit inputs and outputs for manufacturing only. We use these matrices to compute the shares $\omega^{\text{in}}_{\Omega(\ell),i}$ and $\omega^{\text{out}}_{\Omega(\ell),i}$ that sector $\Omega(\ell)$ sells to or buys from sector i. We systematically exclude within-sector transactions as those may capture all sorts of intra-sectoral agglomeration economies

[37] Since plants in our dataset report up to four secondary NAICS codes, we could consider that a plant is a manufacturer if it reports a manufacturing sector as one of its sectors of activity, either primary or secondary. We keep this for future research.

[38] Due to confidentiality reasons, we cannot use the finer W-level matrices which are internally available at Statistics Canada. However, tests ran in Behrens, Bougna, and Brown (2015) using those matrices yielded similar results to those using the matrices constructed by our methodology.

that are conducive to clustering but not correlated with input-output linkages. We use $\omega^{\text{in}}_{\Omega(\ell),i}$ and $\omega^{\text{out}}_{\Omega(\ell),i}$ obtained from the 242 manufacturing industries to compute the manufacturing input and output distances, $\mathcal{I}dist^{\text{mfg}}$ and $\mathcal{O}dist^{\text{mfg}}$.

(ii) All industries. Turning to the non-manufacturing sectors, we do not have comparable detailed sales data for all of them. Hence, the disaggregated input-output matrices at the 6-digit level are constructed as follows. First, we use two complementary industry-level surveys from Statistics Canada to compute employment weights for non-manufacturing industries: (i) The Survey of Employment, Payrolls, and Hours (SEPH); and (ii) the Labour Force Survey estimates (LFS). We continue to use sales data from the Annual Survey of Manufacturers (ASM) for the manufacturing industries. These datasets are used to compute employment and sales shares of the 6-digit subindustries for all industries in the L-level input-output tables. Second, we use these employment and sales shares for the imputation from the L-level to the NAICS 6-digit level, thus obtaining a disaggregated 864 × 864 table for all (concorded) industries.[39] We use $\omega^{\text{in}}_{\Omega(\ell),i}$ and $\omega^{\text{out}}_{\Omega(\ell),i}$ thus obtained to compute our manufacturing input and output distances, $\mathcal{I}dist^{\text{all}}$ and $\mathcal{O}dist^{\text{all}}$ for all industries (including non-manufacturing).

A.3. Distance measures and geography

We compute great circle distances between different plants and between plants and various centroids of census geographical units by using latitude and longitude coordinates of postal code centroids. The latter are obtained from Statistics Canada's Postal Code Conversion Files (PCCF), which associate each postal code with different Standard Geographical Classifications (SGC) used for reporting census data. We match plant-level postal code information with geographical coordinates from the PCCF, using the postal code data for the next year in order to consider the fact 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. We matched the 2003 and 2005 samples with the 2001 Census geography and the corresponding PCCF's. The 2007, 2009, and 2011 samples were matched with the 2006 Census geography, while the 2013 sample was matched with the Census geography of 2011. Hence, our study period draws on geographical data from four different census (1996, 2001, 2006, and 2011), and we use this data for constructing the set of geographical controls (as detailed in Appendix A.4).

[39] In 2010 (our last year), the format of the input-output tables has changed. In order to take this new structure into account, we modify the data to fit the previous year's format. Industries 312C and 312D that were previously combined into one industry (BS3121A) are now split using employment weights. We do the same for industries 3150 and 3160 that were aggregated. We also invert the line and column position for codes BS335A0 and BS33520 to follow the structure of the previous years.

(i) Great circle distance. The great circle distance in kilometers is computed as

$$d_{ij} = 6378.137 \times \cos^{-1}\left(\sin(lat_i)\sin(lat_j) + \cos(|lon_i - lon_j|)\cos(lat_i)\cos(lat_j)\right), \qquad \text{(A-1)}$$

where the latitude (lat) and the longitude (lon) are expressed in radians, and where 6378.137 is the Earth's radius in kilometers.

(ii) Road distance. We reestimated some specifications of the model using input distances that are computed by replacing the great circle distance between plant locations with road distance. To this end, we retrieved about 115,000 postal code based origin-destination travel distances by road using the Google Maps API and the Stata `traveldistance3` module. We then fitted a third degree polynomial of great circle distance to travel distance (without a constant term, so that predicted distances are always positive), and used that filter to convert great circle distance d_{ij} to road distance d_{ij}^{road} in some of the robustness computations:

$$d_{ij}^{\text{road}} = 1.411602 d_{ij} - 0.000129 d_{ij}^2 + 0.281 \times 10^{-7} d_{ij}^3. \qquad \text{(A-2)}$$

Note that using exact road distances between any of the 800,000+ postal codes per year proved computationally infeasible. The reason is that many of our computations require running a large number of bootstrap replications, and for each of these replications we would need to recompute the travel distances, which makes the procedure much too slow. The alternative, namely stocking a 800,000+ × 800,000+ matrix of double-digit numbers in memory proved also infeasible. Accessing the matrix from hard-disk, while faster than accessing a mapping software, is also too slow. Note that using great circle distance is unlikely to matter for our results. Even for relatively small distances (less than 10 kilometers), the correlation between road distance and great circle distance is more than 0.85, and it rises to virtually 1 for longer distances (more than 300 kilometers). Given these high correlations, our results are unlikely to change substantially (as can be seen from Table 5).

(iii) Minimum distance to the US. For each plant, we compute the great circle distance (in kilometers) to the nearest US land border crossing. Crossings allow for either trucks or trains or both. There are 118 such crossings (including crossings to Alaska). They are geocoded at the 6-digit postal code level using the PCCFS. The set of land border crossings is stable over the 2001–2013 period of our analysis.

A.4. Geographical controls and census data

(i) Geographical specialization measures. We discuss in detail the construction of the specialization measures and additional results derived using those measures in Appendix B.

(ii) Occupational employment similarity. To control for geographical concentration related to labor market aspects, we measure the employment similarity between plants and their local environment. To this end, we use Occupational Employment Survey (OES) data from the Bureau of Labor Statistics (BLS) for 2002–2012 to compute the share of each of 554 occupations in each 4-digit NAICS industry.[40] 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 of the OES data allows us to avoid concording the SITC to NAICS data in 2001. Since we use US data – which is more and more employed as an instrument for other countries (see, e.g., Ellison *et al.*, 2010) – associated with Canadian plants, we do not think that there are significant simultaneity problems when using contemporaneous data. All occupational codes are adjusted to reflect changes between the 2000 and 2010 Standard Occupational Classification (SOC).

Let $\theta_{i(j),o}$ denote the share of workers of occupation o in industry i to which plant j belongs. Following Glaeser and Kerr (2009) and Ellison *et al.* (2010), one way to define the employment similarity between plant j and its local environment is as follows:

$$\text{oes}_{j(i)}^{GK} = \sqrt{\sum_o \left(\theta_{i(j),o} - \frac{1}{\sum_{l\in D, l\neq j} e_l} \sum_{k\in D, k\neq j} e_k \theta_{i(k),o}\right)^2}, \tag{A-3}$$

where D denotes the set of plants that are located in a disk with radius d around plant j in industry i; and where e_l denotes the employment of plant l. The measure (A-3) can be interpreted as the 'occupational distance' between plant j and the employment-weighted average plant closer than some distance d from plant j. Whether the distance between plant j and the *average* plant in D is the relevant metric is not clear a priori. Expression (A-3) would derive from a model in which agglomeration benefits require the plant to be similar to the 'average plant' in the local environment. However, it is easy to think about situations where the agglomeration benefits come from being 'on average similar' to the plants in the local environment.[41] We hence also construct measures of the average similarity of the plant to the other plants, both unweighted and employment weighted, as follows:

$$\text{oes}_{j(i)} = \frac{1}{|D|-1} \sum_{k\in D, k\neq j} \sqrt{\sum_o \left(\theta_{i(j),o} - \theta_{i(k),o}\right)^2} \tag{A-4}$$

$$\text{oes}_{j(i)}^{w} = \frac{1}{\sum_{l\in D, l\neq j} e_l} \sum_{k\in D, k\neq j} e_k \sqrt{\sum_o \left(\theta_{i(j),o} - \theta_{i(k),o}\right)^2}. \tag{A-5}$$

[40] The 6-digit level is not comparable between the US and Canada. Since the BLS does not report information for the retail sector, we limit our analysis to the 86 4-digit manufacturing industries. Regressions including all plants thus do not contain the OES similarity control.

[41] Matching models for the micro foundation of urban agglomeration economies that rely on models on the circle usually consider the average mismatch (and not the mismatch with the average) since interactions occur with a nearest firm only (e.g. Duranton and Puga, 2004; Strange, Hejazi, and Tang, 2006).

