DISCUSSION PAPER SERIES

DP11884

THE DETERMINANTS OF COAGGLOMERATION: EVIDENCE FROM FUNCTIONAL EMPLOYMENT PATTERNS

Kristian Behrens and Rachel Guillain

INTERNATIONAL TRADE AND REGIONAL ECONOMICS

ISSN 0265-8003

THE DETERMINANTS OF COAGGLOMERATION: EVIDENCE FROM FUNCTIONAL EMPLOYMENT PATTERNS

Kristian Behrens and Rachel Guillain

Discussion Paper DP11884
Published 28 February 2017
Submitted 28 February 2017

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

THE DETERMINANTS OF COAGGLOMERATION: EVIDENCE FROM FUNCTIONAL EMPLOYMENT PATTERNS

Abstract

Locations differ horizontally in the industry mix they host and vertically in the value-chain functions they perform. Since industry pairs should coagglomerate the functions that interact intensively across industries, analyzing horizontal and vertical patterns can improve our understanding of agglomeration mechanisms. We find that different functions within the same industry pairs display substantially different coagglomeration patterns. While production coagglomerates at longer distances, management and research coagglomerates at short distances. These patterns are consistent with our findings that buyer-supplier links and local labor pools are important for production, whereas they matter less for management and research that rely on shared knowledge. Our results provide support for agglomeration theories and show that extant estimates of average effects based on total employment mask substantial heterogeneity.

JEL Classification: R12, L60

Keywords: coagglomeration, functional specialization, agglomeration mechanisms, Duranton-Overman index

Kristian Behrens - behrens.kristian@uqam.ca
Université du Québec à Montréal (ESG-UQAM), Canada; National Research University Higher School of Economics, Russian Federation and CEPR

Rachel Guillain - guillain@u-bourgogne.fr
Université de Bourgogne Franche-Comté, France

Acknowledgements
We thank our discussant, Ferdinando Monte, as well as Nate Baum-Snow, Théophile Bougna, Mark Brown, Don Davis, Gilles Duranton, Amit Khandelwal, Bill Kerr, Fabian Lange, Julien Martin, Richard Shearmur, Will Strange, and seminar and conference participants at LMU Munich, the 2016 Rotman-Sauder Summer Conference in Real Estate and Urban Economics, the 2015 NARSC Meetings in Portland, the World Bank conference on secondary towns, CIRANO Montréal, HSE Saint Petersburg, and Columbia University for valuable comments and suggestions. We are grateful to Richard Shearmur and Mario Polèse at INRS Montréal (Urbanisation Culture Société) for sharing the special census tabulations from Statistics Canada; and to Bill Kerr for sharing the patent citation data. 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. Guillain gratefully acknowledges financial support from the PARI Programs of the 'Conseil Régional de Bourgogne'. The study has been funded by the Russian Academic Excellence Project '5-100'. The views expressed in this paper, and all remaining errors, are ours.

The determinants of coagglomeration: Evidence from functional employment patterns

Kristian Behrens* Rachel Guillain†

February 25, 2017

Abstract

Locations differ horizontally in the industry mix they host and vertically in the value-chain functions they perform. Since industry pairs should coagglomerate the functions that interact intensively across industries, analyzing horizontal and vertical patterns can improve our understanding of agglomeration mechanisms. We find that different functions within the same industry pairs display substantially different coagglomeration patterns. While production coagglomerates at longer distances, management and research coagglomerates at short distances. These patterns are consistent with our findings that buyer-supplier links and local labor pools are important for production, whereas they matter less for management and research that rely on shared knowledge. Our results provide support for agglomeration theories and show that extant estimates of average effects based on total employment mask substantial heterogeneity.

Keywords: coagglomeration; functional specialization; agglomeration mechanisms; Duranton-Overman index.

JEL Classification: R12; L60.

*Department of Economics, Université du Québec à Montréal (ESG-UQAM), Canada; National Research University Higher School of Economics, Russian Federation; and CEPR, UK. E-mail: behrens.kristian@uqam.ca.

†Université de Bourgogne Franche-Comté, France. E-mail: guillain@u-bourgogne.fr.

1 Introduction

The most salient feature of the economic landscape is how much locations differ in their size and density of economic activity. An almost equally salient feature is how much locations differ in their horizontal composition, i.e., their industry mix. Because some industries tend to operate in many locations, whereas others display strong geographic concentration patterns, some areas are highly specialized in a narrow set of industries while others have a very broad economic tissue. A third salient feature is that, even conditional on their industry mix, locations differ in their horizontal composition: different locations play different roles in the value chain. Following secular changes in technology and the internal organization of firms, many functions like production, management, or research and development can be geographically separated. Consequently, locations are both horizontally differentiated in the industry mix they host, and vertically differentiated in the value-chain functions they perform.

There is no shortage of theories explaining why economic activity tends to agglomerate (e.g., Fujita and Thisse, 2002; Duranton and Puga, 2004; Behrens and Robert-Nicoud, 2015), why different industries tend to colocate (e.g., Anas and Xiong, 2003; Abdel-Rahman and Anas, 2004; Helsley and Strange, 2014), and why locations tend to specialize in different functions in the value chain (e.g., Duranton and Puga, 2001, 2005; Davis and Dingel, 2014). The geographic concentration and functional specialization patterns are the outcome of firms' efforts to minimize the costs of moving goods, people, and ideas. Since these costs are jointly relevant, assessing their relative importance for observed geographic patterns is an empirical question.[1] Although the location patterns of industries *and* functions both subsume relevant information on agglomeration mechanisms, the two have been considered mostly separately until now.[2] We aim to fill that gap by looking at the geographic concentration of functions within industry pairs using detailed microgeographic data. In doing so, we provide new insights into the relative importance of the agglomeration mechanisms underlying the spatial economy.

[1] Agglomeration is driven by multiple factors. The literature usually focuses on the three 'Marshallian' agglomeration forces: input-output links, common labor pools, and knowledge sharing (see Rosenthal and Strange, 2004; and Combes and Gobillon, 2015, for surveys). Yet, 'adaptative aspects' (Duranton and Puga, 2001; Strange, Hejazi, and Tang, 2006), and 'organizational aspects' (Rosenthal and Strange, 2003, 2010) have increasingly been investigated, especially in the strategic management literature (e.g., Alcácer, 2006; and Alcácer and Delgado, 2017).

[2] The empirical literature at the intersection of agglomeration and functional specialization is thin at best. Audretsch and Feldman (1996) document that R&D activities in U.S. industries are geographically concentrated, but that this concentration is mostly orthogonal to that of production. Faggio, Silva, and Strange (2017) dissect coagglomeration patterns along dimensions that reflect organizational differences across industries, yet their analysis does not directly speak to the issue of functional specialization along the value chain. Closer to our analysis, Bade, Bode, and Cutrini (2015) document the coagglomeration patterns of broad functions using German data. Their analysis is, however, geographically more aggregated and purely descriptive. Last, and closest to our work, Alcácer (2006) analyzes firm-level colocation patterns of subsidiaries by separating R&D, production, and sales. His analysis, however, focuses on a specific industry (the cellular handset industry) only.

Previewing our key results, we find that heterogeneity in the location patterns of functions in the value chain provides useful information for the identification of agglomeration mechanisms. The reason is that different functions benefit to a different degree from those different mechanisms, as emphasized in the strategic management literature (e.g., Alcácer, 2006). While the location of production is, for example, sensitive to the presence of vertically linked industries and the composition of local labor pools, these are less important for the location of management and research, which are more sensitive to shared knowledge. Previous studies have considered all functions jointly, thereby estimating the average effects of the agglomeration forces (e.g., Rosenthal and Strange, 2001; Ellison, Glaeser, and Kerr, 2010). These average effects can be misleading. While our proxy for knowledge sharing is, e.g., positively associated with the coagglomeration of total employment in 22% of the specifications that we estimate, the corresponding figures are 68% for the coagglomeration of employment in management and research, but only 3% for employment in production. Consistent with that result, we also find that employment in manangement and research and in production display markedly different spatial profiles of coagglomeration across industry pairs. While the former is mainly coagglomerated at distances of less than 50 kilometers, the latter is mainly coagglomerated at longer distances of about 150-200 kilometers. This reflects the relative importance of knowledge sharing — which operates at shorter distances — for management and research; and the relative importance of vertical links — which operate at longer distances — for production.

Understanding the mechanisms that drive the geographic concentration and the mix of industries and functions across locations is important for at least two reasons. First, it is increasingly recognized that the local economic tissue is an important determinant of many socio-economic outcomes. The agglomeration and coagglomeration of industries drive productivity gains, and speed innovation and technological change (see Combes and Gobillon, 2015; Carlino and Kerr, 2015). They also determine local employment structures and skills, which are important to understand regional inequality and exposure to trade shocks (e.g., Autor, Dorn, and Hanson, 2013). The latter are drivers of a wide range of politically loaded outcomes ranging from firm exit and worker out-migrations to 'rust belt formation' and increased suicide rates among specific subgroups of the population (Pierce and Schott, 2016). Understanding the local determinants of employment composition is important to predict exposure, and the local employment composition is largely driven by agglomeration forces that affect the spatial distribution of industries and functions. Second, the success of local development policies crucially hinges on a good understanding of how changes in the costs of moving goods, people, and ideas will change the size, composition, and function of the local economic tissue. Subsidizing infrastructure to attract specific high-tech industries may attract only their blue-collar jobs, whereas subsidizing knowledge networks may attract only their white-collar jobs. We therefore need evidence on how specific mechanisms jointly affect both the industry mix and functional composition of locations to make better policy recommendations.

Analyzing the mechanisms underlying the location patterns of industry pairs by functions in the value chain is challenging, especially at a microgeographic level. First, it requires access to data that allow us to split plant-level employment by functional type. We achieve this by combining two datasets: detailed microgeographic establishment-level data for Canadian manufacturing plants in 2001, 2003, and 2005; and somewhat more spatially aggregated special census tabulations that split employment by function and by industry within census divisions. These special tabulations allow us to cleanly separate urban from rural areas, thus providing a more accurate picture of the spatial distribution of functions in the value chain. Second, the computational requirements are challenging. We compute microgeographic coagglomeration measures based on point patterns for almost 150,000 industry-function-year pairs, following Duranton and Overman (2005). To our knowledge, we provide the first comprehensive analysis of coagglomeration patterns — and the spatial extent of the mechanisms underlying them — using continuous coagglomeration measures and functional employment splits.[3] Last, we need to address a large number of thorny identification issues, some of which have not been recognized in the literature until now. While the literature has traditionally tried to control for natural advantage and has used instrumentation strategies to deal with reverse causality (Ellison and Glaeser, 1999; Ellison *et al.*, 2010), more subtle confounding factors such as general equilibrium location patterns of multiple industries, coagglomeration of different activities 'within plants', firms' ability to split plants and functions across locations, and unobserved heterogeneity at the industry level have not been considered extensively until now. We show that all of these factors are relevant and should be controlled for in the analysis.

The remainder of this paper is organized as follows. In Section 2, we present some theoretical considerations on how agglomeration forces affect the locations of industries and functions. In Section 3, we present our empirical methodology. Section 4 provides details on our variables and discusses various identification concerns. Section 5 contains our empirical results. Section 6 discusses the limitations and potential shortcomings of our analysis. Finally, Section 7 concludes. We relegate many technical details and a detailed description of our data and variables to an extensive set of appendices. A set of supplemental appendices contains further technical developments, additional results, and robustness checks.

[3] Duranton and Overman (2005, 2008) compute coagglomeration measures based on point patterns for a small subset of UK industries, without investigating their determinants. Ellison *et al.* (2010) look at those determinants, using either the Ellison-Glaeser measure of coagglomeration or a 'lumpy' approximation of the Duranton-Overman measure at the U.S. county level. This amounts to working at a larger spatial scale and in a discrete setting. Howard, Newman, and Tarp (2016) look at coagglomeration patterns and their determinants across discrete regions in Vietnam. Faggio *et al.* (2017) use UK data aggregated to travel-to-work areas, thereby also working in a discrete setting. Other contributions include Kolko (2010), who looks at the coagglomeration of service industries in the U.S., and Gabe and Abel (2016) who look at the coagglomeration of occupations that share common knowledge bases. None of those papers looks at individual functions in the value chain, and none of them has much to say about the spatial extent of the agglomeration mechanisms.

2 Theoretical framework

Assume that firms use $f \in \mathcal{F}$ different functional labor types like management or production workers. A firm with productivity z, in industry $i \in \mathcal{I}$ and city $c \in \mathcal{C}$, solves the cost-minimization problem $\min_{\{\ell^f(z), \forall f \in \mathcal{F}\}} \sum_f w_c^f \ell^f(z)$ s.t. $y_{ic}(z) = 1$.[4] The production function is given by

$$y_{ic}(z) = \left[\sum_{f \in \mathcal{F}} \alpha_{ic}^f \cdot \left(z^{\phi^f} \ell^f \right)^{\frac{\sigma_i - 1}{\sigma_i}} \right]^{\frac{\sigma_i}{\sigma_i - 1}}, \tag{1}$$

where σ_i is the industry-specific elasticity of substitution between functions; w_c^f is the wage of function f in city c; ϕ^f is a parameter that affects the functional composition of the firms; and α_{ic}^f is an industry-function-city-specific production shifter, taken as given by the firms. The optimal input share of each labor type f is given by

$$\theta_{ic}^f(z) \equiv \frac{\ell^f(z)}{\sum_{f'} \ell^{f'(z)}} = \frac{\left(\alpha_{ic}^f / w_c^f \right)^{\sigma_i} z^{(\sigma_i - 1)\phi^f}}{\sum_{f'} \left(\alpha_{ic}^{f'} / w_c^{f'} \right)^{\sigma_i} z^{(\sigma_i - 1)\phi^{f'}}}. \tag{2}$$

A larger ϕ^f increases the share of function f in higher productivity firms.[5] With identical ϕ^f parameters, there is no such change in the functional composition.

As can be seen from (2), $\theta_{ic}^f(z)$ depends positively on α_{ic}^f. Firms take α_{ic}^f as given, yet this term varies across cities. They hence pick city c that minimizes production costs, conditional on (2). Let N_c denote a vector of locational characteristics such as infrastructure, access to natural resources, or proximity to foreign markets. Let $\mathbf{L_c} = \{L_{ic}^f\}_{i \in \mathcal{I}, f \in \mathcal{F}}$ denote the vector of local employment of function f in industry i and city c, denoted L_{ic}^f. Following a long tradition in urban economics, we assume that $\alpha_{ic}^f \equiv \alpha_i^f(N_c, \mathbf{L}_c)$.[6] The literature usually considers that $\partial \alpha_i^f(N_c, \mathbf{L_c}) / \partial L_{ic}^f > 0$, i.e., the productivity of function f in industry i increases with local employment in that function and industry ('localization economies'). Another case considered in the literature is where α_{ic}^f increases with total employment in function f in c, regardless of the industry ('urbanization economies'). Following Helsley and Strange (2014), we can also consider more complex and realistic cases. In what follows, we place no restrictions on

[4] In what follows, we refer to locations as 'cities'. In the empirical analysis, locations will be points on the map.

[5] Alternatively, z can be a plant-specific characteristic (e.g., employment or sales) that is correlated with productivity and that influences the plant's functional labor shares. Larger firms have a different functional structure than smaller firms. Caliendo, Monte, and Rossi-Hansberg (2015) document that large French firms have more hierarchical layers and employ thus more higher-level functions. Davidson, Heyman, Matusz, Sjöholm, and Zhu (2016) show that multinationals have an occupational structure that is skewed towards highly skilled workers. In our application, z is employment, which is positively correlated with productivity. Firms with more employment have a larger share of non-production workers.

[6] See Duranton and Puga (2004), and Behrens and Robert-Nicoud (2015), for reviews of the theory; and Rosenthal and Strange (2004), and Combes and Gobillon (2015), for reviews of the empirical evidence.

the cross-partial derivatives $\partial\alpha_i^f(N_c, \mathbf{L_c})/\partial L_{jc}^f$ that capture how productivity of function f in industry i depends on local employment of function f in industry j. If $\partial\alpha_i^f(N_c, \mathbf{L_c})/\partial L_{jc}^f > 0$, we say that function f displays *coagglomeration economies* between industries i and j.[7] In that case, productivity of function f in industry i is increasing in local employment of function f in industry j, which provides a rationale for the coagglomeration of function f in industries i and j in the same cities c.

Three comments are in order. First, the determinants of coagglomeration and their strength are likely to vary by industry pairs and functions. For example, colocating the production functions of car manufacturers and automobile parts suppliers in the same locations may increase productivity via buyer-supplier links that allow to save transport costs. Colocating the management functions of those same industries may increase productivity via knowledge exchange and the sharing of industry-specific information. Hence, the colocation of each function f between industries i and j should, at least partly, reflect the benefits of greater geographic proximity. The determinants may differ across functions, which provides additional identifying variation in the data that we will exploit.

Second, as shown by Helsley and Strange (2014), the coagglomeration of industries cannot be Pareto efficient under reasonable assumptions. Desirable colocation may not take place, or industries may end up colocated even if $\partial\alpha_i^f(\mathbf{L_c})/\partial L_{jc}^f = 0$. This may obviously be the case when $\alpha_i^f(N_c)$ and $\alpha_j^f(N_c)$ are both increasing functions of N_c.[8] In that case, $\theta_{jc}^f(z)$ is an increasing function of the locational characteristics N_c, so that employment in industry j is also an increasing function of those characteristics. Since α_{ic}^f is increasing in N_c, it then follows that $\mathrm{d}\alpha_i^f(N_c, \mathbf{L_c})/\mathrm{d}L_{jc}^f > 0$ just because the productivity shifters in both industries are positively correlated with N_c. This makes it harder to identify the determinants of the coagglomeration patterns (see Ellison and Glaeser, 1999; Ellison *et al.*, 2010). More subtly, this relationship is also transitive. If α_{ic}^f and α_{jc}^f are both increasing in N_c, and if $\mathrm{d}\alpha_k^f(N_c, \mathbf{L_c})/\mathrm{d}L_{jc}^f > 0$ because of the existence of coagglomeration economies, then the employment of i and k will also covary across space just because i and j and j and k happen to colocate. These *third-industry effects* can induce spurious coagglomeration and need to be controlled for in the empirical analysis. We return to that point in more detail below.

Last, firms can split their functions across plants and cities to better exploit locational advantage such as factor costs (e.g., Fujita and Thisse, 2006) or different agglomeration economies over their life cycle (see Duranton and Puga, 2001, 2005). Expression (2) above assumes that firms pick only a single location c and split their employment across functions based on factor costs, w_c^f, and coagglomeration economies, as subsumed by α_{ic}^f. It neglects the fact that larger

[7] When there is only one function, and when firms and locations are homogeneous ($\phi^f = N_c = 0$), our specification is isomorphic to that in Helsley and Strange (2014): $y_{ic} = g_i(\mathbf{L_c})\ell$, with $g(\mathbf{L_c}) = \alpha_i(\mathbf{L_c})^{\sigma_i/(\sigma_i-1)}$. Our specification adds different functions and firm-level heterogeneity to the basic setup.