These two measures capture the average dissimilarity and the average weighted dissimilarity of plant j and the other plants within a given radius. We exclude the plant itself in the computation. We also compute measures that exclude all plants in the same 4-digit industry in order to pick up cross-industry location patterns that are not driven by the clustering of own 4-digit plants (recall that our measures are constructed at the 4-digit industry level, so that a low value may be confounded with specialization). We lose a number of plants that are in locations where there are no other plants in different 4-digit industries around (in which case we cannot compute the measures (A-4) and (A-5)). We use (A-3) and (A-4) as robustness checks and focus on measure (A-5) in our analysis. Since the correlations for our 319,448 manufacturing plants (pooled across all years, and both including or excluding the own 4-digit industries) between these measures and the Glaeser-Kerr measure are 0.97 (weighted) and 0.96 (unweighted), it is clear that we can use them interchangeably in our regressions.

(iii) Education and occupation variables. We construct occupational specialization and education variables using the most disaggregated Census data at the dissemination area (DA) level (45,000–55,000 geographical units). We fix a distance threshold (e.g., 15 kilometers) and aggregate up all DA's whose centroid is less than this threshold distance from the plant under consideration. For the 'occupation' variable, we compute the number of employees in business and management occupations, and take the ratio with respect to the total population in the same distance. For the 'education' variable, we compute the number of highly educated people ('some college education') and take again the ratio with respect to the total population in the same distance. To avoid problems with zeros, we create the shares by adding one to the numerator and one to the denominator. For example, we create the education variable for plant j as $\ln((\text{edu}_j + 1)/(\text{pop}_j + 1))$, where edu_j is the number of highly educated less than 15 kilometers from plant j and where pop_j is the total population less than 15 kilometers away from plant j.

(iv) Urban type dummies. We construct our 'urban type' dummies from the Statistical Area Classification (SAC) codes provided by Statistics Canada in the Census Geography Files for the 1996, 2001, 2006, and 2011 censuses. The SAC codes indicate how each census subdivision is influenced by the nearby metropolitan areas. The types are, in decreasing order of urbanization: 'urban CMA' (code smaller than 995, which corresponds to the code of the metropolitan area); 'urban strong' (code 996); 'urban moderate' (code 997); 'urban weak' (code 998); and 'rural' (code 999). We use 'urban CMA' as the excluded category.

A.5. Additional industry and product-level measures

We use the following additional variables at the industry- and product-level in our analysis.

(i) Product ubiquity. We used the plant-level data on the reported products to roughly assess their domestic ubiquity, measured as the frequency with which a particular product name appears in the dataset. The average product ubiquity of a plant is the simple average product ubiquity across all products produced by this plant. The average product ubiquity of a firm is the average product ubiquity across all products produced by this firm. Since products are reported in 'free form', we apply standard string cleaning techniques (replacing abbreviations, plural vs singular, obvious typos, capitalization, trimming etc.) before computing the different ubiquity measures.

(ii) Firm- and plant-level diversity. The Scott's database reports information on up to ten products produced at the establishment level. We construct the firm-level diversity measure as the ratio of the sum of the number of products reported at each affiliated establishment to the maximal total number of products that can be listed by those establishments:

$$\text{div}_f = \frac{\sum_j \text{nprod}_{j(f)}}{10 \cdot n(f)}. \tag{A-6}$$

Analogously, we construct the plant-level diversity measure as the ratio of the number of products reported by the establishment to the maximal total number of products that can be listed.

(iii) Core segments and manufacturing durables. For multiunit firms, we consider that a plant belongs to the firm's core segment if it reports a primary NAICS code that is among the codes most frequently reported by any plant of the firm. We also experimented with employment figures (summing the firm's employment across plants by sectors of activity), and results are fairly similar.

(iv) Ad valorem transport cost measures. This measure is provided by Statistics Canada and consists in the ratio of the price index in the transport sector and the price index of each 6-digit NAICS industry. The indices are normalized to 1 in 2007 and they span the period 2001–2013. They are available for 240 6-digit industries (238 in 2013), which explains why we lose a few observations when including them in our estimations.

(v) Share of non-homogeneous inputs. We use Nunn's (2007) 4-digit NAICS classification for the share of non-homogeneous (differentiated or reference-priced) inputs of each industry. We interact this variable with other selected variables of interest.

Appendix B. Microgeographic specialization measures

B.1. Construction

We measure the industrial specialization of a geographical area following, e.g., Rosenthal and Strange (2003, 2010). For each plant j, we fix a radius d and draw a disk of that radius around the plant. We then count the number of establishments in the same industry ('own'), and the number of establishments in other industries ('other') in that disk. Alternatively, we sum the employment across establishments instead of using simple count measures. When computing either type of measure, we exclude the plant itself. Since each plant reports up to five 6-digit NAICS codes, we can measure geographical concentration around each plant using two alternative industry definitions: a *strict* one and an *extended* one. The strict definition only considers the main line of business of an establishment as given by its primary NAICS code. The extended definition considers also all (up to) four secondary NAICS codes when computing the establishment numbers or employment totals. We denote by $\text{own}_{j(i)}$ and by $\text{other}_{j(i)}$ the 'own industry' and 'other industry' measures associated with the local environment of plant j in industry i. We consider up to 16 different types of measures in this paper. More precisely, we consider all combinations of: (i) count- or employment-based measures; (ii) strict and extended industry definitions; (iii) 4- or 6-digit NAICS industry definitions; and (iv) a 5 or 15 kilometers distance threshold. Observe that the strictest measure of geographical concentration is at the 6-digit level, 5 kilometer radius, and using the strict definition of an industry; whereas the laxest measure is at the 4-digit level, 15 kilometer radius, and using the extended definition of an industry. Our main results are robust across measures and vary in a systematic way with the strictness of the measure that we use in the regressions.

Plant counts or employment totals are subject to scale effects that have little to do with geographical specialization per se. We hence construct the specialization measures as the ratio of 'own' to 'other' count or employment. Formally,

$$\ln(\text{spec}_{j(i)}) \equiv \ln\left[(\text{own}_{j(i)}+1)/(\text{other}_{j(i)}+1)\right]. \tag{A-7}$$

Since at a small geographical scale the concentration measures have a number of zeros – there are for example plants without any other plant in the same industry within a radius d of 5 kilometers – we add 1 to the totals before taking the logarithm in (A-7).[42] This is the same procedure as for constructing our education and occupational specialization measures (see Appendix A.4). We also ran robustness checks in levels and when excluding the zeros. This makes little difference for the key quantitative results and no difference for the key qualitative results.

[42] Holmes (1999, footnote 17), for example, also suggests using $\ln(1+\text{neighbor}^{own})$ to have a concave relationship between concentration of the industry and its vertical disintegration. Adding 1 also incidentally deals with the zeros. We ran several robustness checks to confirm that our results do not depend on that choice.

B.2. Additional results for plant types and specialization

As shown in Section 3.2.2 and in Figure 3, multiunit plants are overrepresented in places where the distance to potential input sources is larger. A similar pattern shows up for the degree of industrial specialization of the locations choosen by multiunit vs standalone plants.

Figure 7: Count- and employment-based specialization, multiunit vs standalone mfg plants.

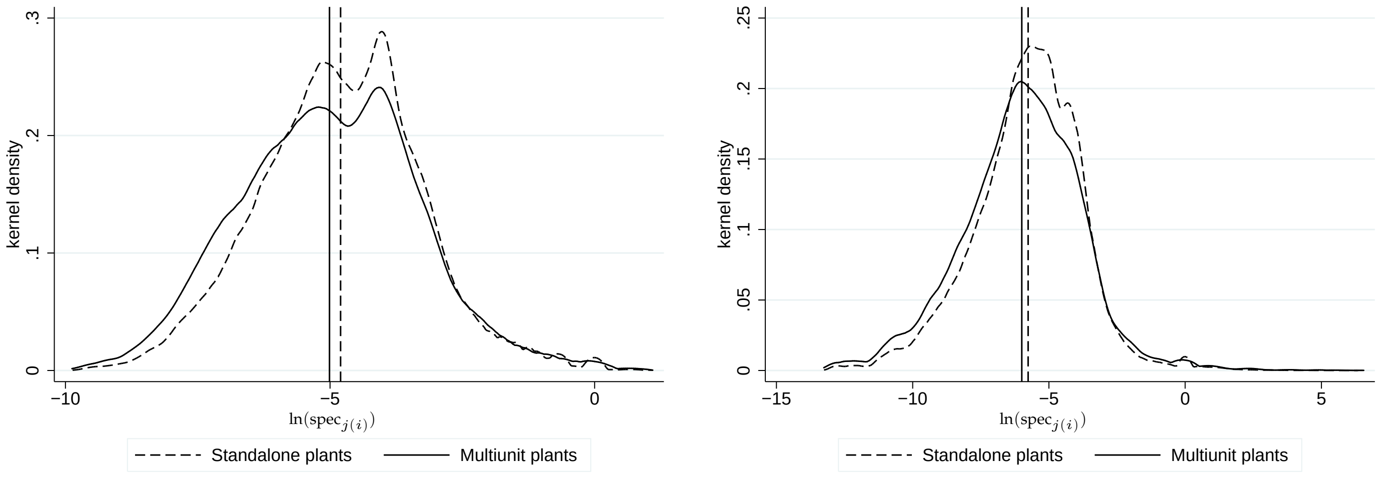


Figure 7 depicts the density distribution of plant-level geographical specialization measures for all manufacturing plants, broken down by standalone plants vs multiunit plants (see Appendix B.1 for details).[43] As can be seen from Figure 7, multiunit plants seem to be located on average in less specialized areas, both in terms of counts and employment. Yet, there is more dispersion in their location patterns in the sense that they are also slightly overrepresented in highly specialized areas, although the difference in the upper tail is less stark. Since specialization is likely to reflect some locational advantage – either locational fundamentals or external agglomeration benefits – this finding again suggests that multiunit establishments are less dependent on access to resources or agglomeration economies generated by their external environment.

Figure 7, albeit suggestive, masks substantial heterogeneity across industries and across plants with heterogeneous characteristics within industries. Indeed, industries differ in their benefits from clustering, and within industries large and small plants display different location patterns since they benefit differently from agglomeration effects (Rosenthal and Strange, 2005, 2010; Holmes and Stevens, 2014; Brown and Rigby, 2015). Table 9 reports results from simple regressions of the logarithm of plant size and a multiplant dummy on the logarithm of the specialization measures at a 5 kilometer distance, including a full set of industry-year fixed effects.[44] We report results separately for manufacturing plants and for all plants in our dataset.

[43] The measures are computed for 6-digit NAICS, strict definition of industries, and a 15 kilometer radius.

[44] Results for a 15 kilometer distance are qualitatively similar, though less sharp. This is due to the fact that,

Table 9: Geographical specialization and plant-level characteristics.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Specialization measure	Count	Count	Count	Count	Empl.	Empl.	Empl.	Empl.
Industries	All	Mfg	All	Mfg	All	Mfg	All	Mfg
ln(employment)	-0.087^a	-0.089^a	-0.088^a	-0.085^a	0.028^a	0.014^a	0.028^a	0.019^a
	(0.002)	(0.004)	(0.002)	(0.004)	(0.003)	(0.005)	(0.003)	(0.005)
multiunit dummy			0.024^b	-0.107^a			-0.019	-0.101^a
			(0.009)	(0.021)			(0.012)	(0.026)
Observations	901,705	321,589	901,705	321,589	901,705	321,589	901,705	321,589
R-squared	0.399	0.275	0.399	0.275	0.393	0.296	0.393	0.297

Notes: The dependent variable, geographical specialization, is measured at a distance of 5 kilometers, for the strict definition of industries, and at the NAICS 6-digit level. See Appendix B.1 for additional details. All regressions include a full set of industry-year fixed effect. Robust standard errors, clustered at the firm level across years, are reported in parentheses. ^{a}: $p < 0.01$; ^{b}: $p < 0.05$; ^{c}: $p < 0.1$

As can be seen from Table 9, large plants tend to cluster with other large plants. Indeed, there is a negative relationship between a plant's own employment and the count-based specialization measure in the same industry within a 5 kilometer radius (columns (1)–(4)). This effect is especially strong for multiunit manufacturing plants. When looking at specialization in terms of employment (columns (5)–(8)), that relationship however turns positive. This shows that large plants tend to locate in more specialized areas in terms of employment, but not in terms of counts. This result is reminiscent of the one in Holmes and Stevens (2014), who find that large plants (in terms of sales) tend to cluster, whereas small plants are more dispersed. It is also in line with the findings of Rosenthal and Strange (2010) who show that small plants benefit more from, and generate more, agglomeration effects. Hence, small and large plants display different location patterns. Yet, as can be seen from Table 9, the multiunit dummy – despite being correlated with plant size – has a separate effect on the specialization measure: large multiunit manufacturing plants are found in less specialized areas, both in terms of counts and in terms of employment. This explains the left tail of the distribution in Figure 7 and shows that multiunit plants behave differently than standalone plants, and that this is more than a mere size effect. We believe that this is important to emphasize since the literature on industrial organization and agglomeration has mostly worked in terms of plant size and little in terms of anything else.[45]

The results in Table 9 suggest that there are clusters of 'large plants' that do 'their own thing', with possibly little interactions. Large plants – and especially large manufacturing plants affiliated with multiunit firms – are in locations where they tend a priori to benefit less from agglomeration externalities that would be generated by concentrations of employment or

as expected, the measures of specialization are both smaller and display less variability at a larger spatial scale. Thus, it is essentially the local composition of the economic environment that matters.

[45] Rosenthal and Strange (2003) are an exception. They look at the effects of affiliated vs unaffiliated plants on births and employment levels at new plants. Their results are, however, inconclusive. More recently, Brown and Rigby (2015) also look at larger, older, foreign-owned, and multiunit firms.

other establishements in the same industry. The reason might be that large multiunit plants can draw on significant internal resources, thus making them more footloose with respect to their external environment.

Appendix C: Propensity score matching procedure

To construct our counterfactual multiunit firms, we use 1:1 nearest neighbor propensity score matching to associate a 'comparable' standalone plant with each multiunit plant. We match each plant in each year-industry on the following variables: log employment, exporter dummy, log geographical controls constructed from the Census data (see Appendix A.4 for details), log plant-level geographical specialization measures (strict 6-digit count measures computed within a 5 kilometer radius; see Appendix B.1 for details), the of log number and the log average ubiquity of products of the plant (see Appendix A.5 for details), and the log relative input distances computed for $N = 5$. Table 10 summarizes the match variables and the balance of controls for 2001, and displays the T-stats for equality of means tests.

Table 10: Balance of controls for the PSM procedure (2001 sample, all plants).

Match variable	# Observations	Mean	Std dev.	Mean PSM	Std dev. PSM	T-stat
ln(employment)	15,619	2.7460	0.0117	2.7470	0.0112	0.9199
exporter dummy	15,619	0.2639	0.0035	0.2621	0.0035	0.6600
ln(minimum distance to US)	15,619	4.4212	0.0096	4.4241	0.0085	0.8067
ln(share of highly educated)	15,619	-2.2337	0.0037	-2.2404	0.0039	0.1744
ln(share of BM occupations)	15,619	-1.9394	0.0021	-1.9451	0.0022	0.0400^b
ln(specialization count)	15,619	-4.0703	0.0119	-4.0656	0.0123	0.7474
ln($\mathcal{I}dist/\mathcal{M}dist$)	15,619	1.8880	0.0010	1.8890	0.0095	0.9405
ln(establishment diversification)	15,619	0.6985	0.0055	0.6890	0.0053	0.1709
ln(average product frequency)	15,619	-7.5105	0.0094	-7.5044	0.0098	0.5802

Notes: Samples are stratified by 3-digit NAICS industries. We drop industry-year pairs with less than 30 plants and less than 10 multiunit plants. We also drop industries in which multiunit plants account for more than 50% of all plants.

Observe that we have 134,170 multiunit plants (treated units) and 767,535 standalone plants (untreated units). Hence, our setting lends itself well to PSM because of the large number of control cases we can draw from. We stratify plants by their primary 3-digit industry, and then match the closest standalone plant in the same industry using the variables in Table 10.[46] We drop industry-year pairs with less than 30 plants, less than 10 multiunit plants, or in which multiunit plants account for more than 50% of the industry. This leads to some attrition of our sample but is unlikely to substantially affect the results (e.g., in 2001 we have 16,824 multiunit plants, 15,619 of which are treated in the PSM). As can be seen from Table 10, the controls are well balanced. All observations are also on common support.

[46] Ideally, we would match on 6-digit industry codes. This leaves us, however, with many industries with small numbers of plants so that good controls are difficult to find so that we have to drop a large number of obervations.

We can also restrict our matching to manufacturing plants only. This has the downside of making us lose a number of plants – all those that belong to multiunit firms yet do not report a manufacturing primary industry code – but has the upside of allowing for a wider range of match controls. In addition to the foregoing controls, we can control for example for occupational employment similarity measures between the plant and its environment, which we can only compute for manufacturing (see Appendix A.4 for details). The controls remain balanced for manufacturing. We do not show these results for the sake of brevity.

Supplemental Appendix, for online publication only

This supplemental appendix is structured as follows. Appendix D contains additional tables and results. Appendix D.1 contains descriptive statistics for external input-output linkages, geographical specialization, and intra-firm linkages. Appendix D.2 summarizes our year-by-year cross-sectional estimates. Appendix D.3 provides additional results on the links between geographical specialization and vertical disintegration. Appendix D.4 summarizes several robustness checks for the inter-firm linkage regressions. Appendix D.5 delves deeper into the geography of multiunit firms and provides supplemental estimates. Last, Appendix D.6 summarizes extra results for the intra-firm linkage regressions for all plants, manufacturing and non-manufacturing, combined. Appendix E proposes a simple model that formalizes the key ideas underlying Figure 1 in the paper.

Appendix D. Additional tables and results

D.1. Descriptive statistics

(i) Descriptive statistics for the different variables used in the analysis. Table 11 summarizes the different variables that we use in our analysis.