[8] See Behrens (2016) for a diagrammatic exposition of the arguments that follow.

multiunit firms are generally less dependent on their economic environment when it comes to exploiting coagglomeration economies (e.g., Rosenthal and Strange, 2010; Brown and Rigby, 2015). Furthermore, as can be seen from (2), the largest α_{ic}^f / w_c^f term may determine the location choice, so that two industries may colocate because of one function. Since all functions are then colocated at the same site — by assumption of single-unit firms — this then induces spurious colocation for the remaining functions. The net effect of these two opposing biases is a priori unclear.[9]

3 Empirical methodology

Let $p(i,c)$ denote plant p in industry i and city c. Our objective is to use the cross-city distribution of employment by function f in industries i and j, given by

$$L_{ic}^f = \sum_{p(i,c)} \theta_{p(i,c)}^f \ell_{p(i,c)}^f \quad \text{and} \quad L_{jc}^f = \sum_{p(j,c)} \theta_{p(j,c)}^f \ell_{p(j,c)}^f, \tag{3}$$

to learn more about coagglomeration economies, $\partial \alpha_i^f(\mathbf{L_c}) / \partial L_j^f$. To this end, we first compute a measure of coagglomeration of employment by function f for industry pairs ij across cities c. Let coagglo_{ijt}^f denote that measure in year t. We provide details on how we construct those measures in subsection 4.1.1 below. We then follow Ellison *et al.* (2010) and Faggio *et al.* (2017) and run various regressions of the following form:

$$\text{coagglo}_{ijt}^f = \alpha_{io}\text{io}_{ijt} + \alpha_{oes}\text{oes}_{ijt}^f + \alpha_{pat}\text{pat}_{ijt} + \mathbf{X}_{ijt}\beta + \xi_{(i,j)} + \delta_t + \epsilon_{ijt}, \tag{4}$$

where $\mathbf{X}_{ijt}$ is a vector of time-varying industry controls; where δ_t are year fixed effects; and where $\xi_{(i,j)}$ are industry fixed effects.[10] Our key variables of interest are the proxies for the Marshallian agglomeration forces: input-output links (io_{ijt}), labor market pooling (oes_{ijt}^f), and knowledge spillovers (pat_{ijt}). If $\mathbf{X}_{ijt}$ and $\xi_{(i,j)}$ in (4) adequately capture confounding factors like natural advantage, third-industry effects, and aspects of industrial organization, and if our explanatory variables are uncorrelated with the error term, then our estimates are the causal effect of the Marshallian factors on the coagglomeration of industries. We discuss in greater detail the identification challenges and strategies in Section 4.

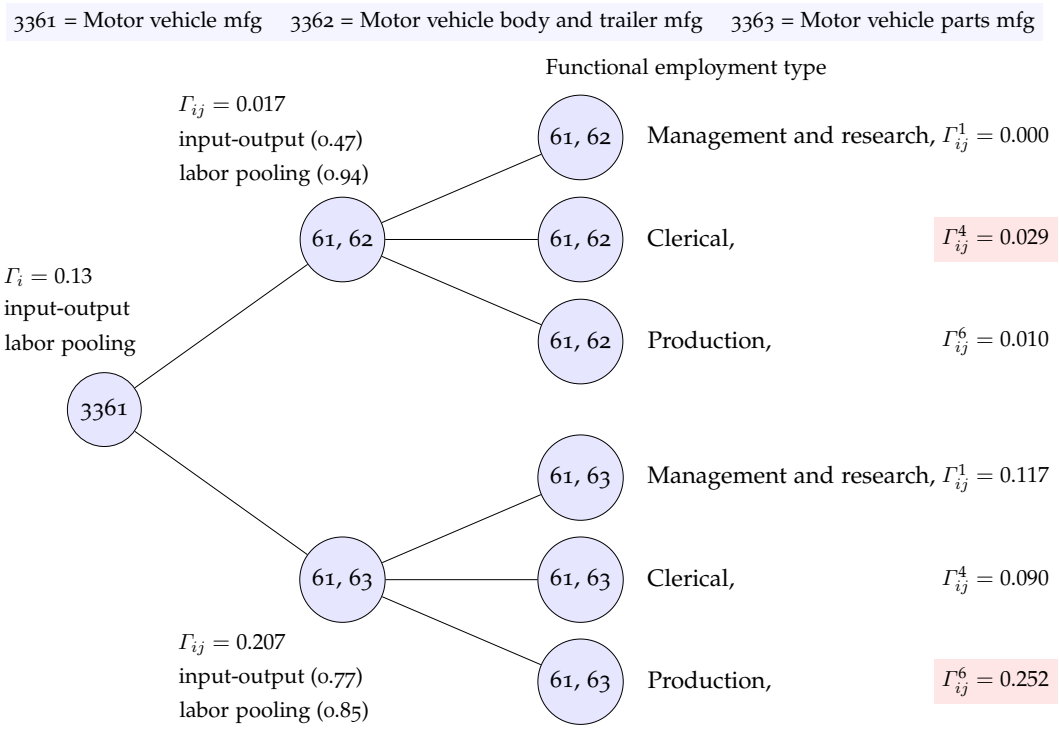
As shown by (4), our empirical approach exploits the differences in the coagglomeration patterns of functional employment types within industry pairs. Using both coagglomeration

[9] We present a simple multi location extension of our framework in the Supplemental Appendix S.1. As is well known, the multi location problem is a complex combinatorial problem, and any more elaborate estimation work on it would require us to know location-function choices at the firm level, which we do not observe.

[10] We construct 'compound' fixed effects as in Ellison *et al.* (2010). We could estimate the same regression using 'standard' industry i and j fixed effects. As shown in the Supplemental Appendix S.2, this yields identical estimates than those with compound fixed effects, provided that the order of the i and j indices in each unique industry pair is assigned sufficiently randomly.

patterns and functional employment splits offers two key advantages for the identification of the determinants of agglomeration. Figure 1 below illustrates these points using a simple example drawn from our data.

Figure 1: Refining the identifying variation of the determinants of agglomeration.



Notes: We report Γ_{ij} as the 'excess coagglomeration' measure of Duranton and Overman (2005) over all distances from 1 to 800 kilometers (see Appendix B.1 for details). Figures are for 2001 and for our baseline functional shares. The value of the Duranton-Overman agglomeration measure Γ_i is from Behrens and Bougna (2015) for the year 2001 (using a slightly different but comparable sample of the same dataset).

First, the coagglomeration of industry pairs — as compared to the agglomeration of individual industries — allows us to exploit industry-pair specific variations in the importance of the determinants of agglomeration. Consider 'Motor vehicle manufacturing' (NAICS 3361). That industry has an agglomeration measure of $\Gamma_i = 0.13$ in 2001, and it is significantly geographically concentrated in Canada (Behrens and Bougna, 2015). That geographic concentration is likely driven by all three Marshallian forces, yet it is hard to isolate their role using only the information of that industry. For example, 'Motor vehicle manufacturing' buys a lot of inputs from some industries, and little from others. Hence, industry measures such as the total value of intermediate inputs as a share of sales of the industry may be weak proxies for agglomeration forces. The same holds true for industry-specific measures of labor market pooling or the importance of knowledge such as R&D expenditures of the industry.

By contrast, coagglomeration measures exploit variations in the colocation patterns of 'Motor vehicle manufacturing' *with many other industries*. Going back to our example, the colocation

patterns of NAICS 3361 with 'Motor vehicle body and trailer manufacturing' (NAICS 3362) and 'Motor vehicle parts manufacturing' (NAICS 3363) are quite different. While the colocation between NAICS 3361 and 3362 is rather weak, with a coagglomeration measure of $\Gamma_{ij} = 0.017$, the colocation between NAICS 3361 and 3363 is fairly strong, with $\Gamma_{ij} = 0.207$. As can be seen from Figure 1, labor market pooling is relatively more important for the former industry pair ($0.47/0.94 = 0.5$); whereas input-output links are relatively more important for the latter industry pair ($0.77/0.85 = 0.91$). This suggests that the former industry pair should display stronger coagglomeration patterns for functions that are strongly dependend on a similar labor force, whereas the latter industry pair should display stronger coagglomeration patterns for functions that are strongly dependend on customer-supplier relationships.

This is where variations in the coagglomeration patterns by functional employment types become useful. As can be seen from the right part of Figure 1, when breaking down total employment into management and research, clerical, and production, the coagglomeration of NAICS 3361 and 3362 is stronger for clerical employment, with $\Gamma^4_{ij} = 0.029$, than for the other employment types. On the contrary, for NAICS 3361 and 3363, the coagglomeration is stronger for production employment, with $\Gamma^6_{ij} = 0.234$, than for the other employment types. These patterns are in line with our priors that: (i) input-output links are especially important for the colocation of production functions, and less important for the colocation of other employment types; and (ii) knowledge sharing and labor market pooling are more important for the colocation of management and research or clerical employment than for production employment. Since different functions are likely to depend to different degrees on different agglomeration forces, pooling all functions to estimate the 'average' determinants of agglomeration does not allow to finely identify the drivers underlying the observed agglomeration patterns. This is especially true for functions like management and research that constitute only a small share of total employment of many manufacturing industries.

4 Variables and identification strategy

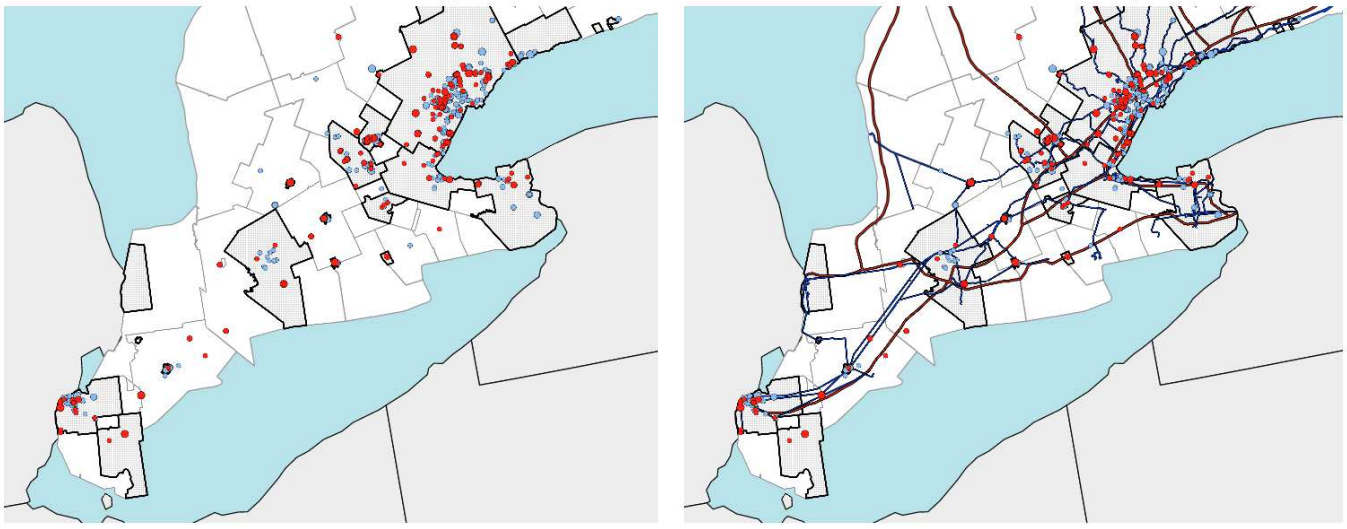
We now explain in detail the construction of our key variables and our identification strategy. A detailed description of the data, the sources, and of several methodological choices are relegated to Appendices A and B. As explained above, our approach exploits the variation of coagglomeration patterns across functions (different employment types) for different industry pairs and relates those to measures of the Marshallian agglomeration forces. Our analysis hence requires two key pieces of information: coagglomeration patterns by functional employment types; and measures of the Marshallian agglomeration forces. We now explain how we construct those two key pieces.

4.1 Construction of the variables

4.1.1 Coagglomeration of functions

To estimate (4), we first construct our measures of coagglomeration. Since we have data that allow us to geolocate manufacturing plants by 6-digit postal code centroid, we estimate the point-pattern based continuous coagglomeration measures of Duranton and Overman (2005, 2008) — henceforth DO K-densities — as explained in greater detail in Appendix B.1.[11] Figure 2 below illustrates the nature of our geographic data for the case of the south-western part of Ontario. The left panel depicts the colocation patterns of plants in 'Motor vehicle manufacturing' (NAICS 3361; red dots), and 'Motor vehicle parts manufacturing' (NAICS 3363; blue dots). Using that data, we compute one coagglomeration measure for each unique 4-digit pair of the 86 manufacturing industries (i.e., $86 \times 85/2 = 3,655$ unique industry-pair coagglomeration measures for each function).

Figure 2: Coagglomeration of NAICS 3361 and 3363, and major infrastructure.



Notes: The left panel depicts the locations of plants in NAICS 3361 (red dots) and 3363 (blue dots). Dots are proportional to employment size. The dark shaded census divisions are either census metropolitan areas, or the urban parts of the census divisions as reported in the special tabulations. The right panel adds major road (red lines) and rail (blue lines) infrastructure.

Using our estimates of the DO K-densities, we then construct two different measures of the extent of coagglomeration of industry pairs that are our dependent variable. As explained in Appendix B.1, our first measure is based on the cumulative distribution of the K-densities, whereas the second is based on the 'excess coagglomeration' of industry pairs. The first measure, $\text{DO\_CDF}_{ij}(d)$, gives the share of bilateral distances between employees in the two indus-

[11] We use continuous microgeographic coagglomeration measures — instead of discrete ones such as the Ellison and Glaeser (1997) index — to obviate the need for arbitrary administrative units. Continuous spatial measures also allow us to assess the effects of coagglomeration at various spatial scales.

tries that is less than d kilometers from each other. The second measure, $\text{DO\_STR}_{ij}(d)$, gives that same information but only for the share of employees agglomerated in 'excess' of what we would observe under a random distribution of plants. Following Duranton and Overman (2005), the benchmark distribution is obtained by randomly reshuffling the plants in industries i and j across all locations occupied by plants in either industry i or industry j. Hence, our second measure controls already for the geographic concentration of the two industries and only looks at how much closer pairs of plants in those two industries are, conditional on their overall structure. This second measure is hence a measure of *relative coagglomeration*, whereas the first measure is one of *absolute coagglomeration*. Both measures can be defined for an arbitrary distance d between 0 and 800 kilometers, where the latter corresponds to the range over which we compute them (see Behrens and Bougna, 2015, for details). In what follows, we will use a baseline distance of $d = 25$ kilometers in our regression analysis, but we also present results where we smoothly vary the distance threshold to look at the spatial extent of the different agglomeration mechanisms.

While computing the coagglomeration measures for overall employment is straightforward, computing them by employment type is more challenging since we observe only total employment — and not the functional splits of employment — at the plant level. We hence combine census division special tabulations from Statistics Canada, which provide a split of employment by industry and function in each census division, with our microgeographic plant-level data.[12] One distinct advantage of our data is that it separately tabulates data for rural and urban parts of some of the census divisions (see Appendix A.1 for details). This is important as cities increasingly specialize by functions rather than by industries.[13] Hence, plants in the same census division are likely to have a different mix of worker types, based on their rural or urban location. The left panel of Figure 2 illustrates the spatial units for which we have employment splits across both industries and functions. The dark shaded (parts of) the census divisions highlight census metropolitan areas or urban parts of census divisions, whereas the other census division are not classified as urban. We present details on the rural-urban specialization differences for Canada in the Supplemental Appendix S.3. In a nutshell, as for the U.S. or France, there is substantial functional specialization in Canada.

Using the census divisions c from the special tabulations, we can associate each plant with a location-specific share θ_{ic}^{f} of function f in industry i. This share can be used to split employment at the plant level. To begin with, assume that all plants in census division c and industry i are identical in terms of their functional composition. This is the case when $\phi^f = \phi^0$ for all

[12] The shares come from special tabulations of Statistics Canada based on the 1996 and 2001 population censuses.

[13] Duranton and Puga (2005, p.344), e.g., report that U.S. cities with a population of more than five million had a 10.2% larger management-to-production workers ratio in 1950 than the national average. That same figure rose to 39% in 1990. Duranton and Puga (2001), Hendricks (2011), and Davis and Dingel (2014) provide product life-cycle and skill-based explanations for functional specialization of large cities, respectively.

functions in our model in Section 2. From (2), we then have

$$\theta_{ic}^{f} = \frac{\left(\alpha_{ic}^{f}/w_{c}^{f}\right)^{\sigma_i}}{\sum_{f'}\left(\alpha_{ic}^{f'}/w_{c}^{f'}\right)^{\sigma_i}}, \tag{5}$$

which is the (overall) share of function f in industry i and census division c. As can be seen from (5), these shares depend on the $\alpha_{ic}^{f} = \alpha_{i}^{f}(N_c, \mathbf{L}_c)$ terms and thus do reflect coagglomeration forces. In that case, the breakdown of employment into functional types f for plant $p(i,c)$ is simply given by:

$$\ell_{p(i,c)}^{f} = \theta_{ic}^{f} \times \ell_{p(i,c)}^{0}, \tag{6}$$

where superscript '0' stands for total employment.

Two issues require clarification. First, the shares θ_{ic}^{f} that we observe are residence-based, and not workplace-based. Hence, we have to proxy workplace-based shares using residence-based ones. This implies that it is not a priori clear what the relevant geographic unit for c should be. Since $\theta_{ic}^{f} = f(\alpha_{ic}^{f}/w_{c}^{f})$ is a function of local wages and employment composition, and since wages and local employment composition are largely determined by supply-side considerations in equilibrium, we have to find the relevant area c for local labor supply. Labor supply elasticities vary substantially across space and depend on commuting patterns (see Monte, Redding, and Rossi-Hansberg, 2015). Hence, we adjust for commuting patterns by smoothing the functional shares θ_{ic}^{f} within a 25 kilometers radius around each plant.[14] We provide technical details on the procedure in Appendix B.1. The bottom line is that smoothing the shares does not significantly alter our results.

Second, when the ϕ^{f} differ across functions, expressions (5) and (6) no longer hold. Instead, the functional split now depends on the plant characteristic z:

$$\ell_{p(i,c)}^{f}(z) = \theta_{ic}^{f}(z) \times \ell_{p(i,c)}^{0}. \tag{7}$$

In our application, z is the employment size of the plant, which is positively linked to productivity. Firms with larger size z should have more non-production workers and a different hierarchical structure (see Caliendo *et al.*, 2015). We show in Appendix B.2 how we can adjust the shares to make them vary by plant size, in order to retrieve $\theta_{ic}^{f}(z)$ in (2) and (6).

In the end, we construct three different types of shares to split plant-level employment by functions: (i) unadjusted shares from the special census tabulations; (ii) spatially smoothed shares within a 25 kilometers radius; and (iii) spatially smoothed and size-adjusted shares. We present results for the baseline shares to save space since they differ little from those with spatial smoothing or size adjustments (see Appendix B.1 for a summary of the correlations

[14] The 25 kilometers threshold is chosen since 90% of Canadians commute less than 25 kilometers to work (Statistics Canada, 2006 Census. Catalogue Number 97-561-XCB2006010).

between the different types of shares). Table 9 in Appendix A.1 summarizes the descriptive statistics for our shares at the plant level.[15]

4.1.2 Marshallian covariates

The second key ingredient to estimate (4) are the measures of the Marshallian agglomeration forces. Following previous work on coagglomeration patterns (see Duranton and Overman, 2005, 2008; Ellison *et al.*, 2010; Faggio *et al.*, 2017) we construct proxies for the three Marshallian determinants of agglomeration: (i) input-output links (customer-supplier links); (ii) common labor pools; and (iii) knowledge sharing.

Starting with input-output links, we use detailed input-output matrices for the years 1998, 2000, and 2002, which we associate with our industry coagglomeration measures in 2001, 2003, and 2005. Next, we compute two measures for common labor pools. The first measure follows Glaeser and Kerr (2009) and Ellison *et al.* (2010) and builds on the 'occupational employment similarity' (henceforth OES) of the workforce between two industries. To this end, we compute the correlation between the vectors of occupational shares of industries i and j. We also decompose this by function using a mapping from the OES functions to our broad functional categories. As a second measure of common labor pools, we compute an index of labor mobility across manufacturing industries. Both of our measures are constructed using U.S. data, which alleviates concerns of endogeneity since labor markets are local and there is very little cross-border labor mobility.[16] Last, we construct proxies for knowledge sharing by using the NBER Patent Citation database and by following previous work by Kerr (2008). We construct two proxies for knowledge flows: (i) a 'make-based' measure, which is the maximum of the shares of patents that industries i or j manufacture and which originate from the other industry; and (ii) a 'use-based' measure, which is the maximum of the shares of patents that industries i or j use and which originate from the other industry. We will use the use-based measure as our benchmark, but the results are very similar for the make-based measure. Table 10 in Appendix A.3 summarizes the descriptive statistics for our Marshallian covariates.

4.2 Identification strategy

As explained in Section 2, there are several confounding factors that may lead to the spurious colocation of industries. First, industries may colocate for reasons unrelated to the existence

[15] We have data for 3 years for a total of 10,965 distinct measures of coagglomeration ($3 \times 3,655$). We compute the coagglomeration measure for total employment and for each of our four functional employment types separately. Each measure by function is further computed based on three different types of shares (baseline, spatially smoothed, and size adjusted), for a total of 142,545 different industry-function-year measures.

[16] Comparable data for Canada is not readily available, which is another reason for using the U.S. data. We conjecture that the same industries in Canada and in the U.S. use a fairly similar labor mix as they are defined in the same way and have access to the same technology.

of agglomeration economies that operate between them because they seek proximity to similar locational advantages such as natural resources or infrastructure. Second, some of the Marshallian covariates may be endogenous to observed coagglomeration patterns. Third, location patterns are inherently a general equilibrium outcome among many industries, and this needs to be addressed when looking at colocation patterns of individual industry pairs. Last, industries may differ technologically in the extent that they can split activities across locations. We now explain succintly how we address these identification challenges. More details are provided in Appendix A.3 and in the Supplemental Appendix S.4.

4.2.1 Locational advantage

As highlighted by Ellison and Glaeser (1999), industries may colocated not because of coagglomeration economies but because they draw on the same locational advantages, such as access to specific infrastructure and differences in factor costs. To control for that, we construct coagglomeration measures based on counterfactual benchmark distributions that are determined using a large range of covariates that may influence the location patterns of industries. We provide details on that procedure in the Supplemental Appendix S.4. The counterfactual coagglomeration measures are then included as a control into our regressions.[17]

One specific form of locational advantage is what Ellison and Glaeser (1999) refer to as 'natural advantage', driven mainly by the unequal spatial distribution of natural resources. Since we do not have detailed geographic data on access to resources, we include as additional controls in our regressions the share of inputs that industries buy from primary industries, and the share of output that the industries sell to primary industries (NAICS 11–22). Furthermore, since our data include only manufacturing industries, we also use as an additional control the share of inputs that industries buy from business service industries, and the share of output that they sell to business service industries (NAICS 52–55). Business services display substantial geographic concentrated, so that industries that source a lot of those services or that sell a lot to those service industries may be coagglomerated because of these buyer-supplier links.

4.2.2 Endogeneity concerns

Another concern is that some of our covariates may be potentially endogenous (see Ellison *et al.*, 2010, for a careful discussion). First, input-output links may reflect geographic coagglomeration patterns instead of driving them. We partly mitigate that issue in our baseline regressions by using three-year lagged input-output tables. Using lagged values of these tables addresses,

[17]We also estimated specifications where we excluded all industry pairs that are significantly coagglomerated or codispersed up to 25 kilometers based on the counterfactual location patterns (i.e., we keep only industry pairs that are randomly distributed up to 25 kilometers based on the counterfactual distributions). The results are fairly comparable to the ones we report below and available upon request.

at least partly, issues of simultaneity between the geographic structure of industries and their input-output links. Of course, since both geographic patterns and input-output relationships change slowly over time, serial correlation may still create an endogeneity problem. Hence, we also instrument the Canadian input-output coefficients with their U.S. counterparts, using the detailed input-output benchmark tables for 1997 and 2002 from the Bureau of Economic Analysis (BEA). Although cross-border trade between Canada and the U.S. — think, e.g., of the automotive industry — may make such instruments less exogenous than in the case of the UK used by Ellison *et al.* (2010), these instruments perform well. Our results are also robust to the exclusion of industries for which cross-border trade is a significant share of total industry output, thus mitigating residual endogeneity concerns.

Second, turning to our two labor market pooling variables (occupational employment correlations and intersectoral labor mobility), we directly use U.S. data for our analysis. Since labor markets are essentially local, and because there is little cross-border labor mobility between Canada and the U.S., we think that contemporaneous U.S. data provides a relatively exogenous source of variation in workforce similarity to investigate Canadian coagglomeration patterns.[18] Our approach is similar in spirit to those in Ellison *et al.* (2010), who use UK labor market areas as instruments for the their U.S. labor pooling variable, and Gabe and Abel (2016), who use Mexican worker data as instruments for similar knowledge in occupations.

Third, there may be endogeneity concerns with the proxy for our shared knowledge variable. Ellison *et al.* (2010) discuss this issue carefully. Since there are no good instruments, they cannot instrument this variable and hence drop it from their instrumental variables regressions. We do not attempt to instrument this variable, as we again use U.S. data to create our shared knowledge variable (see Kerr, 2008, for details).

4.2.3 General equilibrium effects

While the need to control for locational advantage is well understood, the problem that the spatial distribution of economic activity is essentially a general equilibrium outcome has received much less attention in the empirical literature. To understand the nature of the problem, consider three industries. Assume that industries i and j share inputs and are coagglomerated at 25 kilometers distance, with 20% of bilateral distances for plant-pairs in those industries below that threshold. Assume further that industries j and k share knowledge, and that 15% of bilateral distances of plant-pairs in those industries are below 25 kilometers. Then, mechanically, some share of plant-pairs (below 15%) in industries i and k must also be coagglomerated at less than 25 kilometers. This happens not necessarily because those plants interact, but because

[18] As in Ellison *et al.* (2010), the identifying assumption is that the underlying patterns of locational advantage that may drive the colocation of industry pairs by chance — which in turn may lead to the use of a similar workforce or similar inputs — differ between the U.S. and Canada.

they interact with a *third industry* j, which may induce spurious coagglomeration patterns and is potentially a serious threat to identification.[19]

To the best of our knowledge, this general equilibrium problem has not yet been seriously taken into account in the literature. Without a structural model in which multiple industries colocate in equilibrium, it is hard to address. We hence take two reduced-form approaches to provide a first cut at the problem. First, we compute the correlation coefficient between the colocation patterns of two industries i and j with all the *other industries* k. This correlation captures how similar two industries are in terms of their overall location patterns. We then include this variable into our regressions, or we use it to exclude the industry pairs ij that have the overall most similar location patterns. Second, we may view the coagglomeration measures between industry pairs ij as the matrix of an undirected graph, where the coagglomeration measures — which are between 0 and 1 by construction — are weights of the edges. Following Ballester, Calvó-Armengol, and Zenou (2006), we can then compute the *Bonacich centrality* of each industry in that network to measure which industries are the most influentially connected with the other industries. A high value of that measure means that the industry is 'central' in the coagglomeration network, i.e., strongly coagglomerated with industries that are themselves strongly coagglomerated with many other industries. Those industries are thus more likely to be spuriously coagglomerated. Again, we can either include the Bonacich centrality as a control into our regressions, or we can exclude the industry pairs where one of the two industries has a high centrality measure. A more detailed discussion is given in Appendix A.3.

4.2.4 Industrial organization

We finally include controls for industries' organizational structure into our regressions. First, to control for the fact that industries with many multiunit firms can more easily split functions across establishments, we control for the share of multiunit plants in each sector. Second, we control for the fact that agglomeration can also be *internal to plants and firms themselves* (e.g., Alcácer and Delgado, 2017). Consider, for example, a car manufacturer. That car manufacturer may also produce some of the parts that it needs at the same location where it manufactures the cars. If many car manufacturers do that, the proximity of the two activities may not show up in the data just because the firm is labeled 'car manufacturer' based on its main line of activity. If 'car production' and 'parts production' were carried out by two distinct entities at the same location, this would show up in our data. Put differently, if we split the integrated producer into two, this would affect our coagglomeration measures. If the multiple activities that firms internalize at the same location were random, this would not be a problem for our

[19]These third-industry effects are akin to the 'multilateral resistance problem' in the gravity equation literature (see Anderson and van Wincoop, 2003). In that literature, it is shown that trade flows cannot be simply analyzed on an ij basis, since they also depend on trade flows of i and j with third countries k. The coagglomeration pattern between i and j similarly depends on the coagglomeration patterns of i and j with third industries k.

analysis. However, firms internalize activities based on cost savings, and part of those costs savings are due to the geographic proximity of related activities. Hence, it seems important to control for internal agglomerations, as their existence affects our measures of coagglomeration and are correlated with the Marshallian agglomeration forces.[20]

We exploit an interesting feature of our dataset to control for the fact that plants can operate in multiple industries. More precisely, we use our plants' additional (non-primary) NAICS codes to compute the share of plants with primary NAICS code i that report also secondary codes in industry j. This provides us with a 'within-plant coagglomeration metric' of industries. Industry pairs with a high value within_{ij} are industries that are 'coagglomerated more strongly within plants'. As an example, note that 75% of multi-industry plants in 'Apparel knitting mills' (NAICS 3151) also report operating in 'Cut and sew apparel manufacturing' (NAICS 3152). This suggests that technical (or process) complementarities lead that activity to be strongly agglomerated geographically — the strongest possible form of geographic concentration being that the activities are carried out within the same plant.

4.2.5 Others

Last, we also include a full set of industry fixed effects into our regressions (see the Supplemental Appendix S.2 for technical details).[21] The latter will partly control for systematic differences in the degree to which industries colocate with other industries. For example, some industries and functions are more urban so that one industry may seem to systematically coagglomerate more with other urban industries. Industries differ also vastly in their number of plants, their exposure to trade, their transport costs, and the technologies they use. Our fixed effects will at least partly take care of those differences, especially since they are unlikely to vary much over a short time period such as five years that we use in our analysis.[22]

[20]Table 11 in Appendix A.3 shows that the existence of 'within-plant coagglomeration' may create problems since that variable is strongly correlated with the traditional Marshallian factors. Indeed, the simple correlations are 0.45 with our measure of input-output links, 0.39 with our measure of common labor pools, and 0.13 with our measure of knowledge sharing. As expected, plants that carry out several activities in-house bundle those activities where the benefits from geographic proximity are large.

[21]Neither Ellison *et al.* (2010) nor Faggio *et al.* (2017) report fixed effects estimations for their benchmark regressions. In case of the former, the reason they state is that fixed effects suspiciously affect the labor market pooling variable. As we show later when controlling for general equilibrium effects, that variable is indeed suspicious and it is theoretically questionable if it is appropriate to use it in the regressions.

[22]Even if we have panel data, we cannot really use the within variation by industry pair. The reason is that spatial patterns and inter-industry relationships (especially input-output relationships) change only slowly, so that we would need a longer panel to pick up the effects. Furthermore, there is no time variation in our proxy for knowledge sharing. Alternatively, we could construct industry fixed effects as explained in Supplemental Appendix S.2 and interact them with the year dummies to create industry-year fixed effects. Given that we have a short study period (2001-2005) during which there were no major economic shocks, this should not significantly change our key results.

5 Empirical results

We first provide detailed descriptive results on coagglomeration patterns for total employment and by functional employment types. We also examine how these patterns change with distance and over time. We then estimate several variants of (4), using different coagglomeration measures as discussed in Section 4.1.1. We again first present results for total employment and then by functional employment types.

5.1 Coagglomeration patterns

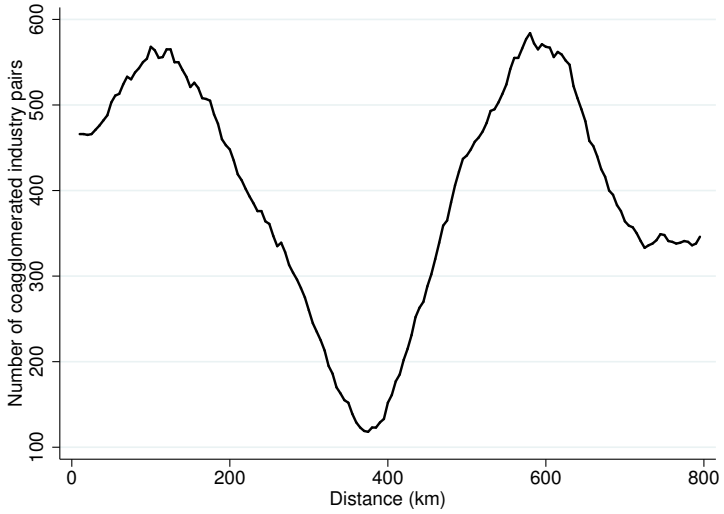
5.1.1 Total employment

The top panel in Table 1 summarizes the shares of coagglomerated, random, and codispersed industry pairs for total employment in 2001, 2003, and 2005. Clearly, a substantial share of industry pairs is coagglomerated, even if the extent of coagglomeration gets weaker between 2001 and 2005 (about 10% fewer coagglomerated industry pairs). Conversely, random colocation patterns become more prevalent (about 35% more random industry pairs). Codispersion patterns remain fairly stable. These trends echo findings by Behrens and Bougna (2015), who have documented that the number of significantly localized manufacturing industries has decreased in Canada between 2001 and 2009.

Figure 3, which depicts the number of significantly coagglomerated industry pairs by distance, shows that the extent and pattern of coagglomeration vary substantially with distance. The number of coaggomerated manufacturing industry pairs increases until about 100-150 kilometers, and then decreases regularly until about 400 kilometers. It then increases again until a second lower peak at about 550-650 kilometers is reached. The latter corresponds roughly to the distance between the two largest Canadian metropolitan areas, Toronto and Montréal.

The peak at about 100-150 kilometers reveals that industries tend to coagglomerate more at intermediate distances, and less at very short ones. Since own-concentration of industries more frequently occurs at short distance (see Behrens and Bougna, 2015), this finding suggests that localization economies — agglomeration economies *within industries* — may operate over shorter distances than coagglomeration economies — agglomeration economies *between industry pairs*. If the geographic concentration of own industry increases land prices and competition in local labor markets, own-industry concentration that benefits more from close proximity of plants will tend to dominate cross-industry concentration that benefits less from (very) close proximity of plants. Observe that the patterns in the data also suggest that input-output relationships, which are less sensitive to distance than common labor pools or knowledge sharing, may be a primary driver of coagglomeration patterns. This observation is consistent with the peak in Canadian shipping patterns at about 100 kilometers, as documented by Behrens and

Figure 3: Significantly coagglomerated industry pairs by distance (2003, total employment).



Brown (2017) using Canadian trucking microdata.[23] In a nutshell, industry pairs that exchange a lot of inputs do not necessarily need to locate in very close proximity, but should not be too far away either. Our results suggest that intermediate distances of about 100-150 kilometers are most frequently observed in the data.

5.1.2 Functional employment types

The bottom four panels of Table 1 summarize the shares of coagglomerated, random, and codispersed industry pairs by broad functional employment types. They show that there is significant heterogeneity in the coagglomeration patterns of different employment types, with clerical and production employment being on average more coagglomerated (about 54–60% of industry pairs) than retail and services (about 30% of industry pairs) and management and research (about 47-55% of industry pairs). As can be further seen from that table, the general trend is toward less coagglomeration, with a slight increase in the number of random and stability of codispersed industry pairs. This finding echoes that for total employment.

Figure 4 depicts the number of significantly coagglomerated industry pairs by distance for the different functional employment types in 2003. It shows that the geographic structure of the coagglomeration patterns for these four functions is markedly different. Whereas management and research displays substantial coagglomeration at short geographic distances (0–150 kilometers), production displays substantial coagglomeration at longer distances (150–250 kilometers) but less coagglomeration at shorter distances. As already argued for the case of total

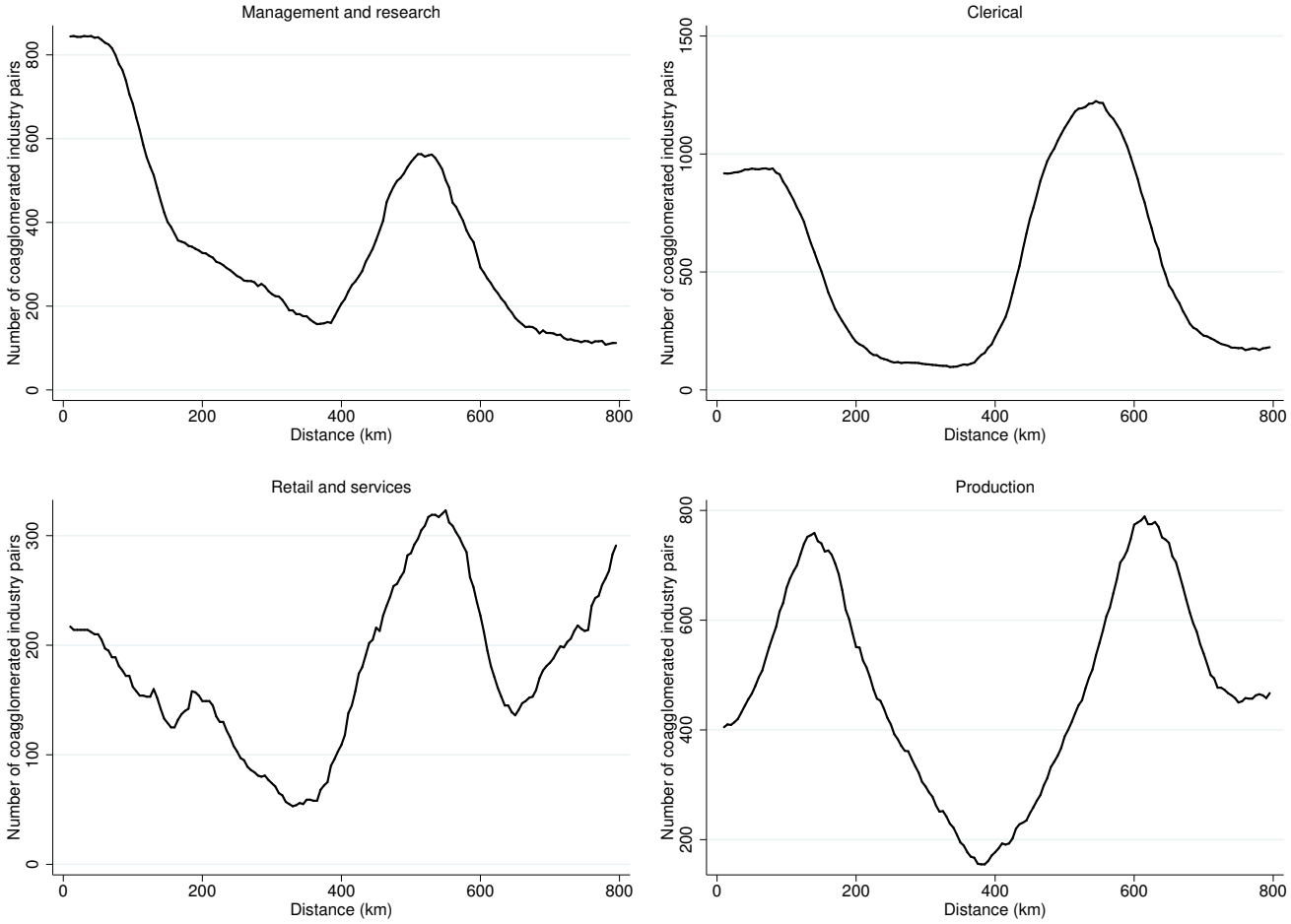
[23]The average value per tonne of goods shipped also starts to rapidly increase after about 100 kilometers distance, while it is fairly stable for distances below 100 kilometers (Behrens and Brown, 2017). This suggests that the mix of goods shipped switches towards higher value goods after that distance, since shipping lower value goods becomes too expensive.