(ii) Additional descriptive statistics for input-output linkages. Table 12 displays the ten NAICS 6-digit manufacturing industries with the shortest and with the longest input distances in 2005, respectively.

As can be seen from Table 12, industries related to computers, automobiles, audio, video, cables, and stamping and moulding make up a large chunk of the top panel (as well as the 'Hosiery and sock mills' industry; see also Table 2 in Holmes, 1999). The bottom part contains industries for which other locational considerations than good access to input and output markets are key. For example, 'Explosives manufacturing' has strong locational constraints – one cannot manufacture explosives anywhere, especially in areas where there are a lot of other firms – whereas 'Seafood product preparation and packaging' or the different 'Pulp mills' obviously strongly rely on access to the sea or other bodies of water.

(iii) Additional descriptive statistics for geographical specialization. Tables 13 and 14 summarize our specialization measures for the strict 6-digit industry definitions and a 5 kilometer radius. Table 13 reports both count and employment-based measures across all manufacturing industries and all non-manufacturing industries on a yearly basis, whereas Table 14 reports the top 10 most specialized industries for either manufacturing or non-manufacturing (where specialization is measured across all years pooled). Observe that manufacturing specialization

Table 11: Summary statistics for the different variables used in the analysis.

Variable	# observations	Mean	Standard dev.	Minimum	Maximum	Sample of establishments
Dependent variables						
input distance, $N = 5$	321,589	177.9827	185.25	5.53	1970.88	Manufacturing
input distance, $N = 5$ (excluding own 4-digit NAICS industry)	321,589	182.71	189.38	5.56	1923.66	Manufacturing
input distance, $N = 5$ (road distance)	321,589	237.35	242.35	7.75	2540.12	Manufacturing
input distance, $N = 5$	901,705	256.29	326.04	5.23	2912.67	All industries
output distance, $N = 5$	321,589	135.94	147.07	3.12	1948.24	Manufacturing
output distance, $N = 5$ (excluding own 4-digit NAICS industry)	321,589	140.00	150.78	2.97	1924.91	Manufacturing
output distance, $N = 5$ (road distance)	321,589	182.43	193.92	4.35	2475.77	Manufacturing
output distance, $N = 5$	887,563[(1)]	262.63	326.88	3.81	3593.20	All industries
average distance to all other plants in the firm	134,170	872.78	968.18	0	5013.54	All multiunit plants
average distance to all other plants in the firm	31,890	875.16	972.81	0	5013.54	Manufacturing multiunit plants
average distance to all other plants in the firm, employment weighted	134,170	854.84	991.40	0	5013.54	All multiunit plants
average distance to all other plants in the firm, employment weighted	31,890	849.87	995.73	0	5013.54	Manufacturing multiunit plants
Key variables of interest						
multiunit dummy	901,705	0.15	0.36	0	1	All industries
multiunit dummy	321,589	0.10	0.30	0	1	Manufacturing
number of plants in the firm	901,705	2.82	9.94	1	145	All industries
number of plants in the firm	321,589	1.71	5.77	1	129	Manufacturing
Firm employment (excluding the plant itself)	102,280	331.81	565.42	1	6091	All multiunit plants
Firm employment (excluding the plant itself)	31,890	329.64	678.92	1	6061	Manufacturing multiunit plants
average input-strength with other plants in the firm	134,170	0.01	0.02	0	0.34	All multiunit plants
average output-strength with other plants in the firm	134,170	0.01	0.02	0	0.37	All multiunit plants
average input-output strength with other plants in the firm	134,170	0.01	0.02	0	0.34	All multiunit plants
average input-strength with other plants in the firm	31,890	0.02	0.03	0	0.34	Manufacturing multiunit plants
average output-strength with other plants in the firm	31,890	0.02	0.04	0	0.37	Manufacturing multiunit plants
average input-output strength with other plants in the firm	31,890	0.02	0.03	0	0.34	Manufacturing multiunit plants
Establishment-level controls ($\mathbf{X}_j$)						
plant age	798,974[(2)]	30.22	24.95	0	409	All industries
exporter dummy	901,705	0.23	0.42	0	1	All industries
exporter dummy	321,589	0.45	0.50	0	1	Manufacturing
headoffice dummy	901,705	0.19	0.39	0	1	All industries
manufacturing durables dummy	321,589	0.63	0.48	0	1	Manufacturing
log employment	321,589	2.50	1.40	0	6.63	Manufacturing
log employment	580,116	2.23	1.27	0	8.99	Non-manufacturing
number of products	901,616	2.62	2.18	1	10	All industries
average product ubiquity (×100), plant-level	90,1616	0.07	0.10	0.00	0.81	All industries
number of products	321,566	3.14	2.36	1	10	Manufacturing
average product ubiquity (×100), plant-level	321,566	0.07	0.11	0.00	0.81	Manufacturing
Geographical controls ($\mathbf{G}_j$)						
average minimum distance to 5 closest plants in each industry	321,589	55.52	94.33	1.80	1804.34	Manufacturing
average minimum road distance to 5 closest plants in each industry	321,589	76.74	127.33	2.54	2292.19	Manufacturing
average minimum distance to 5 closest plants in each industry	901,705	61.59	110.83	1.29	2499.20	All industries
log specialization measure, 6-digit level, 5km, count-based	901,705	-4.2840	1.48	-8.86	2.40	All industries
log specialization measure, 6-digit level, 15km, count-based	901,705	-4.73	1.45	-9.97	2.40	All industries
log specialization measure, 6-digit level, 5km, employment-based	901,705	-5.54	2.11	-12.21	6.55	All industries
log specialization measure, 6-digit level, 15km, employment-based	901,705	-5.59	1.87	-13.36	6.55	All industries
OES similarity (15km) Glaeser-Kerr	319,448[(2)]	0.19	0.08	0	0.61	Manufacturing
OES similarity (15km)	319,448[(2)]	0.26	0.06	0	0.60	Manufacturing
OES similarity (15km) weighted	319,448[(2)]	0.26	0.07	0	0.07	Manufacturing
OES similarity (15km) Glaeser-Kerr, excluding own 4-digit	319,379[(2)]	0.19	0.08	0.04	0.61	Manufacturing
OES similarity (15km), excluding own 4-digit	319,379[(2)]	0.26	0.06	0.05	0.60	Manufacturing
OES similarity (15km) weighted, excluding own 4-digit	319,379[(2)]	0.26	0.07	0.05	0.70	Manufacturing
workers in management occupations, 15 kilometer radius	901,705	45382.06	40452.58	0	163,070	All industries
workers in business occupations, 15 kilometer radius	901,705	88481.46	78445.34	0	279,360	All industries
population with some university education, 15 kilometer radius	901,705	139847.70	138974.30	0	553,775	All industries
population, 15 kilometer radius	901,705	811519.10	718105.60	0	2,560,435	All industries
distance in kilometers to the closest US land border crossing	901,705	116.93	145.83	0	2558.62	All industries
Census Metropolitan Area dummy	901,705	0.88	0.32	0	1	All industries
Strong urban links dummy	901,705	0.02	0.15	0	1	All industries
Moderate urban links dummy	901,705	0.04	0.20	0	1	All industries
Weak urban links dummy	901,705	0.04	0.18	0	1	All industries
Rural area dummy	901,705	0.02	0.13	0	1	All industries

Notes: We report descriptive statistics for all the variables pooled across years. [(1)]We have no information allowing us to compute the output coefficients for NAICS 531 and 813. We therefore exclude the 14,142 zero values for those industries. [(2)]Differences in the number of observations are due to the fact that some variables are missing for some industries.

Table 12: Ten shortest and longest input distances for manufacturing industries in 2005.