Table 1: Coagglomeration patterns of manufacturing industries in Canada.

		Baseline functional shares			**Spatially smoothed shares**			**Size-adjusted shares**			**Locational advantage patterns**		
Year	Employment type	% Coag.	% Rand.	% Codi.	% Coag.	% Rand.	% Codi.	% Coag.	% Rand.	% Codi.	% Coag.	% Rand.	% Codi.
2001	Total employment	56.96	31.52	11.52							51.38	2.22	46.40
2003		52.67	35.84	11.49							51.63	2.35	46.02
2005		47.82	40.14	12.04							50.70	2.11	47.20
2001	Management and research	53.98	30.92	15.10	57.24	29.38	13.38	60.36	26.95	12.69			
2003		48.04	37.35	14.61	50.75	34.72	14.53	57.10	29.44	13.46			
2005		46.02	37.24	16.74	48.78	36.25	14.97	53.19	32.12	14.69			
2001	Clerical	61.83	27.25	10.92	63.75	26.18	10.07	65.09	24.08	10.83			
2003		57.62	29.49	12.89	57.54	28.76	13.71	57.98	30.04	11.98			
2005		55.76	31.00	13.24	56.09	30.67	13.24	53.63	34.34	12.04			
2001	Retail and services	28.24	47.80	23.97	30.48	49.17	20.36	40.49	46.59	12.91			
2003		31.68	52.80	15.51	35.51	49.71	14.77	41.61	46.68	11.71			
2005		30.42	53.87	15.70	32.39	51.57	16.03	36.44	50.70	12.86			
2001	Production	60.19	27.93	11.87	60.36	27.41	12.23	63.17	23.94	12.89			
2003		59.10	29.60	11.30	58.58	30.04	11.38	61.67	26.48	11.85			
2005		54.28	34.06	11.66	53.38	34.45	12.18	56.47	30.67	12.86			

Notes: Results for all 3,655 unique industry pairs obtained from 86 NAICS 4-digit industries. All computations for 0-800 kilometers, 5 kilometers steps and 200 bootstrap replications for the (global) confidence bands. See Appendix B for details on how we construct the different shares and compute the coagglomeration measures; and Supplemental Appendix S.4 for details on how we construct our counterfactual locational advantage distributions. Coag. = significantly coagglomerated; Rand. = not significantly different from random pattern; Codi. = significantly codispersed.

Figure 4: Significantly coagglomerated industry pairs by distance (2003, by employment type).



employment, one explanation for these different location patterns could be that management and research is more sensitive to shared knowledge — which operates at small geographic scales — whereas production may be more sensitive to input-output links — which operate at larger geographic scales. This is exactly what we find in our subsequent regression analysis. Figure 4 further shows that clerical employment displays the strongest coagglomeration patterns at short distances, and at between-city distances of around 550-600 kilometers. Last, retail and services has across the board the smallest number of coagglomerated industry pairs, and the spatial pattern is closest to that of clerical employment (though still markedly different). To summarize, the different functional employment types display spatially distinct coagglomeration patterns. It is these differences that are useful to better understand the determinants of coagglomeration.

5.2 Determinants of coagglomeration

We now provide detailed multivariate results on the determinants of coagglomeration for total employment and by functional employment types. We also analyze how the strength of these

determinants varies with the spatial scope of the coagglomeration patterns. We estimate several variants of (4), using two different coagglomeration measures. In all of our regressions, we pool observations across years to increase the precision of our estimates, but the cross-sectional results are fairly similar. We standardize all variables so that the magnitude of the coefficients are directly comparable. We also exclude industry pairs ij where either industry i or industry j have less than 30 plants. The rationale for excluding those pairs is that the coagglomeration measures are noisier because of small sample sizes. Our results are, however, not sensitive to those exclusions.

5.2.1 Total employment

Table 2 shows our baseline results for total employment. The dependent variable is the CDF of the DO K-density — an absolute measure of coagglomeration, as given by (B.2) — in specifications (1)–(8); and the strength of coagglomeration — a relative measure of coagglomeration, as given by (B.5) — in specifications (9)–(16). Specifications (1) and (9) do not include industry fixed effects, while all other specifications do include such fixed effects. In specifications (3) and (11), we exclude all industry pairs within the same 3-digit industries.[24] Specifications (4)–(5) and (12)–(13) provide robustness checks for separate input or output links, while specifications (6) and (14) use an alternative measure for knowledge spillovers ('make based' instead of 'use based'). Specifications (7) and (15) use an alternative measure for labor market pooling, namely cross-industry labor flows. Last, specifications (8) and (16) instrument the input-output links using our U.S. instruments to correct potentiel endogeneity biases.

Starting with the DO CDF measure of coagglomeration, specifications (1)–(8) in Table 2 show that all coefficients for the input-output links and the labor similarity measure are positive and significant: input-output links and common labor pools are positively associated with the coagglomeration of industries. Observe that labor similarity appears to be the most important driver of the absolute degree of coagglomeration as compared to input-output links or knowledge sharing. Yet, as can be seen from specification (3), the coefficient for input-output links tends to increase when excluding same-industry pairs, whereas the coefficient on labor market pooling remains relatively stable. Including industry pairs within the same 3-digit industries hence tends to understate the importance of input-output links compared to the labor market variable. The knowledge sharing variable is not significantly associated with coagglomeration in specification (1)–(3), (6) and (8). In specifications (3)–(5) and (7) it is borderline significant and has a fairly small effect size compared to the other Marshallian covariates. Last, specification (8) shows that the IV results are fairly similar to the OLS results, with the input-output

[24]The rationale for doing so is that 4-digit industries within the same 3-digit industries tend to be similar along a number of dimensions that may make less clean the identification of the effects of the Marshallian factors (Ellison *et al.*, 2010). Furthermore, some of our variables are constructed at an aggregation level slightly above 4-digit (e.g., the input-output measures, which are at the link level), thus reducing the variability within 4-digit industry pairs.

Table 2: Coagglomeration patterns of manufacturing industries in Canada, total employment.

Dependent variable	CDF of coagglomeration								Strength of coagglomeration							
	(1)	(2)	(3) (Excl3)	(4)	(5)	(6)	(7)	(8) (IV)	(9)	(10)	(11) (Excl3)	(12)	(13)	(14)	(15)	(16) (IV)
Input-output links	0.086^a	0.037^a	0.058^a			0.037^a	0.043^a	0.059^a	0.040^b	0.041^a	0.037^a			0.041^a	0.050^a	0.016
	(0.013)	(0.007)	(0.009)			(0.007)	(0.007)	(0.016)	(0.016)	(0.015)	(0.010)			(0.015)	(0.016)	(0.023)
Labor similarity	0.136^a	0.066^a	0.074^a	0.077^a	0.064^a	0.066^a		0.055^a	0.075^a	0.020	0.037^a	0.027^b	0.019	0.020		0.033^b
	(0.011)	(0.007)	(0.008)	(0.007)	(0.007)	(0.007)		(0.010)	(0.011)	(0.012)	(0.014)	(0.012)	(0.012)	(0.012)		(0.015)
Patent citations	-0.007	0.009	0.010^c	0.010^c	0.009^c		0.010^c	0.007	-0.026^b	-0.004	0.004	-0.003	-0.003		-0.002	-0.002
	(0.011)	(0.005)	(0.006)	(0.005)	(0.005)		(0.005)	(0.005)	(0.010)	(0.008)	(0.008)	(0.008)	(0.007)		(0.007)	(0.008)
Input links				0.016^b								0.028^c				
				(0.007)								(0.015)				
Output links					0.046^a								0.049^a			
					(0.007)								(0.016)			
Patent citations (make based)						0.008								-0.005		
						(0.006)								(0.007)		
Labor movement							0.036^a								-0.006	
							(0.006)								(0.010)	
Year fixed effets	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	10,292	10,292	9,729	10,292	10,292	10,292	10,292	10,292	10,292	10,292	9,729	10,292	10,292	10,292	10,292	10,292
R-squared	0.033	0.846	0.847	0.846	0.847	0.846	0.845	0.846	0.012	0.392	0.389	0.391	0.393	0.392	0.392	0.392
First-stage IV (8) + (16):																
Input-output links, U.S. IV								0.535^a								0.535^a
								(0.045)								(0.045)
Labor similarity								0.306^a								0.306^a
								(0.021)								(0.021)
Patent citations								0.05^b								0.05^b
								(0.011)								(0.011)
First-stage R^2								0.520								0.520
First-stage F statistic								144.35								144.35
Kleibergen-Paap LM								57.964								57.964
Endogeneity p-value								0.104								0.049

Notes: Results for all $3 \times 3,655$ unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All computations for 0-800 kilometers, with 1 kilometer steps for the K-density CDF, and 5 kilometers steps and 200 bootstrap replications for the (global) confidence bands for the strength of agglomeration. The dependent variables are computed at 25 kilometers distance in both panels. In specifications (3) and (11), we exclude all industry pairs that are in the same 3-digit NAICS industry. In specifications (8) and (16), we report 2SLS results instrumenting the input-output links with their U.S. counterparts. Huber-White robust standard errors in parentheses.

links coefficient being stable and the labor similarity coefficient decreasing slightly.

Turning next to the strength of coagglomeration measure, specifications (9)–(16) in Table 2 show that the results are slightly different. The most important driver of the relative coagglomeration of industries are input-output links, which are positive and significant in almost all specifications. The results for labor similarity are less robust. In particular, the magnitude of that variable is much smaller and it is not significantly associated with coagglomeration in specifications (10), (13) and (14). The labor movement variable in specification (15) is also insignificant. Our proxy for knowledge sharing is insignificant in almost all specifications, with point estimates that are very close to zero. Last, specification (16) shows that the input-output links coefficient becomes insignificant in the IV estimation, and that the labor similarity coefficient also decreases and is less precisely estimated.

We present estimations including the full set of controls in Table 19 in Supplemental Appendix S.5.[25] Our key findings from Table 2 are robust to the inclusion of controls for: locational advantage; the shares of inputs and outputs that industries source from primary and from business-service industries; the share of multiunit plants in the industries; and the within-plant agglomeration measures (see Section 4.2.4 for details). Most of those controls are highly significant and have the same sign using both measures — absolute (CDF) and relative (strength) — of coagglomeration as the dependent variable.

One advantage of using continuous measures of coagglomeration is that we can investigate the distance profiles of the Marshallian agglomeration forces. Figure 5 — which summarizes estimates of specifications (2) and (10) in Table 2 when measuring coagglomeration using different distance thresholds — reveals that the effect of input-output links increases with distance, whereas the effect of knowledge spillovers sharply falls with distance, especially below 100 kilometers. The labor similarity variable globally falls with distance when coagglomeration is measured using the CDF. However, it first decreases (though it is insignificant; see (10) in Table 2) and then increases strongly (and becomes significant) when we use the 'strength of coagglomeration' metric. This variable has the largest differences in its distance-based coagglomeration profile, depending crucially on whether we measure coagglomeration in absolute or in relative terms. One possible explanation is related to how that variable is constructed: it captures the similarity of two industries in terms of the workers they employ. Similar industries will hence tend to colocate in absolute terms to take advantage of local labor pools, but conditional on that there is little reason for those industries to be *relatively closer together* because of the labor market channel. It is hence not clear how useful the local labor market variables are for understanding coagglomeration patterns because they are not related to pairwise interactions between industries.

[25]Table 18 in the Supplemental Appendix S.5 summarizes our baseline regressions, where we explain the coagglomeration measures solely as a function of industry fixed effects, locational advantage, industrial organization variables, and the third-industry effects.

Figure 5: Estimated coefficients by distance, using specifications (2) and (10) from Table 2.

(a) CDF of coagglomeration.

(b) Strength of coagglomeration.

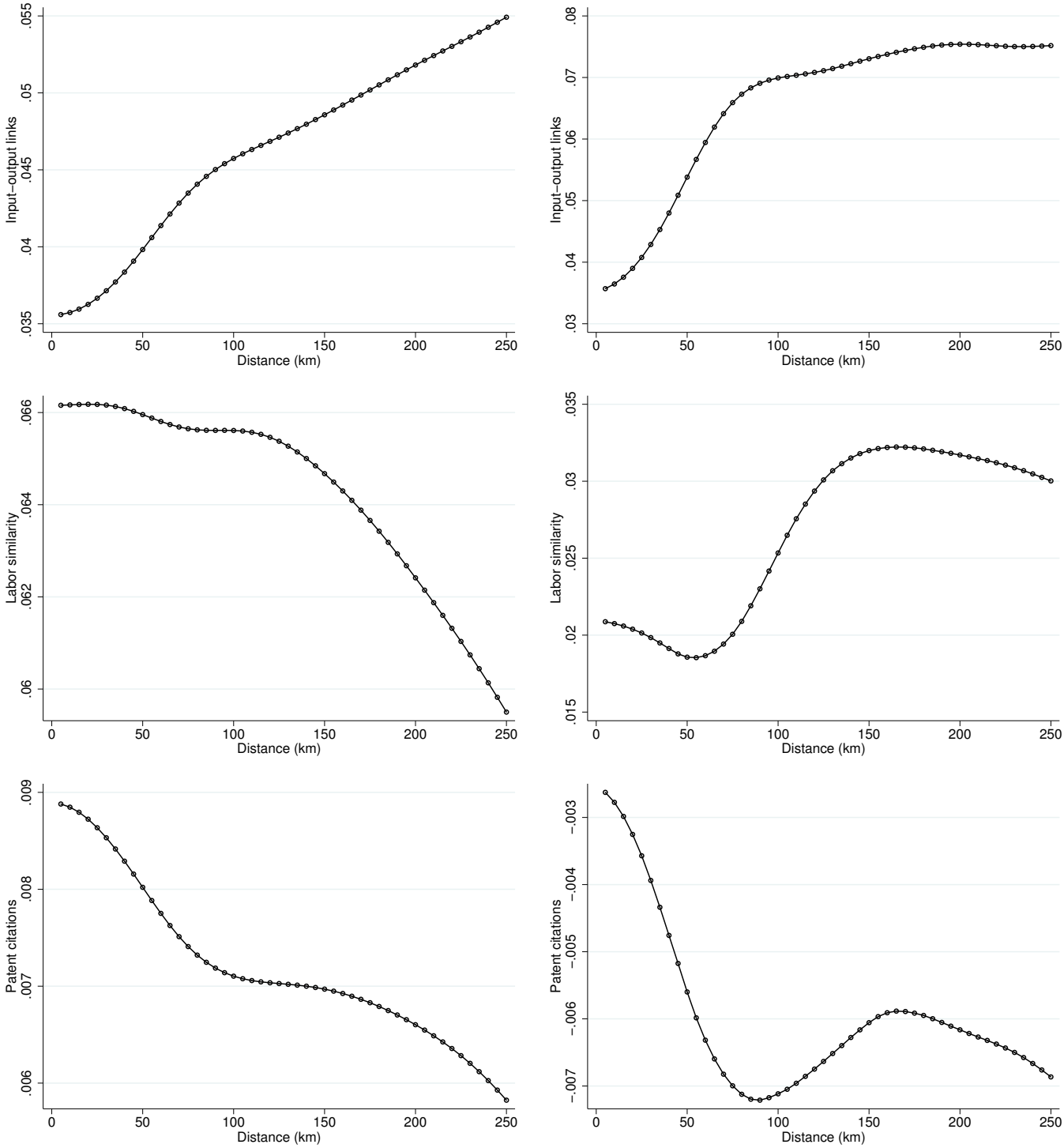


Table 3: Coagglomeration patterns, controlling for locational correlations and Bonacich centrality.

	Controlling for locational correlations							**Controlling for Bonacich centrality**						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Input-output links	0.007	0.018^{b}	0.020^{a}	0.020^{b}	0.017^{c}	0.015	0.013	0.029^{a}	0.026^{a}	0.015^{b}	0.019^{a}	0.030^{a}	0.038^{a}	0.035^{a}
	(0.004)	(0.007)	(0.008)	(0.008)	(0.009)	(0.010)	(0.010)	(0.005)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Labor similarity	0.009	0.026^{a}	0.021^{b}	0.010	0.005	-0.003		0.033^{a}	0.030^{a}	0.031^{a}	0.025^{a}	0.013^{b}	0.012	
	(0.006)	(0.008)	(0.008)	(0.009)	(0.010)	(0.011)		(0.006)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	
Patent citations	0.005	0.004	0.004	0.005	0.009	0.012^{b}	0.012^{b}	0.001	0.002	0.008^{c}	0.008^{c}	0.008	0.010^{b}	0.010^{b}
	(0.004)	(0.004)	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.003)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
Locational correlations	0.516^{a}													
	(0.017)													
Bonacich centrality								-0.394^{a}						
								(0.008)						
Labor movement							0.005							0.016^{b}
							(0.008)							(0.007)
Included deciles	All	1-9	1-8	1-7	1-6	1-5	1-5	All	1-9	1-8	1-7	1-6	1-5	1-5
Controls included	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	10,292	9,263	8,234	7,205	6,176	5,146	5,146	10,292	9,324	8,257	7,261	6,200	5,202	5,202
R-squared	0.895	0.849	0.845	0.848	0.851	0.852	0.852	0.892	0.862	0.869	0.869	0.865	0.859	0.859

Notes: Results for all $3 \times 3,655$ unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All computations for 0-800 kilometers, with 1 kilometer steps. The dependent variable is the CDF of coagglomeration, computed at 25 kilometers distance. The included controls are: Locational advantage (CDF, 25km); input share (primary); output share (primary); input share (business services); output share (business services); multiplant share; and within-plant coagglomeration. In the different specifications, we include only certain deciles of the distribution of the dependent variable, where the inclusion is based on the distribution of locational correlations (left panel) or Bonacich centralities (right panel). For example, decile 1-8 in the left panel means that we exclude the 20% of industry pairs with the highest locational correlations. Huber-White robust standard errors in parentheses.

Table 3 shows results for the CDF measures of coagglomeration where we control for general equilibrium location patterns and third-industry effects in a reduced-form way. Specification (1) in the left panel includes directly our locational correlations as a control. As can be seen, no coefficient on the Marshallian covariates remains significant in that case. Specifications (2)–(7) show results where we progressively exclude the top deciles of industry pairs that have the strongest locational correlations with all other industries, while not including the correlations directly as a control. As one can see, progressively excluding the industry pairs that have the most similar overall location patterns reduces the coefficients on input-output links and on labor similarity, although the former remains more stable and significant, whereas the latter quickly decreases to zero and becomes insignificant. Specifications (8)–(14) use the Bonacich centrality measure of industries, instead of the locational correlations, to control for overall location patterns (see Appendix A.3 for additional details). As can be seen, the results are similar to the ones in the left panel of the table: the input-output links remain fairly stable, whereas the coefficients on labor similarity tend to vanish. As for our results using the strength of coagglomeration as the dependent variable, we find that the labor market variables are sensitive to controls for overall location patterns. This suggests again that those variables are hard to interpret as they tend to predominantly pick up general location patterns. We further discuss these results and their interpretation in Section 6.

5.2.2 Functional employment types

We have so far only estimated average effects of the determinants of coagglomeration by using total employment. Breaking down employment by functional types, we can exploit two additional dimensions of heterogeneity in our data. First, industries and their employment types have an uneven spatial distribution (see Figure 4). Second, coagglomeration patterns differ by employment types. Exploiting that heterogeneity provides better identification of the determinants of coagglomeration.

We hence now run regressions by functional employment type, i.e., we split total employment by employment type and compute type-specific coagglomeration measures (see Appendix B for details). Table 4 shows our baseline results, where we include only the Marshallian covariates and industry and year fixed effects. We present results using the raw functional shares in the top panel. Results for either spatially smoothed or size-adjusted shares are presented in the middle and the bottom panels, respectively. Since the results are similar irrespective of how we compute the shares, we only discuss results using the baseline shares in what follows. To save space, the regressions that include the full set of controls are relegated to Table 20 in the Supplemental Appendix S.5.