		$\mathcal{I}dist_{\ell}^{\text{mfg}}$		$\mathcal{O}dist_{\ell}^{\text{mfg}}$		$\mathcal{M}dist_{\ell}^{\text{mfg}}$	
NAICS	Industry name	Mean	Std dev.	Mean	Std dev.	Mean	Std dev.
Ten shortest manufacturing input distances in 2005							
315110	Hosiery and sock mills	22.92	19.64	23.60	18.82	13.17	18.21
311830	Tortilla manufacturing	45.44	27.75	39.52	25.29	19.17	22.34
333220	Rubber and plastics industry machinery manufacturing	52.22	82.17	56.20	69.46	16.56	31.01
334310	Audio and video equipment manufacturing	52.34	72.03	50.62	66.40	20.02	31.25
335920	Communication and energy wire and cable manufacturing	58.34	69.42	60.61	58.59	25.52	33.91
332118	Stamping	60.50	88.46	47.00	67.66	20.61	30.82
311615	Poultry processing	65.34	64.65	47.40	51.23	42.80	64.62
336110	Automobile and light-duty motor vehicle manufacturing	65.74	121.90	81.90	99.19	26.13	46.65
334110	Computer and peripheral equipment manufacturing	67.17	96.42	82.69	95.85	25.68	54.20
333511	Industrial mould manufacturing	67.17	66.58	72.69	57.39	30.30	27.80
Ten longest manufacturing input distances in 2005							
325220	Artificial and synthetic fibres and filaments manufacturing	287.47	99.80	111.24	127.12	49.36	58.29
325190	Other basic organic chemical manufacturing	288.87	130.99	254.71	209.39	43.91	81.83
325210	Resin and synthetic rubber manufacturing	294.31	119.33	102.89	125.68	42.86	71.21
325181	Alkali and chlorine manufacturing	325.37	141.88	338.72	225.74	77.48	73.26
311710	Seafood product preparation and packaging	331.94	194.37	253.72	187.40	207.42	187.21
325120	Industrial gas manufacturing	334.89	227.23	324.37	285.55	85.38	183.80
322112	Chemical pulp mills	335.73	173.03	310.63	187.36	192.14	171.35
322111	Mechanical pulp mills	341.96	177.74	314.16	184.20	179.93	149.53
325189	All other basic inorganic chemical manufacturing	372.28	177.02	346.71	243.83	78.94	118.67
325920	Explosives manufacturing	377.97	219.38	277.84	205.61	169.20	186.51

Notes: Descriptive statistics for input and output distances, using 242 (concorded) manufacturing industries. We report results for $N = 5$ nearest plants in each industry. The ten shortest and longest input distance industries are determined in increasing order of their input distances. We report industry averages, based on plant-level measures, in 2005.

has dramatically fallen in Canada over the years, both in terms of plant counts and in terms of employment totals (see Behrens, Bougna, and Brown, 2015; and Behrens and Bougna, 2015, for additional evidence for Canada). Observe further that, as expected, resource-based industries make up a fair share of the list, both for manufacturing and for non-manufacturing. The reason for this is that our measure of specialization is a topographic measure: industries that are overrepresented in thinly populated regions will rank among the most specialized ones.

Note also that there is substantial variation in the specialization measures at the plant level. In other words, within narrowly defined 6-digit industries, plants in highly specialized areas coexist with plants in poorly specialized areas.

(iv) Additional descriptive statistics for intra-firm linkages. Table 15 summarizes the different internal distance measures by year. As can be seen from that table, employment and input-output weighted distance measures are smaller than 'raw' distance measures, thus suggesting that bigger and more strongly vertically linked plants are closer to each other (see also Figure 4 in the paper).[47] The average distance across plants of multiunit firms hovers around

[47] Johnson and Noguera (2012) use similar distance-weighted measures to show that the international fragmentation of the value chain is geographically localized among nearby countries.

Table 13: Summary and descriptive statistics for specialization measures (6-digit, strict, 5 kilometer).

		$\text{own}_{j(i)}^{\text{count}}$		$\text{own}_{j(i)}^{\text{empl}}$		$\text{spec}_{j(i)}^{\text{count}}$		$\text{spec}_{j(i)}^{\text{empl}}$	
Year	Observations	Mean	Std dev.	Mean	Std dev.	Mean	Std dev.	Mean	Std dev.
		Geographical concentration and specialization (all manufacturing industries)							
2001	52,031	9.59	24.44	256.94	800.87	0.07	0.18	0.13	2.58
2003	51,876	8.80	21.31	237.76	654.93	0.06	0.16	0.11	3.79
2005	49,223	7.97	18.36	218.54	554.29	0.06	0.16	0.10	2.93
2007	46,243	6.70	13.91	193.33	440.59	0.05	0.15	0.11	2.66
2009	44,681	6.31	12.94	181.73	403.86	0.05	0.17	0.10	2.62
2011	42,210	5.75	11.69	161.73	356.59	0.05	0.15	0.11	2.87
2013	35,325	4.30	8.56	128.47	275.89	0.04	0.13	0.09	3.30
		Geographical concentration and specialization (all non-manufacturing industries)							
2001	56,666	29.98	51.32	664.91	1447.14	0.04	0.16	0.04	1.47
2003	62,610	27.13	48.28	640.28	1382.36	0.04	0.17	0.07	3.18
2005	72,606	26.74	53.18	665.10	1581.65	0.04	0.18	0.05	2.08
2007	87,673	24.90	48.24	643.08	1464.66	0.03	0.09	0.02	0.31
2009	93,773	24.12	46.66	647.42	1491.01	0.03	0.08	0.02	0.27
2011	97,570	22.36	43.64	629.74	1455.43	0.04	0.12	0.03	0.74
2013	109,218	17.90	37.54	532.94	1318.64	0.04	0.11	0.03	0.78

Notes: Descriptive statistics for all years (2001, 2003, 2005, 2007, 2009, 2011, 2013) and 242 (concorded) manufacturing industries or all 622 (concorded) non-manufacturing industries. We report results for $d = 5$ and for the strict 6-digit industry definitions. Results for other choices are available from the authors upon request. We report industry averages across time.

Table 14: Top ten industries in terms of count-based specialization measures (6-digit, strict, 5 kilometer).

Top 10 specialization measures for manufacturing industries (all years)			
NAICS	Industry name	$\text{spec}_{j(i)}^{\text{count}}$	$\text{spec}_{j(i)}^{\text{empl}}$
311710	Seafood product preparation and packaging	0.47	4.01
321111	Sawmills (except shingle and shake mills)	0.30	1.33
327410	Lime manufacturing	0.22	0.14
321217	Waferboard mills	0.21	0.77
312130	Wineries	0.21	0.71
321112	Shingle and shake mills	0.19	0.30
333110	Agricultural implement manufacturing	0.17	0.36
311214	Rice milling and malt manufacturing	0.15	0.10
321992	Prefabricated wood building manufacturing	0.15	0.12
311611	Animal (except poultry) slaughtering	0.14	0.12
Top 10 specialization measures for non-manufacturing industries (all years)			
111140	Wheat farming	0.42	0.09
113210	Forest nurseries and gathering of forest products	0.36	0.52
913910	Other local, municipal and regional public administration	0.35	0.33
212397	Peat extraction	0.33	0.30
111999	All other miscellaneous crop farming and aquaculture	0.32	0.28
212395	Gypsum mining	0.30	0.20
111330	Non-citrus fruit and tree nut farming	0.27	0.19
713930	Marinas	0.27	0.35
113311	Logging (except contract)	0.26	0.90
113312	Contract logging	0.24	0.25

Notes: Descriptive statistics pooled across all years (2001, 2003, 2005, 2007, 2009, 2011, 2013) for 242 (concorded) manufacturing industries or 622 (concorded) non-manufacturing industries. We report results for $d = 5$ and for the strict 6-digit industry definitions. Results for other choices are similar. The top 10 industries are determined in decreasing order of their specialization measures. We report time averages across industries.

700–800 kilometers, depending on the year. Note that this figure is comparable to that reported in Aarland *et al.* (2007) for the US, where firms have an average distance between plants of about 635 kilometers – this is slightly less than in our case, but Canada has a more dispersed geography (especially in terms of the density distribution of the population) than the US.

Table 15: Summary and descriptive statistics for internal distance measures of multiunit firms.

All industries						
Year	# firms	Average # plants	Average $\overline{d}_f$	Average $\overline{d}_f^w$	Average $\overline{d}_f^{IO}$	Average $\overline{IO}_f$
2001	4,874	3.45	763.40	731.90	756.32	0.82%
2003	4,970	3.54	731.30	707.31	725.01	0.80%
2005	5,026	3.56	694.67	669.62	689.94	0.72%
2007	5,439	3.55	858.49	836.04	848.04	0.68%
2009	5,559	3.62	860.47	840.96	851.48	0.63%
2011	5,584	3.66	857.27	845.37	844.45	0.62%
2013	5,797	3.79	850.08	837.86	834.82	0.43%

Notes: Summary statistics of average internal distances of plants within the same firms. All internal input-output measures (last column) are multiplied by 100 and thus displayed as percentages.

D.2. Cross-sectional estimates

Table 16 summarizes the results for year-by-year cross-sectional regressions. The simple average coefficient for the multiunit dummy is 0.107, which is very close to the 0.111 coefficient in the benchmark estimation in Table 4, column (4). The corresponding figures for the number of establishments and the firm employment are 0.038 and 0.015, respectively (which should be compared to 0.033 and 0.013 from colums (5) and (6) in Table 4). As can be further seen from Table 16, the multiunit dummy coefficient is increasing in magnitude over time. This may reflect changes in communication technologies or international sourcing that have a larger effect on multiunit firms than on single unit firms. The same trends hold when we measure the 'internal size' of the firm using the number of plants (middle panel), or total firm employment excluding the plant itself (bottom panel). In all cases, multiunit establishments are farther away from potential input sources and this trend is stronger in more recent years. We have also estimated our baseline specification (see Table 4) by interacting the variables of interest – the multiunit dummy, the number of establishments, or the firm's employment – with a time trend. The results are consistent with the cross-section estimates of Table 16: the magnitude of the coefficients is not decreasing over time.

D.3. Geographical specialization and vertical disintegration

Table 17 shows that our key results are highly robust to how we measure the geographical specialization around establishments.[48] As can be seen from Table 17, the multiunit dummy

[48] To save space, we do not report results for input linkages constructed using $N = 7$ nearest neighbors. Those results are qualitatively the same. Also, results using output (as opposed to input) links are qualitatively the

Table 16: Detailed results for year-by-year cross-sectional estimates.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	2001	2003	2005	2007	2009	2011	2013
				multiunit dummy			
multiunit dummy	0.072^{a}	0.075^{a}	0.053^{a}	0.132^{a}	0.126^{a}	0.144^{a}	0.144^{a}
	(0.014)	(0.014)	(0.014)	(0.019)	(0.019)	(0.020)	(0.024)
ln(specialization count)	-0.056^{a}	-0.031^{a}	-0.003	-0.075^{a}	-0.075^{a}	-0.077^{a}	-0.073^{a}
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	51,692	51,532	48,872	45,939	44,362	41,896	35,026
R-squared	0.688	0.711	0.720	0.516	0.545	0.516	0.459
				number of establishments			
ln(number of establishments)	0.027^{c}	0.023^{c}	0.022^{c}	0.042^{a}	0.042^{a}	0.054^{a}	0.058^{a}
	(0.015)	(0.014)	(0.013)	(0.014)	(0.016)	(0.017)	(0.018)
ln(specialization count)	-0.056^{a}	-0.031^{a}	-0.003	-0.075^{a}	-0.075^{a}	-0.077^{a}	-0.073^{a}
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	51,692	51,532	48,872	45,939	44,362	41,896	35,026
R-squared	0.687	0.711	0.720	0.515	0.544	0.516	0.458
				firm employment			
ln(firm employment)	0.008^{b}	0.006^{c}	0.007^{b}	0.020^{a}	0.019^{a}	0.022^{a}	0.022^{a}
	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.005)	(0.005)
ln(specialization count)	-0.056^{a}	-0.031^{a}	-0.003	-0.075^{a}	-0.075^{a}	-0.077^{a}	-0.073^{a}
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	51,692	51,532	48,872	45,939	44,362	41,896	35,026
R-squared	0.687	0.711	0.720	0.515	0.544	0.516	0.458

Notes: The dependent variable in all regressions is $\ln(\mathcal{I}dist)$ and it is constructed using the $N = 5$ nearest neighbors in each industry. Regressions for manufacturing establishments only. All regressions are of the same form as in (9), and include industry fixed effects. All regressions include the same establishment- and firm-level controls, and detailed geographical controls, as in Table 4. Robust standard errors, clustered at the firm level, in parentheses. Significance levels: ^{a}: $p < 0.01$, ^{b}: $p < 0.05$, ^{c}: $p < 0.1$.

is positive and very stable across all specifications, irrespective of whether or not we include geographical controls (minimum distance to the US, share of highly educated within a 15 kilometer radius, share of workers in management and business occupations within a 15 kilometer radius, and occupational employment similarity within 15 kilometers weighted by plant employment). Turning to the link between specialization and vertical disintegration, a number of interesting insights emerge. First, in 12 out of 16 regressions using count-based measures of specializations, the coefficient on the specialization measure is negative and highly significant: at close distance (5 kilometers), areas with a larger concentration of plants in the same industry tend to offer better access to inputs. That effect goes away at a larger spatial scale (15 kilometers) when we include the geographical controls. The reason is that areas that are strongly specialized in manufacturing industries tend to be also areas in which there are less highly-educated workers and less workers in management and business occupations: the coefficients on the share of highly educated, on the share of workers in management and business occupations, and on the occupational employment similarity are all positive and significant at 1% in all regressions, whereas the minimum distance to the US coefficient is negative and

same. All results are available from the authors upon request.

significant at 1% all the time. Thus, our specialization measures at large spatial scale pick up that effect. Turning to employment-based measures of specialization, the right part of the table shows that the negative and significant association between specialization and vertical disintegration continues to hold at short range (5 kilometers) when geographical controls are not included, but completely disappears once we control for geographical aspects.

Table 17: Geographical specialization and vertical disintegration.

			Count-based measures				**Employment-based measures**			
			Establishment controls		**+ Geographic controls**		**Establishment controls**		**+ Geographic controls**	
Type of specialization measure			γ_2	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2	γ_1
6-digit	strict	5 km	-0.074[a]	0.125[a]	-0.023[a]	0.111[a]	-0.008[a]	0.133[a]	0.006[a]	0.112[a]
			(0.002)	(0.015)	(0.003)	(0.014)	(0.001)	(0.015)	(0.001)	(0.014)
6-digit	extended	5 km	-0.071[a]	0.126[a]	-0.016[a]	0.112[a]	-0.005[a]	0.133[a]	0.007[a]	0.112[a]
			(0.002)	(0.015)	(0.003)	(0.014)	(0.001)	(0.015)	(0.001)	(0.014)
6-digit	strict	15 km	-0.041[a]	0.131[a]	0.039[a]	0.111[a]	0.009[a]	0.132[a]	0.018[a]	0.110[a]
			(0.003)	(0.015)	(0.003)	(0.014)	(0.002)	(0.015)	(0.002)	(0.014)
6-digit	extended	15 km	-0.036[a]	0.132[a]	0.048[a]	0.111[a]	0.010[a]	0.132[a]	0.017[a]	0.110[a]
			(0.003)	(0.015)	(0.003)	(0.014)	(0.002)	(0.015)	(0.002)	(0.014)
4-digit	strict	5 km	-0.064[a]	0.126[a]	-0.005[b]	0.112[a]	-0.007[a]	0.133[a]	0.006[a]	0.112[a]
			(0.002)	(0.015)	(0.003)	(0.014)	(0.001)	(0.015)	(0.001)	(0.014)
4-digit	extended	5 km	-0.068[a]	0.127[a]	-0.007[b]	0.112[a]	-0.008[a]	0.133[a]	0.003[b]	0.112[a]
			(0.002)	(0.015)	(0.003)	(0.014)	(0.001)	(0.015)	(0.001)	(0.014)
4-digit	strict	15 km	-0.029[a]	0.132[a]	0.063[a]	0.111[a]	-0.000	0.133[a]	0.008[a]	0.111[a]
			(0.003)	(0.015)	(0.003)	(0.014)	(0.002)	(0.015)	(0.002)	(0.014)
4-digit	extended	15 km	-0.037[a]	0.132[a]	0.056[a]	0.111[a]	-0.006[a]	0.134[a]	0.002	0.112[a]
			(0.003)	(0.015)	(0.004)	(0.014)	(0.002)	(0.015)	(0.002)	(0.014)

Notes: Pooled cross section results for manufacturing plants for the years 2001, 2003, 2005, 2007, 2009, 2011, and 2013. The dependent variable in all regressions is $\ln(\mathcal{I}dist)$ and it is constructed for $N = 5$. All regressions are of the same form as in (9), and include industry-year fixed effects. All regressions include the same establishment- and firm-level controls as in Table 4. The columns '+ Geographical controls' report estimates that include geogaphical controls (minimum distance to the US, share of highly educated within 15 kilometers, share of workers in management and business occupations within a 15 kilometer radius, occupational employment similarity within 15 kilometers weighted by plant employment, and a full set of 'urban type' dummies). Robust standard errors, clustered at the firm level across years, in parentheses. The number of observations in each regression is 321,589 without geographic controls, and 319,319 with geographical controls. Significance levels: [a]: $p < 0.01$, [b]: $p < 0.05$, [c]: $p < 0.1$.

Our results suggest that specialization tends to be associated with better access to inputs when specialization is narrow (6-digit industries), at a small spatial scale (5 kilometers), associated with small plants (count- vs employment-based results): the geographical scope of agglomeration economies is limited (see also Rosenthal and Strange, 2003). Note that, as shown in Table 5 and as substantiated by the foregoing 4-digit results, access to vertically linked firms seems primarily driven by clustering within NAICS industries at a higher level of aggregation than the 6-digit level.

When taken together, our results provide indirect evidence for more vertical disintegration in more geographically specialized locations. This is consistent with previous findings by Holmes (1999) for the US and by Li and Lu (2009) for China. Disintegration seems to operate at a narrow industrial definition, at a small geographical scale, and when there are many small firms (as opposed to a few big ones). The latter result suggests that industrial organization is key to understanding the specialization-disintegration link, as previously suggested in the literature (e.g., Rosenthal and Strange, 2010; or Holmes and Stevens, 2014). See also Table 9, which provides results consistent with the ones reported in Table 17 above.

Note that the coefficient on specialization is robust for all measures when geographical controls are not included. When the latter are included, the results remain robust for count-based measures at small geographical scales (5 kilometers). The coefficients on the specialization variables become insignificant or turn positive when employment-based measures and geographical controls are included. This finding suggests that smaller establishments are more dependend on external links (Chinitz, 1961; Rosenthal and Strange, 2010): more specialized areas – in terms of counts – offer generally better access to potential input suppliers, whereas more specialized areas – in terms of employment – offer worse access to potential inputs. The latter effect may be due to the fact that more specialized areas in terms of size are areas with larger establishments (Holmes and Stevens, 2014; see also Table 9 in Appendix B). It is also in part mechanically driven by the fact that our input-distance variables are based on plant counts and do not control for the size of plants.