Note first that the coagglomeration patterns of all employment types in Table 4 are positively and significantly associated with the input-output links. In general, the effect is the

Table 4: Coagglomeration patterns of manufacturing industries in Canada, by employment types.

Dependent variable	CDF of coagglomeration				Strength of coagglomeration			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Management and research	Clerical	Retail and services	Production	Management and research	Clerical	Retail and services	Production
Baseline shares								
Input-output links	0.038^{a}	0.064^{a}	0.046^{a}	0.039^{a}	0.033^{a}	0.054^{a}	0.038^{a}	0.043^{a}
	(0.006)	(0.007)	(0.007)	(0.007)	(0.010)	(0.011)	(0.010)	(0.015)
Labor similarity	0.016^{c}	0.070^{a}	0.066^{a}	0.064^{a}	-0.044^{a}	-0.044^{a}	0.021^{c}	0.007
	(0.009)	(0.008)	(0.007)	(0.007)	(0.015)	(0.012)	(0.011)	(0.017)
Patent citations	0.013^{a}	0.017^{a}	0.009^{c}	0.007	0.011	0.017^{b}	-0.008	-0.015^{b}
	(0.005)	(0.005)	(0.005)	(0.005)	(0.008)	(0.008)	(0.008)	(0.007)
Observations	10,292	10,292	10,292	10,292	10,292	10,292	10,292	10,292
R-squared	0.844	0.827	0.810	0.818	0.469	0.446	0.393	0.315
Smoothed shares								
Input-output links	0.039^{a}	0.064^{a}	0.046^{a}	0.038^{a}	0.038^{a}	0.063^{a}	0.031^{a}	0.032^{a}
	(0.006)	(0.007)	(0.007)	(0.007)	(0.008)	(0.011)	(0.010)	(0.011)
Labor similarity	0.018^{b}	0.074^{a}	0.055^{a}	0.064^{a}	-0.044^{a}	-0.035^{a}	0.021^{c}	0.003
	(0.009)	(0.008)	(0.007)	(0.007)	(0.014)	(0.012)	(0.012)	(0.017)
Patent citations	0.011^{b}	0.018^{a}	0.009^{c}	0.007	0.006	0.021^{a}	-0.013	-0.011
	(0.005)	(0.005)	(0.005)	(0.005)	(0.008)	(0.007)	(0.009)	(0.007)
Observations	10,292	10,292	10,292	10,292	10,292	10,292	10,292	10,292
R-squared	0.849	0.829	0.817	0.816	0.468	0.457	0.405	0.310
Size-adjusted shares								
Input-output links	0.039^{a}	0.059^{a}	0.046^{a}	0.040^{a}	0.041^{a}	0.060^{a}	0.032^{a}	0.036^{a}
	(0.006)	(0.007)	(0.007)	(0.007)	(0.011)	(0.013)	(0.011)	(0.011)
Labor similarity	0.042^{a}	0.077^{a}	0.078^{a}	0.066^{a}	-0.056^{a}	-0.030^{b}	0.010	0.005
	(0.010)	(0.008)	(0.006)	(0.007)	(0.016)	(0.012)	(0.011)	(0.014)
Patent citations	0.009^{c}	0.014^{a}	0.009^{c}	0.006	0.005	0.015^{b}	-0.005	-0.015^{b}
	(0.005)	(0.005)	(0.005)	(0.005)	(0.008)	(0.007)	(0.008)	(0.008)
Observations	10,292	10,292	10,292	10,292	10,292	10,292	10,292	10,292
R-squared	0.840	0.834	0.843	0.813	0.446	0.437	0.420	0.311

Notes: Results for all $3 \times 3,655$ unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All computations for 0-800 kilometers, with 1 kilometer steps for the K-density CDF, and 5 kilometers steps and 200 bootstrap replications for the (global) confidence bands for the strength of agglomeration. The dependent variables are computed at 25 kilometers distance in both panels. All regressions include industry and year fixed effects, but no other controls. See Appendix B for details on how we construct the different shares. The 'Labor similarity' variable is specific to each employment type. Huber-White robust standard errors in parentheses.

smallest for management and research, for both the CDF and the strength of coagglomeration measures. When the controls are introduced into the model (see Table 20 in the Supplemental Appendix S.5), the input-output links still significantly influence coagglomeration for the CDF measure. When considering the strength of coagglomeration measure, however, the input-output links do no longer have any effect for management and research and for retail and service employment, but continue to significantly influence the coagglomeration of the clerical and production employment types. These results suggest that input-output links are a robust predictor of both absolute and relative coagglomeration, but less so for the coagglomeration of employment in management and research than for employment in production.

Turning to the labor similarity variable, it is positively and significantly associated with absolute coagglomeration for all employment types, but to a larger extent for clerical, retail and service, and production employment. This ranking continues to hold true when looking at relative coagglomeration in the right panel of Table 4. Yet, as shown, labor similarity seems to be more important for production and retail and service employment than for either of the other two functions. In a nutshell, labor similarity seems to be important for production, retail and services and, to a lesser extent for clerical, but not for management and research. All these results continue to hold when controls are introduced into the regressions.

Table 5: Different patterns of significantly positive coefficients.

	Management and research	**Production**
Input-output links	82%	95%
Labor similarity	32%	61%
Patent citations	68%	3%

Notes: This table reports the share of significantly positive coefficients (at the 10% level at least) across the total number of 38 specifications that we estimate. We report results across all specifications (without controls, with controls, pooled, and IV).

Last, and quite interestingly, knowledge sharing is positively and significantly associated with the coagglomeration of management and research and clerical employment, whereas it is not for the other two employment types. This result continues to hold when controls are introduced into the regression. It suggests that knowledge is not significantly associated with the coagglomeration of total employment in manufacturing, yet that it is significantly associated with the coagglomeration of employment in the functions that are a priori knowledge intensive (see also Howard *et al.*, 2016, who show for the case of Vietnam that the results for knowledge are sensitive to how coagglomeration is measured). Table 5 highlights the patterns of differences in the significance of coefficients across different functional employment types (management and research versus production) for the 38 different regressions that we have es-

timated in this paper. While input-output links seem to be important for the coagglomeration of both labor types, labor similarity is clearly more often significantly positively associated with the coagglomeration of production employment, whereas knowledge sharing is clearly more often significantly positively associated with the coagglomeration of management and research employment. These results suggest that estimating average effects does not allow to well identify the drivers of coagglomeration because those drivers differ across functions. Different functions rely to a different extent on the various agglomeration forces, and estimating the average across functions masks substantial heterogeneity. Coagglomeration patterns between industry pairs differ significantly across functions, with industries coagglomerating the functions that interact intensively within the pair (e.g., knowledge for management and research; and input-output links and labor similarity for production).

Table 6 — which focuses on management and research versus production — shows that our main results for the coagglomeration of industries are robust to different alternative specifications. The general patterns are the same as in our baseline regressions: input-output links and common labor pools matter for production more than for management and research; whereas knowledge sharing matters more for the latter. Table 21 in Supplemental Appendix S.5 replicates the results of Table 6 with all controls included.

Finally, Table 7 provides estimates where we pool the coagglomeration measures of all employment types and include interaction terms. The excluded category in those regressions is production (i.e., all interaction terms capture differences with the coagglomeration patterns of production employment). As can be seen from Table 7, the interaction term between our knowledge sharing measure and the management and research dummy is always positive and highly significant, whereas it is not for the other employment categories. Also, the input-output links have the largest effect on the coagglomeration of production employment when considering the strength of coagglomeration as the dependent variable, whereas the effect is weaker (or even negative) for the other categories. Hence, the pooled regressions across functional types yield qualitatively similar results to the separate regressions by type.

6 Discussion

Our key findings are largely in line with those in the relatively thin existing literature on the role of the Marshallian agglomeration forces for coagglomeration patterns. We now discuss a number of limitations and caveats of our analysis and provide some additional interpretations of our main findings.

First, the generally weak effect of shared knowledge as a determinant of coagglomeration could be linked to our choice of proxies for knowledge sharing. The microeconomic foundations of the role of knowledge flows in agglomeration are still thin (see Duranton and Puga, 2004). Many important issues regarding the empirical aspects of knowledge flows originate

Table 6: Coagglomeration patterns for management and research versus production (robustness).

	(1)	(2)	(3) (Excl3)	(4)	(5)	(6)	(7)	(8)	(9) (IV)
Management and research employment									
Input-output links	0.095^a	0.038^a	0.048^a			0.038^a	0.030^a	0.025^a	0.054^a
	(0.012)	(0.006)	(0.009)			(0.006)	(0.006)	(0.006)	(0.011)
Labor similarity	-0.038^a	0.016^c	0.038^a	0.025^a	0.015^c	0.017^c			0.007
	(0.009)	(0.009)	(0.011)	(0.009)	(0.009)	(0.009)			(0.010)
Patent citations	0.020	0.013^a	0.014^b	0.015^a	0.013^a		0.012^b	0.011^b	0.012^b
	(0.012)	(0.005)	(0.006)	(0.005)	(0.005)		(0.005)	(0.005)	(0.005)
Input links				0.021^a					
				(0.006)					
Output links					0.046^a				
					(0.006)				
Patent citations (make based)						0.012^b			
						(0.005)			
Labor movement							0.025^a		
							(0.005)		
Labor similarity (total employment)								0.048^a	
								(0.006)	
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	10,292	10,292	9,729	10,292	10,292	10,292	10,292	10,292	10,292
R-squared	0.011	0.844	0.845	0.843	0.845	0.844	0.845	0.845	—
Production employment									
Input-output links	0.071^a	0.039^a	0.060^a			0.039^a	0.042^a	0.035^a	0.063^a
	(0.013)	(0.007)	(0.009)			(0.007)	(0.007)	(0.007)	(0.017)
Labor similarity	0.174^a	0.064^a	0.065^a	0.073^a	0.063^a	0.064^a			0.052^a
	(0.011)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)			(0.010)
Patent citations	-0.012	0.007	0.010^c	0.008	0.007		0.008	0.007	0.005
	(0.011)	(0.005)	(0.006)	(0.005)	(0.005)		(0.005)	(0.005)	(0.006)
Input links				0.020^a					
				(0.007)					
Output links					0.046^a				
					(0.007)				
Patent citations (make based)						0.006			
						(0.006)			
Labor movement							0.043^a		
							(0.007)		
Labor similarity (total employment)								0.075^a	
								(0.008)	
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	10,292	10,292	9,729	10,292	10,292	10,292	10,292	10,292	10,292
R-squared	0.042	0.818	0.820	0.817	0.818	0.818	0.817	0.818	—

Notes: Results for all $3 \times 3,655$ unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All computations for 0-800 kilometers, with 1 kilometer steps. The dependent variables is the CDF of coagglomeration, computed at 25 kilometers distance and using the baseline shares (see Appendix B for details on the different shares). Controls are not included. The 'Labor similarity' variable is specific to each employment type. The 'Labor similarity (total employment)' variable is common to all employment type. In specification (3), we exclude all industry pairs that are in the same 3-digit NAICS industry. In specification (9), we report 2SLS results instrumenting the input-output links with their U.S. counterparts. Huber-White robust standard errors in parentheses.

Table 7: Coagglomeration patterns, pooled across all employment types.

Dependent variable	CDF of coagglomeration				Strength of coagglomeration			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Input-output links	0.027^{a}	0.008	-0.009	0.011^{c}	0.058^{a}	0.044^{a}	0.035^{b}	0.045^{a}
	(0.007)	(0.007)	(0.006)	(0.006)	(0.014)	(0.015)	(0.014)	(0.014)
Input-output links × Management and research	-0.004	-0.007	-0.008	-0.007	-0.062^{a}	-0.063^{a}	-0.064^{a}	-0.063^{a}
	(0.009)	(0.008)	(0.007)	(0.007)	(0.017)	(0.018)	(0.018)	(0.017)
Input-output links × Clerical	0.041^{a}	0.036^{a}	0.035^{a}	0.038^{a}	-0.001	-0.003	-0.003	-0.002
	(0.010)	(0.009)	(0.008)	(0.008)	(0.018)	(0.018)	(0.018)	(0.018)
Input-output links × Retail and services	0.029^{a}	0.028^{a}	0.027^{a}	0.031^{a}	-0.018	-0.019	-0.020	-0.017
	(0.011)	(0.010)	(0.009)	(0.010)	(0.017)	(0.018)	(0.018)	(0.017)
Patent citations	0.005	-0.000	0.001	-0.002	-0.020^{b}	-0.021^{a}	-0.020^{a}	-0.023^{a}
	(0.006)	(0.005)	(0.005)	(0.004)	(0.008)	(0.008)	(0.008)	(0.007)
Patent citations × Management and research	0.025^{a}	0.023^{a}	0.023^{a}	0.024^{a}	0.057^{a}	0.056^{a}	0.056^{a}	0.056^{a}
	(0.007)	(0.007)	(0.006)	(0.005)	(0.011)	(0.011)	(0.011)	(0.010)
Patent citations × Clerical	0.008	0.007	0.007	0.008	0.039^{a}	0.038^{a}	0.038^{a}	0.039^{a}
	(0.008)	(0.007)	(0.006)	(0.006)	(0.011)	(0.011)	(0.011)	(0.010)
Patent citations × Retail and services	-0.008	-0.010	-0.010	-0.009	-0.017	-0.018	-0.018	-0.018^{c}
	(0.008)	(0.007)	(0.007)	(0.006)	(0.011)	(0.011)	(0.011)	(0.011)
Labor similarity	0.095^{a}	0.074^{a}	0.069^{a}	0.078^{a}	0.015	0.008	0.005	0.009
	(0.006)	(0.006)	(0.006)	(0.006)	(0.012)	(0.012)	(0.012)	(0.012)
Labor similarity × Management and research	-0.097^{a}	-0.090^{a}	-0.087^{a}	-0.091^{a}	-0.001	0.004	0.006	0.004
	(0.007)	(0.007)	(0.007)	(0.007)	(0.013)	(0.013)	(0.013)	(0.013)
Labor similarity × Clerical	-0.032^{a}	-0.022^{a}	-0.018^{a}	-0.038^{a}	-0.008	-0.004	-0.002	-0.013
	(0.008)	(0.007)	(0.007)	(0.007)	(0.014)	(0.014)	(0.014)	(0.014)
Labor similarity × Retail and services	-0.012	-0.015^{c}	-0.013^{c}	-0.034^{a}	0.057^{a}	0.059^{a}	0.060^{a}	0.048^{a}
	(0.008)	(0.008)	(0.007)	(0.007)	(0.014)	(0.014)	(0.014)	(0.014)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls included	No	Yes	Yes	Yes	No	Yes	Yes	Yes
General equilibrium controls	No	No	Correlation	Bonacich	No	No	Correlation	Bonacich
Observations	41,168	41,168	41,168	41,168	41,168	41,168	41,168	41,168
R-squared	0.772	0.779	0.807	0.806	0.347	0.352	0.359	0.362

Notes: Results for all $4 \times 3 \times 3,655$ unique industry pairs and four functional employment types obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All computations for 0-800 kilometers, with 1 kilometer steps for the K-density CDF, and 5 kilometers steps and 200 bootstrap replications for the (global) confidence bands for the strength of agglomeration. The dependent variables are computed at 25 kilometers distance in both panels. Regression for all employment types pooled, where the excluded category is production employment. The shares we use are the baseline shares. The included controls are: Locational advantage (CDF, 25km) *or* Locational advantage (Strength, 25km); input share (primary); output share (primary); input share (business services); output share (business services); multiplant share; and within-plant coagglomeration. The general equilibrium controls are either the locational correlations or the Bonacich centrality. Huber-White robust standard errors in parentheses.

in the literature about the geography of innovation. Jaffe (1989) provides an estimation of the knowledge production function — based on R&D expenses and patents — to evaluate the extent to which research generates knowledge spillovers. Many extensions of this framework have been proposed, and they conclude more or less all that knowledge spillovers have a quite limited spatial extent. The use of patent data to track knowledge flows is now very common. Nevertheless, patents do not reveal all possible knowledge sharing for two main reasons: (i) sectors do not have the same propensity to patent because patents are expensive and also reveal information to competitors, so that not all technological spillovers between industries are

embodied in patents (Audretsch and Feldman, 1996; Carlino and Kerr, 2015); and (ii) there are other sources of knowledge spillovers — patents only pertain to the formal exchange of knowledge, while informal exchanges, through face-to-face contacts or social events, are also important channels. Several case studies stress the importance of up-to-date information about products, services, competitors or production processes, and the effectiveness of its exchanges through face-to-face contacts, which are easier in close spatial proximity (see, e.g., Saxenian, 1994, or Arzaghi and Henderson, 2008). Storper and Venables (2004) refer to the 'buzz' of large cities — the particular atmosphere of information exchanges due to a dense environment of diverse people that lead to the cross-fertilization of ideas.

An alternative to the use of patent data is the use of indices that aim to capture the 'content of knowledge'. Based on Spanish survey data, Jofre-Monseny, Marín-López, and Viladecans-Marsal (2011) develop an index of technological similarity that measures the shared knowledge between firms. Gabe and Abel (2012, 2016) construct an index that measures the similarity of knowledge required to perform a job, using the O*NET database (from the U.S. Department of Labor's Occupational Information Network). Although these indices capture different aspects of knowledge flows, they suffer from disadvantages too: they rely on specific questions in available surveys, and it is difficult to apply them more broadly to other empirical studies. Testing empirically the role of knowledge spillovers as a determinant of coagglomeration is, therefore, still challenging. Both formal and informal knowledge flows follow channels that cannot be easily tracked. Hence, our generally weak effect of patent citations should not be interpreted in the sense that knowledge flows play no role in driving coagglomeration patterns.[26] There are clearly measurement problems but, as shown in this paper, disentangling knowledge-intensive functions from other functions also provides another cut at these difficulties.

Second, our results may be affected by the close link between labor movements and knowledge flows. Indeed, knowledge is embodied in workers who thus support knowledge transfers. Several surveys of high-skilled workers, mainly engineers as in Dahl and Pedersen (2004), identify the importance of knowledge flows for clusters through formal and informal contacts between workers. These case studies deal with a small number of clusters by focusing on a particular industry or region or city, thereby making generalizations of their results difficult. To overcome this limitation, several reviews of the literature on clusters have been conducted. For Canada, Wolfe and Gertler (2004) survey 26 case studies of industrial clusters conducted by the Innovation Systems Research Network. They underline that labor movements and informal collaborations are crucial for the success of clusters. Combes and Duranton (2006) go a step further concerning the overlap between the role of labor markets and of knowledge flows in coagglomeration patterns. They argue that the labor force and knowledge spillovers are not

[26]When taken together, most studies find results that are in line with ours: the effects of knowledge spillovers are smaller than those of the others Marshallian determinants (see, e.g., Rosenthal and Strange, 2003; Jofre-Monseny *et al.*, 2011).

distinct determinants of agglomeration. Indeed, their model shows that agglomeration could be linked to the transfer of knowledge through labor movements in local labor markets. The model is based on the stylized fact that most workers switch firms mainly locally. Several case studies corroborate the local nature of labor movements (for example, Saxenian, 1994, or Fallick, Fleischman, and Rebitzer, 2006, on Silicon Valley). While these arguments are relevant, we think that labor market pooling and knowledge flows cannot be considered the same determinants in our analysis since they have different impacts on coagglomeration patterns. Our results by functional employment types corroborate this assertion. Indeed, the labor similarity variable is significantly associated with coagglomeration in most regressions, while the knowledge sharing variable is only associated with coagglomeration for the 'knowledge intensive' functions. This finding reveals the relevance of using patent citations to measure knowledge sharing, at least for functions that have a higher propensity to rely on patents, such as management and research. Our results are also consistent with the argument that a part of knowledge is shared through formal mechanisms, independently from the knowledge that is embodied in workers and exchanged by them through labor movements.