D.4. Additional robustness checks for the inter-firm linkage regressions

Table 18 below summarizes the results for a series of additional robustness checks concerning the inter-firm linkage regressions. These robustness checks essentially consist in trimming the sample, dropping zeros, and estimating the model in level instead of logs.

D.5. Geographical patterns of multiunit establishments

We now dig still a bit deeper into the geographical distribution of manufacturing multiunit plants. Table 19 reveals that multiunit plants are disproportionally located in either fairly rural or weakly urban areas, or in the most urbanized census metropolitan areas (CMAs). They are, however, underrepresented in strong or moderately urban areas, which corresponds to the outskirts of major metropolitan regions or medium-sized cities.

Table 20 below runs a number of regressions where we supplement the continuous geographical controls that we use with economic region (ER) fixed effects. As can be seen from columns (2)–(4) of that table, the introduction of ER effects – when combined with industry-year fixed effects – substantially reduces and even eliminates the coefficient on the multiunit dummy for the manufacturing establishments. Yet, as can be seen from columns (6)–(8), the coefficient survives for all establishments, although it falls in magnitude.

The fact that the multiunit dummy coefficient disappears in columns (2)–(4) is probably due to manufacturing being a fairly spatially concentrated activity in Canada. The location choices of manufacturing multiunit establishments do display a strong geographical pattern at a broad

Table 18: Additional robustness checks for the inter-firm linkage regressions.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	NAICS cluster	5-5 $\mathcal{I}dist$	5-5 spec	ln($\mathcal{O}dist$)	ln($\mathcal{O}dist$)	no zeros	levels
multiunit dummy	0.111^a	0.110^a	0.116^a	0.098^a	0.099^a	0.117^a	20.132^a
	(0.019)	(0.014)	(0.015)	(0.014)	(0.014)	(0.017)	(3.217)
core segment dummy	-0.015	-0.020	-0.022	-0.010	-0.012	-0.047^b	-8.026^c
	(0.021)	(0.018)	(0.019)	(0.018)	(0.018)	(0.024)	(4.228)
ln(avg product frequency)	0.006	0.006^b	0.007^b	0.007^a	0.007^a	0.008^b	2.183^a
	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.579)
ln(specialization count)	-0.023^c	-0.026^a	-0.051^a	-0.032^a		-0.016^a	-6.485^a
	(0.014)	(0.002)	(0.003)	(0.002)		(0.004)	(0.667)
ln(specialization employment)					-0.000		
					(0.001)		
Observations	319,319	287,683	277,921	319,319	319,319	201,072	201,072
R-squared	0.551	0.470	0.540	0.606	0.606	0.543	0.406

Notes: Pooled cross section results for manufacturing plants for the years 2001, 2003, 2005, 2007, 2009, 2011, and 2013. All regressions include a full set of industry-year fixed effects. We report selected variables only. All specifications include the following establishment- and firm-level controls: log of plant employment, exporter dummy, headoffice dummy, plant-level diversity measure, firm-level diversity measure. We also include a full set of geographical controls: minimum distance to the US, employment-weighted occupational employment similarity, oes^w, share of highly educated within a15 kilometer distance, share of workers in management or business occupations within a 15 kilometer distance, and a full set of 'urban type' dummies. **Specifications:** **(1)**: standard errors are clustered by 6-digit industries; **(2)**: top and bottom 5% of the dependent variable are trimmed; **(3)**: top and bottom 5% of the specialization variable are trimmed; **(4)** to **(5)**: output distances $\mathcal{O}dist$ are used as the dependent variable; **(6)**: observations where one of the independent variables is zero are dropped (see Appendix A); **(7)**: estimation in levels, not in logs. Robust standard errors, clustered by firm across years (except for **(1)**), in parentheses. Significance levels: ^{a}: $p < 0.01$, ^{b}: $p < 0.05$, ^{c}: $p < 0.1$.

Table 19: Distribution of manufacturing plants by urban type and multiunit status.

	Total	Rural	Urban weak	Urban moderate	Urban strong	Urban CMA
Standalone	289,699	10,124	13,026	18,540	10,120	237,889
Multiunit	31,890	1,127	1,441	1,723	877	26,722
Share of multiunit plants	9.92%	10.02%	9.96%	8.50%	7.97%	10.10%
(excluding 2001)	(9.71%)	(9.66%)				(9.89%)
Total	321,589	11,251	14,467	20,263	10,997	264,611

Notes: Breakdown of the distribution of plant types (standalone or multiunit) by 'urban type' dummies. In 2001, the classification is 'census metropolitan area' or 'rural', detailed urban types are only available starting in 2003.

spatial scale. Yet, conditional on our exhaustive set of continuous geographic controls, the within economic region variation by industry is more limited. Thus, multiunit establishments are found in locations with certain characteristics, but conditional on being overrepresented in those locations the within ER-industry variation does not allow to pick up clear standalone versus multiunit establishment effects. For all plants, which includes non-manufacturing industries that are much less spatially concentrated than manufacturing industries, the multiunit effect survives since there is more within ER-industry variation that allows for identification.[49]

[49] It is unclear (to us at least) how continuous measures of the geographical environment interact with discrete area fixed effects. The reason is that a lot of firms will be associated with 5 or 15 kilometer discs that cut across borders of the discrete areas. Thus, there is an inconsistency in spatial scale in that the same establishment will be associated with variables that span different areas for the purpose of purging unobserved heterogeneity.

Table 20: Inter-firm linkage regressions and additional geographical controls.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
	Mfg	Mfg	Mfg	Mfg	All	All	All	All
	(Base)	(ER effects)	(ER effects)	(ER effects)	(Base)	(ER effects)	(ER effects)	(ER effects)
multiunit dummy	0.111^a	0.031^b	0.001	0.001	0.099^a	0.019^a	0.015^a	0.012^a
	(0.014)	(0.014)	(0.006)	(0.006)	(0.010)	(0.006)	(0.003)	(0.003)
core segment dummy	-0.015	0.018	0.009	0.006	-0.045^a	-0.013^c	-0.003	-0.004
	(0.018)	(0.017)	(0.007)	(0.007)	(0.011)	(0.007)	(0.004)	(0.004)
ln(average product frequency)	0.006^b	0.042^a	-0.003^a	-0.003^a	-0.002	-0.000	0.008^a	0.004^a
	(0.003)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
ln(specialization count)	-0.023^a	-0.003	-0.001	0.000	0.002	-0.003^a	-0.002^b	-0.000
	(0.003)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
mfg × multiunit dummy					-0.030^b	0.071^a	-0.014^b	-0.010
					(0.014)	(0.016)	(0.006)	(0.006)
mfg dummy						-0.040^a		
						(0.002)		
Industry-year fixed effects	Yes	No	No	Yes	Yes	No	No	Yes
Industry fixed effects	No	No	Yes	No	No	No	Yes	No
Year fixed effects	No	Yes	Yes	No	No	Yes	Yes	No
Economic region fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	319,319	319,319	319,319	319,319	901,324	901,324	901,324	901,324
R-squared	0.551	0.794	0.918	0.924	0.586	0.786	0.852	0.885

Notes: Pooled cross section results for manufacturing and non-manufacturing plants for the years 2001, 2003, 2005, 2007, 2009, 2011, and 2013. The dependent variable in all regressions is $\ln(\mathcal{I}dist)$ and it is constructed using the $N = 5$ nearest neighbors in each industry. All specifications include the following establishment- and firm-level controls: log of plant employment, exporter dummy, headoffice dummy, plant-level diversity measure, firm-level diversity measure. We also include a full set of geographical controls: minimum distance to the US, employment-weighted occupational employment similarity, oes^w, share of highly educated within a 15 kilometer distance, share of workers in management or business occupations within a 15 kilometer distance, and a full set of 'urban type' dummies. **Specifications: (1)**: reports the baseline results from column (1) in Table 4; **(2)** to **(4)**: include 86 economic region fixed effects, as well as different combinations of industry and year effects; **(5)** to **(8)**: replicate columns (1) to (4) using all plants. Significance levels: ^{a}: $p < 0.01$, ^{b}: $p < 0.05$, ^{c}: $p < 0.1$.

D.6. Intra-firm linkage regressions for all plants.

Finally, Table 21 summarizes the results for the intra-firm linkage regressions using all establishments, including non-manufacturing. As can be seen, the coefficient on the internal input-output strength is negative and highly significant for the interaction with the manufacturing dummy, but not for non-manufacturing establishments. If anything, the effect seems positive for non-manufacturing establishments, which would be in line with the fact that locational considerations for establishments not involved in the production of goods are much less dependent on good access to intermediate inputs.

Appendix E. A simple model of location choice and spatial sorting by multiunit status

We present a simple model to illustrate the key forces at work in our analysis. Consider two types of plants: multiunit plants (superscripted by m) and standalone plants (superscripted by s). For simplicity, all plants have the same production function, f, irrespective of the indus-

Table 21: Intra-firm linkage regressions for all plants.

	(1)	(2)	(3)	(4)
	unweighted	weighted	unweighted	weighted
log(number of establishments)	0.461^a	0.461^a	0.464^a	0.464^a
	(0.029)	(0.031)	(0.029)	(0.031)
mfg × log(number of establishments)	0.312^a	0.297^a	0.289^a	0.273^a
	(0.051)	(0.045)	(0.049)	(0.044)
headoffice dummy	-0.267^a	-0.235^a	-0.280^a	-0.248^a
	(0.028)	(0.028)	(0.028)	(0.028)
mfg × headoffice dummy	-0.011	-0.049	0.026	-0.008
	(0.046)	(0.047)	(0.046)	(0.046)
log(average product frequency)	-0.132^a	-0.134^a	-0.016	-0.028
	(0.014)	(0.014)	(0.060)	(0.062)
log(specialization count)	0.010	0.017	0.010	0.017
	(0.012)	(0.013)	(0.012)	(0.013)
$\ln(\overline{IO}_{j(i,f),t})$	0.063^c	0.064^c	0.208^a	0.194^b
	(0.036)	(0.035)	(0.076)	(0.077)
mfg × $\ln(\overline{IO}_{j(i,f),t})$	-0.108^b	-0.113^b	-0.201^a	-0.203^a
	(0.045)	(0.045)	(0.056)	(0.057)
core segment dummy	-0.130	-0.131	0.010	0.052^b
	(0.183)	(0.181)	(0.025)	(0.026)
mfg × core segment dummy	-0.067	-0.022		
	(0.256)	(0.255)		
core segment dummy × $\ln(\overline{IO}_{j(i,f),t})$	-0.012	-0.017		
	(0.024)	(0.024)		
mfg × core segment dummy × $\ln(\overline{IO}_{j(i,f),t})$	-0.046	-0.043		
	(0.038)	(0.038)		
log(avg product frequency) × $\ln(\overline{IO}_{j(i,f),t})$			0.020^b	0.018^b
			(0.008)	(0.008)
mfg × log(avg product frequency) × $\ln(\overline{IO}_{j(i,f),t})$			-0.010^b	-0.010^c
			(0.005)	(0.005)
Observations	131,559	131,559	131,559	131,559
R-squared	0.376	0.358	0.376	0.358

Notes: Pooled cross section results for the years 2001, 2003, 2005, 2007, 2009, 2011, and 2013. The dependent variables in the regressions are either $\ln(\overline{d}_{j(i,f),t})$ or $\ln(\overline{d}^{w}_{j(i,f),t})$. All regressions include a full set of industry-year fixed effects. We report selected variables only. All specifications include the following establishment- and firm-level controls: log of plant employment, exporter dummy, plant-level diversity measure, firm-level diversity measure, monoindustry dummy. We also include a full set of geographical controls: minimum distance to the US, employment-weighted occupational employment similarity, oes^w, share of highly educated within a 15 kilometer distance, share of workers in management or business occupations within a 15 kilometer distance, and a full set of 'urban type' dummies. **Specifications: (1)** and **(2)**: interactions with core segment dummy; **(3)** and **(4)**: interactions with product ubiquity measure. Robust standard errors, clustered at the firm level, reported in parentheses. Significance levels: ^{a}: $p < 0.01$, ^{b}: $p < 0.05$, ^{c}: $p < 0.1$.

try they operate in and of their multiunit status. They choose their inputs $X = (X_1, X_2, \ldots, X_n)$ and location, ℓ. Each location offers some local agglomeration benefits, ε^ℓ (e.g., local labor market thickness, local access to intermediates, and knowledge spillovers). Furthermore, for each multiunit plant, location ℓ is at a distance $\overline{\delta}_\ell$ from the other plants of the firm (including the plant's headquarter). The profit of a standalone plant and of a multiunit plant in location ℓ are given by:

$$\pi^s(X,\ell) = A^s(\varepsilon^\ell) f(X^s) - \sum_i P_i^s(\varepsilon^\ell) X_i^s \quad \text{and} \quad \pi^m(X,\ell) = A^m(\varepsilon^\ell, \overline{\delta}_\ell) f(X^m) - \sum_i P_i^m(\varepsilon^\ell, \overline{\delta}_\ell) X_i^m,$$

respectively. In the above expression, $A^m(\cdot)$ and $A^s(\cdot)$ denote the TFP of the two types of plants. We assume that

$$\frac{\partial A^s}{\partial \varepsilon^\ell} > \frac{\partial A^m}{\partial \varepsilon^\ell} > 0, \ \forall \overline{\delta}_\ell \quad \text{and} \quad \frac{\partial A^m}{\partial \overline{\delta}_\ell} < 0. \tag{E-1}$$

In words, both types benefit from agglomeration economies, but standalone plants more than multiunit plants. Also, multiunit plants are less productive if they are more remote from the other plants of the firm (see, e.g., Giroud, 2013; Kalnins and Lafontaine, 2014). Because more productive clusters have higher factor prices (see Greenstone *et al.*, 2010), whereas remote transactions within firms are impeded by distance, we also assume that

$$\frac{\partial P_i^s}{\partial \varepsilon^\ell} = \frac{\partial P_i^m}{\partial \varepsilon^\ell} > 0, \ \forall \overline{\delta}_\ell \quad \text{and} \quad \frac{\partial P_i^m}{\partial \overline{\delta}_\ell} > 0. \tag{E-2}$$

Assume, for simplicity, that the possible set of locations can be described by the couples $(\varepsilon_\ell, \overline{\delta}_\ell(\varepsilon_\ell))$. In words, firms pick locations based on two characteristics that are functionally related.[50] The first-order conditions for standalone and for multiunit plants with respect to location choices are then:

$$\frac{\partial \pi^s}{\partial \varepsilon_\ell} = \frac{\partial A^s}{\partial \varepsilon_\ell} f(X^s) - \sum_i \frac{\partial P_i^s}{\partial \varepsilon_\ell} X_i^s = 0$$

and

$$\frac{\mathrm{d}\pi^m}{\mathrm{d}\varepsilon_\ell} = \left(\frac{\partial A^m}{\partial \varepsilon_\ell} + \frac{\partial A^m}{\partial \overline{\delta}_\ell} \frac{\mathrm{d}\overline{\delta}_\ell}{\mathrm{d}\varepsilon_\ell} \right) f(X^m) - \sum_i \left(\frac{\partial P_i^m}{\partial \varepsilon_\ell} + \frac{\partial P_i^m}{\partial \overline{\delta}_\ell} \frac{\mathrm{d}\overline{\delta}_\ell}{\mathrm{d}\varepsilon_\ell} \right) X_i^m = 0.$$

Taking the difference, we get

$$\begin{aligned} \frac{\mathrm{d}\pi^m}{\mathrm{d}\varepsilon_\ell} - \frac{\partial \pi^s}{\partial \varepsilon_\ell} &= \left(\frac{\partial A^m}{\partial \varepsilon_\ell} + \frac{\partial A^m}{\partial \overline{\delta}_\ell} \frac{\mathrm{d}\overline{\delta}_\ell}{\mathrm{d}\varepsilon_\ell} \right) f(X^m) - \frac{\partial A^s}{\partial \varepsilon_\ell} f(X^s) \\ &\quad - \sum_i \left[\frac{\partial P_i^m}{\partial \varepsilon_\ell} (X_i^m - X_i^s) + \frac{\partial P_i^m}{\partial \overline{\delta}_\ell} \frac{\mathrm{d}\overline{\delta}_\ell}{\mathrm{d}\varepsilon_\ell} X_i^m \right] = 0, \end{aligned} \tag{E-3}$$

where we have made use of (E-2). The signs of both terms are a priori undetermined. The first line of (E-3) tends to be negative if the gap between $\frac{\partial A^s}{\partial \varepsilon^\ell}$ and $\frac{\partial A^m}{\partial \varepsilon^\ell}$ is large enough, whereas the difference in the production scales is not too large. It is also more likely to be negative if there is a positive correlation between the agglomeration benefits of locations and their remoteness from the other plants of the multiunit firm ($\mathrm{d}\overline{\delta}_\ell / \mathrm{d}\varepsilon_\ell > 0$). The second line of (E-3) is more likely to be negative if the plants' input choices are not too different, and if $\mathrm{d}\overline{\delta}_\ell / \mathrm{d}\varepsilon_\ell > 0$. Hence, standalone plants are generally more profitable than comparable multiunit plants in locations that offer more agglomeration benefits and that are farther away from the rest of the multiunit firms. Of course, the sign of $\mathrm{d}\overline{\delta}_\ell / \mathrm{d}\varepsilon_\ell$ is an empirical question.

[50] In the case of two independent characteristics, not much can be said about location choices.