While the role of workers in mediating knowledge spillovers has been widely analyzed in the literature, the part of knowledge transferred through the buyer-seller relationships in intermediate or final products has attracted less attention. In a recent contribution, Howard *et al.* (2016) use questions from the Vietnamese enterprise survey to construct an index that measures the technology transfer between suppliers and clients. We cannot capture that dimension with our data. It could explain why input-output links are significantly associated with coagglomeration for management and research: a part of knowledge transfers remains in the input-output links, even when we split employment by functional types.

Third, our results suggests that input-output links are the most robust and strongest determinant of coagglomeration, with more systematic effects than labor similarity. One reason for that result may be that input-output links are better measured. Another reason is directly related to the construction and inclusion of the labor similarity variable which, as explained before, may be problematic. Indeed, the latter is not directly tied to interactions between two industries. As such, it is more likely to capture general location patterns and the urban environment, contrary to the input-output variable that is specific to interactions between two industries. Note that in Table 3, the labor movement variable, which captures pairwise interactions between two industries, remains significant in specification (14), although the labor similarity variable is not in the similar specification (13), which lends some support to that interpretation. Furthermore, our results using the strength of coagglomeration — a relative measure — as dependent variable show that input-output links are systematically more important that labor similarity. Since the strength of coagglomeration captures the coagglomeration in excess of the general location patterns of the industries, this finding suggests that labor similarity may be more a driver of the locations of individual industries, not of industry pairs in

particular. Thus, caution is required when interpreting the large coefficients for that variable, especially when measuring coagglomeration in absolute terms.

To summarize, our results are in line with previous work on the determinants of coagglomeration, and similar critiques about the proxies that we use remain relevant. Given our data, it is not possible to separate labor market pooling from knowledge spillovers in the presence of workers flows across industries, and to capture cleanly the transfer of knowledge through input-output links. Some of our results highlight the complex links between them. For example, in Table 3, the knowledge sharing variable gets stronger and more significant when we exclude the industry pairs that have the most similar overall location patterns. This may be due to the fact that the industries with the most similar overall location patterns are also very similar along all Marshallian dimensions, which then makes it difficult to pick up the estimates for knowledge sharing as they are soaked up by the input-output links or labor similarity variables. To make further progress here, improvements in measures are required. However, taking into account the functional employment type in the regressions is a first step to capture heterogeneity in the coagglomeration patterns and the respective importance of the determinants driving them. As expected, major differences emerge in coagglomeration patterns, and those differences can be exploited in an informative way.

7 Conclusions

We analyze horizontal location patterns of industry pairs and vertical location patterns of functions using detailed microgeographic data. By jointly analyzing those patterns, we provide new insights into the relative importance of the agglomeration mechanisms underlying the spatial economy. We find that heterogeneity in the location patterns of functions in the value chain provides useful information for the identification of agglomeration mechanisms, because different functions benefit to a different degree from those mechanisms. While the location of production is sensitive to the presence of vertically linked industries and the composition of the local labor pool, these are less important for the location of management and research, which are more sensitive to shared knowledge. Consistent with that result, employment in management and research displays a markedly different spatial profile of coagglomeration than employment in production, with the former mainly coagglomerated at distances of less than 50 kilometers, and the latter mainly coagglomerated at distances of about 150-200 kilometers.

Our results provide support for agglomeration theories and show that extant estimates of average effects using total employment mask substantial heterogeneity. In particular, the proxy for shared knowledge is sensitive to whether or not we consider functional splits. While patent citations across industry pairs are positively associated with the coagglomeration of total employment in 22% of the specifications that we estimate, the corresponding figures are 68% for the coagglomeration of employment in management and research, but only 3%

for employment in production. Clearly, average effects mask substantial heterogeneity and may lead us to believe that knowledge sharing does not matter or is badly captured by the proxy. While more work is required to better measure knowledge sharing and to more cleanly separate knowledge sharing from both input-output links and cross-industry labor mobility, our results suggest that we already have some mileage by more carefully considering what functions are carried out in what locations.

Acknowledgements. We thank our discussant, Ferdinando Monte, as well as Nate Baum-Snow, Théophile Bougna, Mark Brown, Don Davis, Gilles Duranton, Amit Khandelwal, Bill Kerr, Fabian Lange, Julien Martin, Richard Shearmur, Will Strange, and seminar and conference participants at LMU Munich, the 2016 Rotman-Sauder Summer Conference in Real Estate and Urban Economics, the 2015 NARSC Meetings in Portland, the World Bank conference on secondary towns, CIRANO Montréal, HSE Saint Petersburg, and Columbia University for valuable comments and suggestions. We are grateful to Richard Shearmur and Mario Polèse at INRS Montréal (Urbanisation Culture Société) for sharing the special census tabulations from Statistics Canada; and to Bill Kerr for sharing the patent citation data. 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*. Guillain gratefully acknowledges financial support from the PARI Programs of the 'Conseil Régional de Bourgogne'. The study has been funded by the Russian Academic Excellence Project '5-100'. The views expressed in this paper, and all remaining errors, are ours.

References

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

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

[3] Alcácer, Juan, and Mercedes Delgado. 2017. "Spatial organization of firms and location choices through the value chain." *Management Science*, forthcoming

[4] Anas, Alex, and Kai Xiong. 2003. "Intercity trade and the industrial diversification of cities." *Journal of Urban Economics* 54**(2)**: 258–276.

[5] Anderson, James E., and Eric van Wincoop. 2003. "Gravity with gravitas: A solution to the border puzzle." *American Economic Review* 93**(1)**: 170–192.

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

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

[8] Autor, David H., David Dorn, and Gordon H. Hanson. 2013. "The China syndrome: Local labor market effects of import competition in the United States." *American Economic Review* 103**(6)**: 2121–2168.

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

[10] Ballester, Coralio, Antoni Calvó-Armengol, and Yves Zenou. 2006. ""Who's who in networks. Wanted: The key player." *Econometrica* 74**(5)**: 1403–1417.

[11] Behrens, Kristian. 2016. "Agglomeration and clusters: Tools and insights from coagglomeration patterns." *Canadian Journal of Economics* 49**(4)**: 1293–1339.

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

[13] Behrens, Kristian, and W. Mark Brown. 2017. "Transport costs, trade, and geographic concentration: Evidence from Canada." In: Blonigen, Bruce A. and Wesley W. Wilson (eds.), *Handbook of International Trade and Transportation*, forthcoming.

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

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

[16] Caliendo, Lorenzo, Ferdinando Monte, and Esteban Rossi-Hansberg. 2015. "The anatomy of French production hierarchies." *Journal of Political Economy* 123**(4)**: 809–852.

[17] Carlino, Gerald, and William R. Kerr, 2015 "Agglomeration and innovation." In: Duranton, Gilles, J. Vernon Henderson, and William C. Strange (eds.) *Handbook of Regional and Urban Economics, vol. 5*. North-Holland: Elsevier B.V., pp. 349–404.

[18] Combes, Pierre-Philippe, and Gilles Duranton. 2006. "Labor pooling, labor poaching, and spatial clustering." *Regional Science and Urban Economics* 36**(1)**: 1–28.

[19] 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.

[20] Dahl, Michael S., and Christian O.R. Pederson. 2004. "Knowledge flows through informal contacts in industrial clusters: Myth or reality?" *Research Policy* 33**(10)**: 1673–1686.

[21] Davidson, Carl, Fredrik Heyman, Steven Matusz, Fredrik Sjöholm, and Susan Chun Zhu. 2016. "Global engagement and the occupational structure of firms." Mimeographed, Michigan State University.

[22] Davis, Donald R. and Jonathan I. Dingel. 2014. "The comparative advantage of cities." NBER Working Paper #20602, National Bureau of Economic Research, MA.

[23] 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.

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

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

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

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

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

[29] Ellison, Glenn D., and Edward L. Glaeser. 1997. "Geographic concentration in U.S. manufacturing industries: A dartboard approach." *Journal of Political Economy* 105**(5)**: 889–927.

[30] 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.

[31] Faggio, Giulia, Olmo Silva, and William C. Strange. 2017. "Heterogeneous agglomeration." *Review of Economics and Statistics*, forthcoming.

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

[33] Fujita, Masahisa, and Jacques-François Thisse. 2006. "Globalization and the evolution of the supply chain: Who gains and who loses?" *International Economic Review* 47**(3)**: 1937–1958.

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

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

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

[37] 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.

[38] 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.

[39] Hendricks, Lutz. 2011. "The skill composition of US cities". *International Economic Review* 52**(1)**: 1–32.

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

[41] Jaffe, Adam B. 1989. "Real effects of academic research." *American Economic Review* 79**(1)**: 957–970.

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

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

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

[45] Monte, Ferdinando, Stephen J. Redding, and Esteban Rossi-Hansberg. 2015. "Commuting, migration and local employment elasticities." NBER Working Paper #21706, National Bureau of Economic Research, MA.

[46] Pierce, Justin R., and Peter K. Schott. 2016. "Trade liberalization and mortality: Evidence from U.S. counties." NBER Working Paper #22849, National Bureau of Economic Research, MA.

[47] 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.

[48] Rosenthal, Stuart S., and William C. Strange. 2004. "Evidence on the nature and sources of 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. 2119–2172.

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

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

[51] Saxenian, Annalee. 1994. " *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*." Harvard University Press: Cambridge, MA.

[52] Storper, Michael, and Anthony J. Venables. 2004. "Buzz: face-to-face contact and the urban economy." *Journal of Economic Geography* 4**(4)**: 351–370.

[53] 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.

[54] Wolfe, David A., and Meric S. Gertler. 2004. "Clusters from the inside and out: Local dynamics and global linkages." *Urban Studies* 41**(5/6)**: 1071–1093.

Appendix

Appendix **A** presents our data and discusses details of how we construct different variables. Appendix **B** explains how we compute our coagglomeration measures and how we adjust the functional employment shares.

A. Data

A.1. Special census tabulations and plant-level data

To construct our coagglomeration measures, we combine two datasets. The first are special census tabulations from Statistics Canada, which split industry-level employment by census division and functional type (see the left panel of Figure 2 for an illustration of the geographic structure of our data). There are six aggregate functional employment types, which are based on the 1991 Standard Occupational Classification (SOC): 'Managers, directors, and related occupations' (type 1); 'Natural sciences, engineering, mathematics, social sciences' (type 2); 'Religion, education, health care, arts, recreation' (type 3); 'Administration and related activities' (type 4); 'Retail and services' (type 5); and 'Agriculture, fishing, forestry, mines, construction, transport' (type 6).

Table 8: Functional employement categories.

Type	Occupation title, special tabulations	Our classification	1991 SOC categories
0		Total employment	All
1	Managers, directors and related occupations	Management and research	A, B0, B1, B3
2	Natural sciences, engineering, mathematics, social sciences	Management and research	C, E0, E211, E212, E213
3	Religion, education, health care, arts, recreation	(excluded)	D, E1, E214, E215, E216, F
4	Administration and related activities	Clerical	B2, B4, B5
5	Retail and services	Retail and services	G
6	Agriculture, fishing, forestry, mines, construction, transport	Production	H, I, J

Notes: Relationship between the 1991 SOC classification and our functional employment categories. See Polèse and Shearmur (2005) for additional details on the data.

We exclude all unclassified functions and employment of type 3, and work with four broad functional employment types: 'Management and research' (sum of types 1 and 2); 'Clerical' (type 4); 'Retail and services' (type 5); and 'Production' (type 6). 'Total employment' (type 0) is the sum of types 1 to 6. Table 8 summarizes the categories and shows their relations to the SOC classification.[27] Each type is reported by industry and by census division. Concerning industries, the special tabulations split employment by functions at an intermediate level between the 3- and the 4-digit NAICS. We create a concordance that associates each 4-digit NAICS

[27]Type 3 is included in our total employment, but we do not report detailed coagglomeration results for it. 'Religion, education, health care, arts, recreation' is only sparsely reported and relates little to manufacturing.

code with an industry code from the special tabulations. We focus on the 86 manufacturing industries only, because we have detailed microgeographic data for those industries (see below). Data are for the 1996 and the 2001 censuses, and the geographic units are time consistent.

One of the key aspects of these special census tabulations is that they split census divisions into rural and urban parts.[28] Census divisions are indeed relatively large administrative constructs that are not clearly linked to any urban or rural divide. This poses problems when using such administrative units to work on functional specialization because urban and rural areas, as well as cities of different sizes, differ substantially in the functions they perform.

Using the special census tabulations, we compute the shares θ_{ic}^{f}, separating the rural from the urban parts. We then apply these shares to the plants to split their employment by functional type. Our plant-level data comes from the *Scott's National All* database. This establishment-level database builds on the Business Register and contains information on plants operating in Canada. It contains about 47,000–50,000 manufacturing plants per year and extensively covers all manufacturing industries. Although the Scott's dataset is only a large sample and not the universe of manufacturing plants, it has a very wide (85–90%) and representative coverage. It contains almost all of the large plants and many small plants. Behrens and Bougna (2015, Appendix A) provide detailed information on the data quality and its representativeness — both in terms of provinces and industries — of the manufacturing portion of the database.

We use data for the years 2001, 2003, and 2005. For every establishment, we have information on its primary 6-digit NAICS code, up to four secondary NAICS codes, its total employment, and its 6-digit postal code.[29] 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). We use the postal code data for the next year in order to take into account that there is a six months delay in the updating of postal codes. For example, the census geography of 1996 and the postal codes as of May 2002 (818,907 unique postal codes) were associated with the 2001 Scott's data, whereas we matched the 2003 and 2005 Scott's data with the 2001 census geography and the corresponding PCCF's. We further associate standard geographical identifiers of the postal code's census division (CD) and census metropolital area (CMA) with each plant: the 1996 census with the 2001 Scott's data, and the 2001 census with the 2003

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

[29]The *Scott's All* database unfortunately does not provide firm identifiers (only plant identifiers). Yet, it reports information on the legal name of the entity that owns the plant, and we use this information to group plants into firms. The shares of multiunit plants we obtain for the different industries are similar to those of the manufacturing portion of the Business Register or the Annual Survey of Manufacturers. We cross-checked our measure against aggregate measures from Statistics Canada — which we ordered as special tabulations from confidential data — and the correlations are high and in the range of 0.8 to 0.9.

and 2005 data. We then assign plants to census divisions based on their postal code centroid latitude and longitude coordinates (see the left panel of Figure 2 for an illustration) and compute the employment splits. To mitigate the role of outliers, we trim the top 1% of plants in terms of employment. Doing so takes care of obvious errors in the data and also avoids our employment-based coagglomeration measures to be driven by a few very large observations (since we weight observations by the product of their employment).

Panel (a) of Table 9 summarizes the plant-level employment figures broken down by functional types using the baseline shares from the special tabulations. Panels (b) and (c) report the same information for spatially smoothed (see Appendix B.2) and size-adjusted (see Appendix B.3) functional shares.

Table 9: Summary of plant-level employment by functional type.

	2001				2003				2005			
	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.
Total employment	32.835	65.946	1	630	33.619	68.115	1	692	34.602	68.734	1	685
	(a) Baseline (unadjusted) functional shares											
Management and research	5.874	14.983	0	400	6.827	16.783	0	351.351	7.038	16.936	0	354.436
Clerical	3.802	8.699	0	450	3.078	7.362	0	500	3.168	7.178	0	150.748
Retail and services	2.249	6.588	0	600	1.871	5.735	0	210.446	1.961	6.285	0	268.780
Production	20.185	44.177	0	542.482	21.070	46.270	0	611.765	21.646	46.368	0	611.765
	(b) Commuting pattern-adjusted shares											
Management and research	5.986	15.045	0	397.796	6.935	16.727	0	351.352	7.161	17.0111	0	399.021
Clerical	3.841	8.515	0	188.403	3.077	7.071	0	251.076	3.182	7.174	0	150.748
Retail and services	2.288	6.597	0	598.330	1.891	5.884	0	262.999	1.970	6.182	0	268.783
Production	19.988	43.574	0	529.341	20.938	45.835	0	610.429	21.492	45.871	0	610.455
	(c) Commuting pattern and size-adjusted shares											
Management and research	5.884	13.906	0.129	393.773	7.180	16.916	0.131	329.620	7.385	16.907	0.131	367.196
Clerical	7.197	16.540	0.122	358.991	6.421	15.208	0.124	434.679	6.620	15.167	0.125	306.410
Retail and services	7.222	18.263	0.115	593.009	6.660	17.665	0.116	496.235	6.909	18.224	0.116	503.805
Production	6.461	14.347	0.022	359.497	6.701	14.829	0.034	417.390	6.898	14.778	0.060	417.920
Number of plants	50,109				50,357				47,786			

Notes: Figures based on the manufacturing part of the *Scott's National All* 2001, 2003, and 2005 datasets, as well as the special tabulations from Statistics Canada. In the baseline case (a), total employment is broken down using the weights $\theta_{i,c}^{f}$ computed from the special tabulations of *Statistics Canada*. 'Management and research', 'Clerical', 'Retail and services', and 'Production' do not sum to total employment because we exclude the category 'Religion, education, health care, arts, recreation'. See Appendices B.1 and B.2 for a description of cases (b) and (c).

Because we split census divisions into rural and urban parts, two plants with the same industry-year-census division can have different employment breakdowns if one of them is located in the rural part and the other is located in the urban part of the census division. As Table 9 shows, plants had on average 32.83 employees in 2001, of which 5.94 were in 'management and research', 3.82 in clerical, 2.27 in 'retail and services', and 20.06 in production. Clearly, production workers constitute the lion's share of total employment in the manufacturing sector. The average plant size increased slightly between 2001 and 2005, whereas the number of plants decreased. The increase in size is due to a slight growth in 'management and research' and production employment, and a slight decrease in the other categories.

A.2. Marshallian covariates

Our Marshallian covariates consist in proxies for input-output links, labor market pooling, and knowledge spillovers. Table 10 provides descriptive statistics.

Table 10: Summary statistics for the Marshallian covariates, U.S. instruments, and controls.

Variable	Description	Mean	S.D.	Min.	Max.
	Marshallian covariates				
Input-output links	Maximum input-output share between i and j	0.014	0.036	0.000	0.806
Input links	Maximum input share between i and j	0.011	0.0288	0.000	0.630
Output links	Maximum output share between i and j	0.009	0.029	0.000	0.806
Labor similarity	Labor similarity between i and j	0.317	0.210	0.006	0.993
Labor similarity (Mngmt and research)	Labor similarity, Management and research	0.785	0.198	0.102	0.996
Labor similarity (Clerical)	Labor similarity, Clerical	0.847	0.107	0.245	0.993
Labor similarity (Services and retail)	Labor similarity, Services and retail	0.677	0.224	0.002	0.999
Labor similarity (Production)	Labor similarity, Production	0.281	0.231	-0.018	0.996
Labor movement	Labor movement between i and j	0.018	0.036	0.000	0.516
Patent citations	Share of cross-industry citations, use based	0.018	0.047	0.000	0.798
Patent citations (make based)	Share of cross-industry citations, make based	0.019	0.050	0.000	0.860
	U.S. instruments				
Input-output links (U.S. IV)	Maximum input-output share (US BEA) between i and j	0.011	0.037	0.000	0.937
Input links (U.S. IV)	Maximum input share (US BEA) between i and j	0.010	0.033	0.000	0.589
Output links (U.S. IV)	Maximum output share (US BEA) between i and j	0.009	0.034	0.000	0.937
	Controls				
Locational advantage (CDF, 25km)	CDF of natural advantage-based coagglo. measure, 25km	0.017	0.072	0.000	1.001
Locational advantage (Strength, 25km)	Strength of natural advantage-based coagglo. measure, 25km	-0.008	0.020	-0.534	0.380
Multiplant share	Max of shares of multiunit plants in i and j	0.156	0.103	0.000	0.447
Within-plant coagglomeration	Max of shares of plants in i reporting activity in j	0.013	0.036	0.000	0.750
Input share (primary)	Max of shares of inputs sourced from primary industries (NAICS 11–22)	0.205	0.193	0.003	0.806
Output share (primary)	Max of shares of output sold to primary industries (NAICS 11–22)	0.114	0.163	0.006	0.851
Input share (business services)	Max of shares of inputs sourced from business services (NAICS 52–55)	0.082	0.027	0.030	0.186
Output share (business services)	Max of shares of output sold to business services (NAICS 52–55)	0.062	0.041	0.004	0.267
Locational correlations	Correlation coefficient of K-densities with 3rd industries	0.791	0.316	-0.721	0.999
Bonacich centrality	Bonacich centrality measure of the coagglomeration matrix	1.791	0.217	1.156	2.216

Notes: Descriptive statistics based on 10,965 4-digit industry pairs (86 4-digit manufacturing industries) and three years (2001, 2003, and 2005) pooled.

Input-output links. We use detailed input-output matrices for the years 1998, 2000, and 2002, which we associate with the plant-level data in 2001, 2003, and 2005, respectively. These matrices are constructed using the finest public release of the Canadian input-output tables at the L-level (link level), which is between NAICS 3- and 4-digit. We first disaggregate the input-output matrices to the W-level (NAICS 6-digit) using sales or employment data as sectoral weights, and then reaggregate them to the 4-digit level.[30] Let $\omega_{i,j}^{\text{in}}$ denote the share of inputs sourced by industry i from industry j. Conversely, let $\omega_{i,j}^{\text{out}}$ denote the share of output sold by industry i to industry j. These shares are computed taking into account all industries (including primary industries and services, but excluding private consumption and the different

[30]Due to confidentiality reasons, we cannot directly use the W-level matrices that are internally available at *Statistics Canada*. However, tests we ran in Behrens *et al.* (2015) using those matrices yielded similar results to those using the matrices constructed by our methodology.

government aggregates and imports/exports). Our measure of the strength of input-output links between industries i and j is $\text{io}_{ij} \equiv \max\{\omega_{ij}^{\text{in}}, \omega_{ji}^{\text{in}}, \omega_{ij}^{\text{out}}, \omega_{ji}^{\text{out}}\}$. We also use the strength of input links $\text{in}_{ij} \equiv \max\{\omega_{ij}^{\text{in}}, \omega_{ji}^{\text{in}}\}$ and of output links $\text{out}_{ij} \equiv \max\{\omega_{ij}^{\text{out}}, \omega_{ji}^{\text{out}}\}$ separately as robustness checks in our analysis.

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

Labor market pooling. We first construct a measure of occupational employment similarity of the workforce in the different industries. To this end, we use Occupational Employment Survey (OES) data from the Bureau of Labor Statistics (BLS) for 2002, 2003, and, 2005 to compute the share of each of 554 occupations in each 4-digit NAICS industry.[31] We use 2002 data for the 2001 plant sample, and then data for each year t for the plant sample in year t. Using 2002 as the starting year for the OES data allows us to avoid the difficult concordance from SITC to NAICS. Our measure of occupational employment similarity for total employment, OES_{ij}^{0}, is computed as the correlation between the vectors of occupational shares of industries i and j. Turning to individual functions, we further decompose occupational employment similarity by functional types. To do so, we assign each occupation code from the 554 OES occupations to one of our six functional employment categories. We then recompute the pairwise industry correlations using the subvectors of the employment shares of these functional categories only. This yields four additional measures of occupational similarity within each functional type. We refer to them as OES_{ij}^{1}, OES_{ij}^{4}, OES_{ij}^{5}, and OES_{ij}^{6}.

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

[31] There are 808 occupations in total in the OES data. We only use occupations for which there is at least some employment in manufacturing (e.g., there are no 'Surgeons' in manufacturing industries, hence we exclude them completely from our data).

use the same values of mov_{ij} across the three years of our geographic data. This explains why we use this measure as a robustness check only.

Knowledge flows. Last, we construct proxies for 'knowledge spillovers' using the NBER Patent Citation database, following previous work by Kerr (2008). We construct two proxies: (i) know_{ij}^{m}, which is the maximum of the shares of patents that industry i (or j) manufacture and which originate from the other industry; and (ii) know_{ij}^{u}, which is the maximum of the shares of patents that industry i (or j) use and which originate from the other industry.

A.3. Controls

Our controls include variables that capture locational advantage, different aspects of industrial organization, and general equilibrium effects.

Locational advantage. First, we construct benchmark coagglomeration measures based on counterfactual plant distributions predicted from a location choice model including a large number of covariates. These coagglomeration measures are used to control for locational advantage that stems from a wide range of factors related to infrastructure (roads and rail), access to the sea, to ports and so on.[32] See the Supplemental Appendix S.4 for all details and a summary of results. Second, we build proxies for the industries' dependence on natural resources and business services. We construct 'primary' and 'service' input and output shares for each manufacturing industry, based on our disaggregated input-output tables. The primary input shares are the shares of inputs that industries source from primary industries (NAICS 11, 21, and 22). Similarly, the primary output shares are the shares of their output that industries sell to primary industries. We construct similar input and output shares for business services (NAICS 52, 53, 54, and 55). For each industry pair ij we include the maximum of the two industry shares into our analysis to control for natural advantage and service dependency broadly defined.[33] The idea is that manufacturing industries that rely heavily on primary inputs or business services may have location patterns that deviate from those predicted by the underlying logic of coagglomeration, because resource-based industries and business service industries are themselves concentrated in — or dispersed across — specific places.

Industrial organization. When analyzing the coagglomeration of specific employment types by industry, we further have to take into account the fact that the organizational structure of

[32] We may view infrastructure as a Marshallian agglomeration force related to 'sharing'. Hence, our benchmark coagglomeration measures may soak up some Marshallian factors, yet they should not encapsulate our key variables of interest, namely input-output links, labor market pooling, and knowledge spillovers.

[33] Because the ordering of the indices ij in each pair is essentially arbitrary, all variables in our analysis are symmetric with respect to i and j (see Ellison *et al.*, 2010, for a discussion of that point).

firms in the industry may influence their coagglomeration patterns (see the Supplemental Appendix S.1 for a simple extension of our framework to the case of multiunit firms). Consider, e.g., an industry in which all establishments are standalone firms with a single plant only. In that case, the coagglomeration of management and research employment and of production employment necessarily follow each other and that of total employment because firms cannot separate their functions geographically. Conversely, an industry in which all firms are multiunit may display very different colocation patterns by function. For example, while production employment may be located in different rural areas and therefore not be coagglomerated, 'management and research' functions may be coagglomerated in the same urban areas. We conjecture that coagglomeration patterns differ systematically by function and organizational form of firms in industries. To control for that, we build proxies for firms' ability to geographically split their functional structure. More specifically, we include for each industry pair ij the maximum of the two sectoral shares of multiunit plants. Table 10 shows that there is substantial variation across industries in their organizational structure. The average share of multiunit plants is 15.7% (pooled across years), with a low of 0% for 'Other Transportation Equipment Manufacturing' (NAICS 3369) and a high of 44.7% for 'Basic chemicals manufacturing' (NAICS 3251).

Table 11: Correlations between industrial organization controls and Marshallian covariates.

	Multiplant share	Within-plant coagglo.	Input-output links	Patent citations	Labor similarity
Within-plant coagglo.	-0.084*				
Input-output links	-0.032*	0.451*			
Patent citations	0.011	0.130*	0.098*		
Labor similarity	-0.1617*	0.390*	0.294*	0.104*	
Labor movement	-0.049*	0.491*	0.391*	0.123*	0.358*

Notes: Simple correlation coefficients, results for all 10,965 industry pairs pooled across the three years 2001, 2003, and 2005. * denotes significant at 1%.

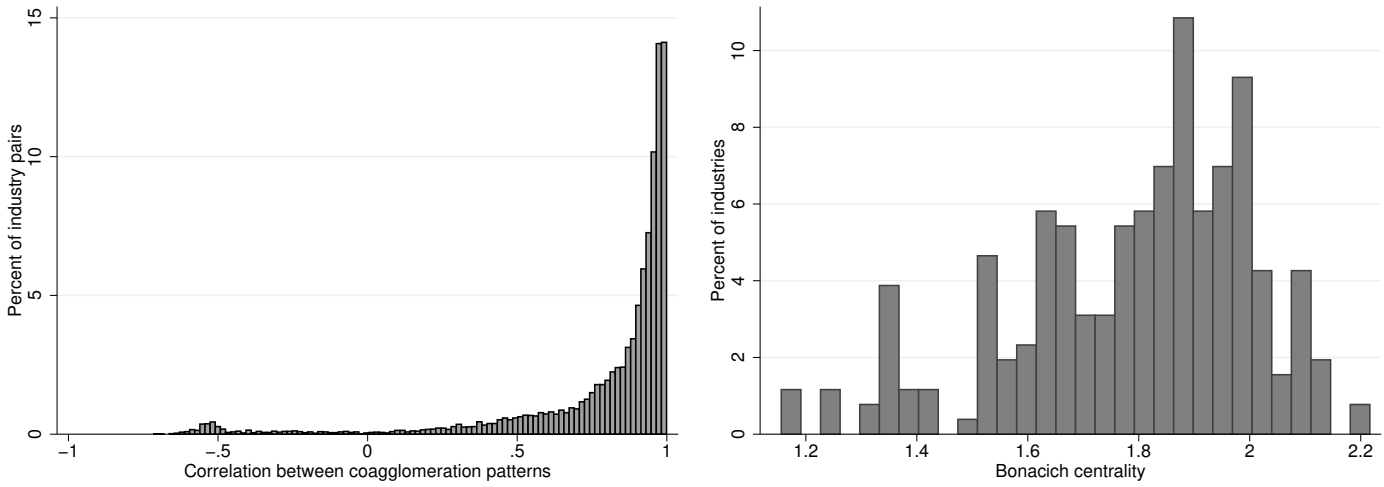
Next, we exploit an interesting feature of our dataset to control for the fact that plants can operate in multiple industries. More precisely, we use our plants' secondary NAICS codes to compute the share of plants with primary NAICS code i that report also secondary codes in industry j. This provides us with a 'within-plant coagglomeration metric' of industries. Industries pairs with a high value within_{ij} are industries that are 'coagglomerated more strongly within plants'. Controlling for this within-plant coagglomeration is important for two reasons. First, the within-metric is strongly correlated with the Marshallian covariates, so that omitting it may bias the estimates (see Table 11; and Behrens, 2016, for a discussion). Second, within-plant coagglomeration affects the coagglomeration measures that we use as dependent variable. Indeed, depending on which activities are coagglomerated within the boundaries of the plant, and which are coagglomerated outside the boundaries of the plant, the coagglomeration measure will be different. If two plants in industries i and j operate separately next to each other, their small distance d_{ij} will enter the computation of the coagglomeration measure;

whereas if the two plants operate as one entity, their distance d_{ij} does not enter.

General equilibrium effects. Controlling for general equilibrium effects is important since the coagglomeration of industries does not only take place between industry pairs ij, but involves multiple other industries. We have to take into account the fact that two industries i and j may end up colocated without any coagglomeration economies and without any locational fundamentals just because of the existence of a third industry k that they are both coagglomerated (or colocated) with. These 'third industry effects' are one of the reasons why inefficient coagglomeration may occur (Helsley and Strange, 2014). Controlling for those network effects in a theory-based way requires a structural model that captures the interdependence in the location patterns of all industries. This is beyond the scope of this paper, but we will show how to control (at least partly) for those effects in a reduced-form way.

Figure 6: Histogram of third-industry effects and Bonacich centralities.

(a) Third-industry effects. (b) Bonacich centralities.



Notes: Histogram of correlation coefficients ρ_{ij}, computed as the simple correlation between the coagglomeration measures of industry i with all industries except j, and the coagglomeration measures of industry j with all industries except i. Bonacich centrality computed on the matrix of coagglomeration measures for each year, following Ballester *et al.* (2006).

We proceed in two ways. First, for each industry pair ij, we compute the correlation coefficient between the coagglomeration of industries i and j with all other industries $k \neq i, j$. A high correlation means that industries i and j have very similar location patterns with respect to the other industries which, by transitivity, then implies that they should have a very similar location pattern with respect to each other. We then either include that measure into our regressions, or we use it to exclude the top x% of industry pairs that have the highest correlations — i.e., the pairs that have location patterns that are very similar to those of many other industries — from the regressions. Figure 6 shows that many industries display very similar

overall coagglomeration patterns with respect to the other industries. This suggest that a large share of the observed coagglomeration may be spurious and simply induced by the presence of third industries. As shown by our regressions, third industry effects are highly correlated with pairwise coagglomeration patterns. This suggest that a large share of coagglomeration may be spurious.

Second, we may view the coagglomeration measures between industry pairs ij as the matrix of an undirected graph, where the coagglomeration measures — which are between 0 and 1 — are weights on the arcs. Following Ballester *et al.* (2006) we can then compute the *Bonacich centrality* of each industry in that network to measure which industries are the most influentially connected with the other industries. A high value of that measure means that the industry is 'central' in the coagglomeration network, i.e., strongly coagglomerated with industries that are themselves strongly coagglomerated with many other industries. Those industries are thus likely to be spuriously coagglomerated. Again, we can either include the Bonacich centrality as a control in our regressions, or we can exclude the industry pairs where one of the two industries has a high centrality measure. When we include the Bonacich centrality into the regressions, we take the maximum of the measure for the two industries i and j. The correlation between the locational correlations ρ_{ij} and the centrality measures — $\max\{\text{bonacich}_i, \text{bonacich}_j\}$ — is 0.2, which shows that those two measures capture fairly different things despite being correlated.

B. Measuring coagglomeration and adjusting functional shares

In this appendix, we explain how we measure coagglomeration and how we adjust our functional shares to control for commuting patterns and differences in firm sizes.

B.1. Measuring coagglomeration

We compute continuous point-pattern based coagglomeration measures following Duranton and Overman (2005, 2008). Let ℓ_k^f denote the employment of function f in plant k in industry i, and ℓ_l^f the corresponding employment in plant l in industry $j \neq i$. The DO coagglomeration measure for the industry pair ij (with n_i plants in industry i and n_j plants in industry j), and function f, is given by

$$\widehat{K}_{ij}^f(d) = \frac{1}{h \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \ell_k^f \ell_l^f} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \ell_k^f \ell_l^f f\left(\frac{d - d_{kl}}{h}\right), \tag{B.1}$$

where d_{kl} is the great circle distance between plants k and l; where h is the optimal bandwidth of the kernel, set according to Silverman's rule; and where f is a Gaussian kernel function. We estimate (B.1) for each of the 3,655 unique NAICS 4-digit industry pairs in each year, using an

800 kilometers distance cutoff and a step size of 1 kilometer. We then use (B.1) to construct our dependent variables that capture the extent of coagglomeration between industry pairs. It is a priori unclear how to extract a measure of the strength of coagglomeration from the DO K-densities. In what follows, we use two approaches.

First, we follow Behrens *et al.* (2015) and take the cumulative of (B.1) up to some distance threshold $\overline{d}$. Formally,

$$\text{DO\_CDF}_{ij}^{f}(\overline{d}) = \sum_{d \leq \overline{d}} \widehat{K}_{ij}^{f}(d). \tag{B.2}$$

The measure (B.2) can be interpreted as the (kernel-smoothed) probability that a pair of employees working in two plants randomly drawn from industries i and j are working less than $\overline{d}$ kilometers away from one another. Note that we can smoothly vary the threshold $\overline{d}$ to compute coagglomeration measures at various distances. This will allow us to estimate the importance of the determinants of coagglomeration at various spatial scales.

Alternatively, Duranton and Overman (2005) suggest to measure the strength of localization by summing the K-density in excess of some appropriately defined confidence band.[34] Let $\overline{K}_{ij}^{f}(d)$ denote the upper bound of that confidence band. Formally, the strength of coagglomeration is then defined as

$$\Gamma_{ij}^{f}(\overline{d}) = \sum_{d \leq \overline{d}} \gamma_{ij}^{f}(d), \quad \text{where} \quad \gamma_{ij}^{f}(d) = \max\left\{\widehat{K}_{ij}^{f}(d) - \overline{K}_{ij}^{f}(d), 0\right\}. \tag{B.3}$$

The measure (B.3) can be interpreted as the excess probability of finding employees of the two industries closer together than distance $\overline{d}$ under a random location pattern while accepting some level of risk (e.g., 5%). The strength of codispersion is analoguously defined by

$$\Psi_{ij}^{f}(\overline{d}) = \sum_{d \leq \overline{d}} \psi_{ij}^{f}(d), \quad \text{where} \quad \psi_{ij}^{f}(d) = \max\left\{0, \underline{K}_{ij}^{f}(d) - \widehat{K}_{ij}^{f}(d)\right\}, \tag{B.4}$$

where $\underline{K}_{ij}^{f}(d)$ is the lower bound of the confidence band. Note from (B.3) and (B.4) that, by construction, all industry pairs ij that fall in between the confidence bands will have measures of Γ_{ij} and Ψ_{ij} that are zero at distance $\overline{d}$.

Finally, to turn (B.3) and (B.4) into a meaningful dependent variable defined for all industries, we reflect the strength of the codispersion measure Ψ into the negative numbers by taking

[34] We construct (global) confidence bands $[\underline{K}_{ij}^{f}(d), \overline{K}_{ij}^{f}(d)]$ at each distance d by simulating 200 counterfactual industry distributions (see Duranton and Overman, 2005, 2008, for details). We use a fine grained 1-by-1 kilometer estimate of the K-densities to construct our dependent variable. Yet, for our measures of the 'strength' of agglomeration, we need to compute confidence bands using Monte Carlo replications. Computing bands for five employment types, 3 years, 800 kilometers, 1,000 replications, and 3,655 industry pairs proved to be computationally too time-consuming. Hence, the computations for the confidence bands are based on 5km steps (between 0 and 800 kilometers) and 200 bootstrap replication only. Note that we use no approximation techniques and compute the exact K-densities, using a fairly optimized C++ algorithm. Yet, several months are required to compute all 142,545 different coagglomeration measures.

the minus of it (see Kerr and Kominers, 2015). Our second dependent variable — based on the DO measure of the strength of coagglomeration — is then given by:

$$\text{DO\_STR}_{ij}^{f}(\overline{d}) = \begin{cases} \Gamma_{ij}^{f}(\overline{d}) & \text{if} \quad (i,j) \text{ is significantly coagglomerated} \\ 0 & \text{if} \quad (i,j) \text{ is random} \\ -\Psi_{ij}^{f}(\overline{d}) & \text{if} \quad (i,j) \text{ is significantly codispersed} \end{cases} \tag{B.5}$$

Our baseline estimates of (B.2) and (B.5) are for total employment, i.e., we do not split plant-level employment by function f. Once we split plant-level employment by functional types f (see Subsection 4.1 and Appendices B.2 and B.3 below), we can also compute these measures for each function f separately.

Table 12 summarizes our two coagglomeration measures, based on the DO K-densities for the different years and functions. Two points are worth noting. First, across all employment types and for all measures of coagglomeration, the extent of coagglomeration is decreasing over time. Hence, coagglomeration gets weaker, which echoes the findings of Behrens and Bougna (2015) and Behrens *et al.* (2015) who document that geographic concentration in manufacturing has been decreasing in Canada. Second, there is substantial variation in these measures, especially for DO_STR_{ij}^{f}. By construction, the variation of DO_CDF_{ij}^{f} is smaller but still sizable.

Note, finally, that the correlation between the different CDF coagglomeration measures by functions is very large: within each function (and for total employment), the yearly correlations by distance vary between 0.95 and 0.99. The correlation is very high between the spatially smoothed shares and the benchmark case, and it is still very high (though a bit less) between the spatially smoothed and size-adjusted shares and the benchmark shares. This shows that our results will not be really sensitive to spatial smoothing and size adjustments. These findings are fairly similar for the strength of coagglomeration measures, where the correlations exceed 0.90 between the different measures based on different shares.

B.2. Spatially smoothed shares

The shares of functional employment types by census division are based on special census tabulations. As such, these shares are residence-based and not workplace based. This creates potentially problems if there is a lot of commuting, which breaks the links between residence-based shares and composition of the labor force at the plants' locations. To partly control for this, we adjust our shares by smoothing them spatially. To this end, we compute for each plant the smoothed occupational shares in a 25 kilometers radius around the plant. We choose 25 kilometers since 90% of Canadians commute within this distance.

Our procedure works as follows. For each plant, we count the number of postal code centroids that are within 25 kilometers around the plant. We then generate weights based on the number of these postal codes by census division around the plant. These weights are used

Table 12: Summary statistics for all coagglomeration measures, 25 kilometers distance.

	2001				2003				2005			
	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.
	(a)											
	Cumulatives, baseline homogeneous											
Total employment	0.039	0.020	0.001	0.146	0.039	0.019	0.001	0.115	0.038	0.018	0.001	0.113
Management and research	0.042	0.021	0.001	0.119	0.043	0.021	0.001	0.139	0.042	0.020	0.001	0.133
Clerical	0.046	0.022	0.001	0.180	0.044	0.023	0.001	0.187	0.044	0.021	0.001	0.139
Retail and services	0.035	0.017	0.002	0.090	0.035	0.017	0.000	0.282	0.034	0.016	0.001	0.115
Production	0.035	0.017	0.002	0.090	0.038	0.019	0.001	0.122	0.038	0.018	0.001	0.156
	Strength Γ, baseline homogeneous											
Total employment	0.001	0.005	-0.034	0.068	0.000	0.004	-0.036	0.026	0.000	0.003	-0.031	0.030
Management and research	0.002	0.006	-0.038	0.048	0.001	0.006	-0.034	0.047	0.001	0.005	-0.033	0.034
Clerical	0.003	0.007	-0.036	0.095	0.002	0.006	-0.027	0.047	0.002	0.006	-0.023	0.039
Retail and services	0.000	0.003	-0.037	0.032	-0.000	0.002	-0.025	0.022	-0.000	0.002	-0.024	0.018
Production	0.001	0.006	-0.034	0.166	0.000	0.004	-0.032	0.030	0.000	0.004	-0.028	0.045
	(b)											
	Cumulatives, spatially smoothed											
Management and research	0.042	0.021	0.119	0.131	0.043	0.020	0.001	0.131	0.042	0.019	0.001	0.124
Clerical	0.046	0.022	0.001	0.176	0.045	0.023	0.001	0.240	0.044	0.021	0.001	0.134
Retail and services	0.036	0.017	0.002	0.100	0.036	0.017	0.000	0.183	0.034	0.016	0.001	0.129
Production	0.039	0.022	0.001	0.244	0.038	0.019	0.001	0.126	0.038	0.019	0.001	0.170
	Strength Γ, spatially smoothed											
Management and research	0.001	0.006	-0.036	0.045	0.001	0.015	-0.034	0.043	0.001	0.005	-0.027	0.033
Clerical	0.003	0.007	-0.037	0.076	0.002	0.006	-0.031	0.046	0.002	0.005	-0.028	0.040
Retail and services	0.000	0.004	-0.030	0.043	0.000	0.003	-0.026	0.023	-0.000	0.003	-0.024	0.019
Production	0.001	0.007	-0.033	0.168	0.002	0.004	-0.034	0.035	0.000	0.004	-0.031	0.084
	(c)											
	Cumulatives, smoothed and size adjusted											
Management and research	0.041	0.021	0.001	0.140	0.042	0.020	0.001	0.127	0.041	0.020	0.001	0.154
Clerical	0.044	0.022	0.001	0.160	0.043	0.022	0.001	0.145	0.042	0.021	0.001	0.144
Retail and services	0.039	0.019	0.001	0.129	0.038	0.018	0.001	0.138	0.037	0.017	0.001	0.122
Production	0.037	0.021	0.001	0.228	0.036	0.018	0.001	0.114	0.036	0.018	0.001	0.148
	Strength Γ, smoothed and size adjusted											
Management and research	0.001	0.005	-0.033	0.059	0.001	0.005	-0.039	0.043	0.001	0.005	-0.034	0.059
Clerical	0.002	0.006	-0.027	0.086	0.001	0.005	-0.033	0.054	0.001	0.005	-0.029	0.037
Retail and services	0.000	0.004	-0.041	0.032	0.000	0.003	-0.030	0.027	0.000	0.003	-0.026	0.022
Production	0.001	0.006	-0.034	0.149	0.001	0.004	-0.036	0.030	0.000	0.004	-0.027	0.067

Notes: Descriptive statistics based on 3,655 unique 4-digit industry pairs (86 4-digit manufacturing industries). For the case of the Duranton-Overman localization strength, Γ, we report all values, including for industry pairs whose colocation patterns do not significantly deviate from randomness (i.e., where Γ is zero at a distance of 25 kilometers by construction). All figures are reported for a distance threshold of 25 kilometers. Total employment is just reported in panels (A.1) and (A.2) since it does not change for spatially-smoothed or size-adjusted shares. See Appendix B.2 for details on the spatial smoothing, and Appendix B.3 for details on the size adjustments of shares.

to compute the weighted occupational shares. Weighting by postalcode frequency attributes larger weights to census divisions that are close to the plant. Panel (b) of Table 12 shows that the spatially smoothed shares are fairly similar to the unsmoothed ones.

B.3. Employment-adjusted shares

Larger firms are likely to have smaller shares of production and higher shares of non-production workers (Caliendo *et al.*, 2015). Unfortunately, we do not observe the functional employment split of plants in our data. Yet, we can devise a simple procedure that allows us to adjust shares based on plant size.[35] This procedure works as follows.

Starting from (2), we take the ratio of employment in functions f and 0 for plant z to obtain:

$$\frac{\theta_{ic}^{f}(z)}{\theta_{ic}^{0}(z)} = \frac{\theta_{ic}^{f}}{\theta_{ic}^{0}} \times z^{(\sigma-1)(\phi^{f}-\phi^{0})}. \tag{B.6}$$

Clearly, when $\phi^{f} \neq \phi^{0}$, shares will differ across plants within industry-region pairs ic, based on plant size z. We have data on $\theta_{i,c}^{f}/\theta_{i,c}^{0}$ and plant size z. However, we do not observe the plant-specific shares $\theta_{ic}^{f}(z)$ and the function-specific bundles of parameters $(\sigma-1)(\phi^{f}-\phi^{0})$. To adjust our employment shares at the plant level, we first rewrite equation (B.6) as

$$\ln\left(\theta_{ic}^{0}/\theta_{ic}^{f}\right) = (\sigma-1)(\phi^{f}-\phi^{0})\ln z + \ln\left(\theta_{ic}^{0}(z)/\theta_{ic}^{f}(z)\right). \tag{B.7}$$

Letting $\varepsilon_{ic}(z)$ denote the unobserved component, and since z is employment, we have

$$\ln\left(\theta_{ic}^{0}/\theta_{ic}^{f}\right) = \beta^{f}\ln \text{empl}_{ic}(z) + \varepsilon_{ic}(z). \tag{B.8}$$

We estimate (B.8) by OLS using fixed effects for functional types and years.[36]

The specification in Section 2 assumes that all employment types are essential, i.e., $L^{f} > 0$ for all functions f for all plants in all locations and industries. While this may hold at the firm level, it usually fails to hold at the plant level since firms can split functions across plants and locations. That all functions can be found everywhere also fails to hold geographically: θ_{ic}^{f} can be zero since no workers of function f working in industry i are located in c. To cope with 'geographical zeros' in our data (e.g., 21% of industry-location pairs have zero shares of 'Management and research' in 1996), we construct the right-hand side variables as

$$\frac{\theta_{i,c}^{0}}{\theta_{i,c}^{f}} = \frac{1+\text{empl}_{i}^{0}}{1+\text{empl}_{i}^{f}}. \tag{B.9}$$

[35]The link between firm- and plant-level functional employment is unclear. The model by Caliendo *et al.* (2015) is silent about what happens at the plant level. If all firms were single plant, then there would be no problem. Since there are multiunit firms, it is not easy and fully clear how to make adjustments at the plant level.

[36]We experimented with a several alternative specifications, including ones with industry, census division, industry-year, and census division-year fixed effects. The results are very similar. We hence stick to the specification that is closest to our model.

In (B.8), we use the spatially smoothed shares from Appendix B.2 for each plant (which reduces the number of zeros). We further include zero dummies for the different relative employment shares in our regressions. We have a total of 593,008 plant-year-function observations, and the estimation yields an R^2 of 0.578. Given the three year dummies and 86×232 industry-location dummies, this is quite high. We set production as our baseline employment 0 in the regressions, but the results are independent of that choice.

To retrieve the unobserved terms $\widehat{\varepsilon}_{ic}(z)$, note that $\widehat{\varepsilon}_{ic}(z) = \ln\big[\theta^f_{ic}(z)/\theta^0_{ic}(z)\big]$, which implies that $\widehat{\theta}^f_{ic}(z)/\widehat{\theta}^0_{ic}(z) = \exp(\widehat{\varepsilon}_{ic}(z))$. Thus, since by definition $\frac{\sum_{f>0}\widehat{\theta}^f_{ic}(z)}{\widehat{\theta}^0_{ic}(z)} = \frac{1-\widehat{\theta}^0_{ic}(z)}{\widehat{\theta}^0_{ic}(z)}$, we obtain

$$\widehat{\theta}^0_{ic}(z) = \frac{1}{1+\sum_{f>0}\exp(\widehat{\varepsilon}_{ic}(z))]} \quad \text{and} \quad \widehat{\theta}^f_{ic}(z) = \frac{\exp(\widehat{\varepsilon}_{ic}(z))}{1+\sum_{f>0}\exp(\widehat{\varepsilon}_{ic}(z))}. \tag{B.10}$$

Using the estimated coefficients $\widehat{\theta}^f_{ic}(z)$ from (B.10), we then compute the plant-specific size-ajdusted employment of function f as in (7).

Table 13 summarizes key aspects of that procedure and of our results. The top panel shows that the correlation between the plant-level spatially smoothed and the size-adjusted shares is very high. In a nutshell, controlling for plant size using our methodology does not tremendously alter the relative composition of plants. The correlations are a bit lower for clerical employment and retail and service employment, but production employment constitutes the lion's share of employment and this barely changes. Note that the correlation with the baseline functional employment shares — i.e., the raw shares from the special census tabulations — is also very high.

Table 13: Correlations and direction of the size adjustment by employment type at the plant level.

	Management and research	Clerical	Retail and services	Production
Correlations (all years):				
Baseline and smoothed employment	0.979	0.956	0.959	0.996
Smoothed and size-adjusted employment	0.908	0.867	0.827	0.920
Size-adjusted and baseline employment	0.931	0.904	0.860	0.929
Regression of smoothed on size-adjusted empl. and plant size:				
Coefficient $\widehat{\beta}_1$ of smoothed employment	0.689^a	1.030^a	1.654^a	0.533^a
Coefficient $\widehat{\beta}_2$ of plant size	0.066^a	0.111^a	0.129^a	-0.161^a
R-squared	0.892	0.887	0.890	0.912

Notes: The top panel reports the correlation between the different employment splits at the plant level (pooling our three years of data). The regressions in the bottom panel include fixed effects for functions, industries, census divisions, and years.

The bottom panel of Table 13 reports results of OLS regressions of the form

$$\text{empl}^f_{p(i,c,z),t} = \beta_0 + \beta_1 \text{empl\_sizeadj}^f_{p(i,c),t} + \beta_2 z_{p(i,c),t} + \xi_i + \gamma_c + \delta_t + \epsilon^f_{p(i,c,z)},$$

where ξ_i, γ_c, and δ_t are year-, industry-, and census division fixed effects, respectively. The results in Table 13 show that, as expected, larger plants are weighted down in terms of production employment, and weighted up in terms of management and research, clerical, and retail

and service employment. Put differently, larger plants are assigned a relatively larger share of employment in non-production categories, consistent with previous findings in the literature.

One last comment is in order. While our simple procedure produces good correlations for employment shares, it does not get the levels right (see Table 9 for descriptive statistics). Although employment in management and research is on average at the correct level, employment levels in clerical and retail and services are too high, whereas employment in production is too low. We conjecture that these problems with the levels arise because the Cobb-Douglas functional form implies that solutions must be interior: the convexity of the isosets pushes towards a solution where different functions are used in approximately same levels. Yet, the fact that the model does not get the levels rights is *not problematic for our exercise*. Indeed, the Duranton-Overman K-densities used to measure coagglomeration are distribution functions and as such insensitive to any constant employment scaling across all plants. Put differently, only relative employment for each functional type across plants matter, and not the overall scale of employment. This implies that the K-densities we estimate are not affected by the level problems.

Additional references

[1] 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.

[2] Kerr, William R., and Scott D. Kominers. 2015. "Agglomerative forces and cluster shapes." *Review of Economics and Statistics* 97**(4)**: 877–899.

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

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

Supplemental appendix material

Appendix **S.1** extends our simple framework to the case of multiunit firms and shows that no simple benchmark shares can be derived. Appendix **S.2** discusses the construction of 'compound fixed effects' and establishes conditions for their equivalence with standard industry fixed effects. Appendix **S.3** provides additional information on our special census tabulations and shows evidence for functional specialization in Canada. Appendix **S.4** presents the procedure used to construct the locational advantage benchmark distributions. Last, Appendix **S.5** contains additional tables and results for our robustness checks.

S.1. Multiple locations

Assume that firms can split their functions across locations. Let $C \subseteq \mathcal{C}$ denote the set of chosen locations, and $|C|$ be the number of locations where the firm operates. Let χ_c^f denote the share of the firm's function f in location c. Hence, C corresponds to the set of locations where $\chi_c^f > 0$. Each firm solves the cost-minimization problem: $\min_{\{\chi_c^f, \ell^f(z), \forall f \in \mathcal{F}, c \in \mathcal{C}\}} \sum_f \sum_c w_c^f \chi_c^f \ell^f(z)$ s.t. $y_{i\chi}(z) = 1$, $\chi_c^f \geq 0$, and $\sum_c \chi_c^f = 1$. We assume that the production function is given by

$$y_{i\chi}(z) = \left[|C|^{\frac{\gamma_i(\sigma_i-1)-1}{\sigma_i}} \sum_f \sum_c \alpha_{ic}^f \cdot \left(z^{\phi^f} \chi_c^f \ell^f(z) \right)^{\frac{\sigma_i-1}{\sigma_i}} \right]^{\frac{\sigma_i}{\sigma_i-1}}, \tag{S.1}$$

where $\gamma_i \geq 0$ is an industry-specific 'organizational cost' of splitting functions across locations (we could make that cost depend on the distance between locations, but this would make notation heavy without adding additional insights). If there were only benefits to splitting its activity, each firm would always split across cost-minimizing locations. If $\gamma_i = 0$, the first term in (S.1) reduces to $|C|^{-1/\sigma_i}$. In that case, the 'love-of-variety effect' in the production function that increases output as the firm splits its activities across more locations is neutralized (as in Benassy, 1996). If $\gamma_i < 0$, there are decreasing returns to splitting activities across locations. In that case, the firm will do so only if picking sites with higher α_{ic}^f / w_c^f outweigh the additional organizational costs.

Conditional on a location profile $\chi = \{\chi_c^f, f \in \mathcal{F}, c \in \mathcal{C}\}$, the firm chooses functional shares:

$$\theta_{iC}^f(z) \equiv \frac{\sum_c \chi_c^f \ell^f}{\sum_{c,f'} \chi_{c,f'} \ell^{f'}} = \frac{\sum_c \left(\alpha_{ic}^f / w_c^f \right)^{\sigma_i} (\chi_c^f)^{(\sigma_i-1)} z^{(\sigma_i-1)\phi^f}}{\sum_{c,f'} \left(\alpha_{ic}^{f'} / w_c^{f'} \right)^{\sigma_i} (\chi_c^{f'})^{(\sigma_i-1)} z^{(\sigma_i-1)\phi^{f'}}}. \tag{S.2}$$

As can be seen from (S.2), if all ϕ^f are the same, we have

$$\theta_{iC}^f(z) = \frac{\sum_c \left(\alpha_{ic}^f / w_c^f \right)^{\sigma_i} (\chi_c^f)^{(\sigma_i-1)}}{\sum_{c,f'} \left(\alpha_{ic}^{f'} / w_c^{f'} \right)^{\sigma_i} (\chi_c^{f'})^{(\sigma_i-1)}}, \tag{S.3}$$

which still depends on z via the χ_c^f terms. Hence, there is no simple way to separate the firm-specific terms from the industry-region specific terms which would allow for a simple adjustment procedure such as that in Appendix B.2.

S.2. Construction of fixed effects

As argued by Ellison *et al.* (2009, supplementary online material), including industry fixed effects in the coagglomeration regressions is not straightforward. The reason is that the co-agglomeration measures and the independent variables are symmetric, but that only 'unique industry pairs' ij are selected. The results should not depend on the ordering of the ij pairs. If there is no specific structure in the data, then the ordering of pairs should not matter. However, there is — by construction — a lot of specific structure in the industry-level data. Consider the case (as in our application) where industries are ranked in increasing order, from NAICS 3111 to NAICS 3399. Industry 3111 then appears 85 times as industry i, and never as industry j; industry 3112 appears 84 times as industry i, and only once as industry j; and so on, until industry 3399, which appears 85 times as industry j and never as industry i. This specific assignment is problematic for the inclusion of standard industry fixed effects, because the explanatory variables do not take random values with respect to this specific industry ordering. Ellison *et al.* (2009, 2010) thus propose to use 'compound' fixed effects ξ_k, where $\xi_k = 1$ if either $i = k$ or $j = k$ in coagglo_{ijt}. Actually, those fixed effects are equivalent to 'standard' industry (i and j) fixed effects, *provided that for each pair ij the pairing order is sufficiently random.*

To see this, we take our unique industry pairs — which correspond to the upper triangular part of the matrix $\{\text{coagglo}\}_{ijt}$ — and randomly switch the ij ordering for each coagglomeration pair with some probability p. In a nutshell, for each pair ij we randomly reshuffle the ordering with some probability that is the same for all pairs. We then include standard i and j fixed effects into the regression and run it 200 times.

Table 14 reports the mean coefficient estimates for the 200 replications, as well as the standard error of their distribution. As one can see, there is an upward bias for the input-output coefficient and a downward bias for the labor pooling coefficient when standard industry fixed effects are used and when the ordering within pairs is not 'random enough'. With a probability of 0.01 and 0.05 for switching the i and j indices, this bias is still there but decreases compared to the 'no shuffle' scenario. It completely disappears when the pairs are completely random (i.e., when $p = 0.5$). Thus the Ellison *et al.* (2010) fixed effects do the same job as standard fixed effects when the pairs are randomized. Either approach will work. The differences between the i and j (non-randomized) fixed effects coefficients for both input-output and labor pooling, and the compound or randomized i and j fixed effects coefficients, suggest that there is a strong structure in both the input-output and labor metrics with respect to the ordering of i and j. Indeed, as can be seen from the bottom part of Table 14, industries that feature

Table 14: Coefficient estimates using standard industry fixed effects.

'Shuffle prob.'	Replications	Input-output links	Labor similarity	Patent citations
		α_{io}	α_{oes}	α_{pat}
		Reshuffling estimations		
No shuffle		0.046	0.046	0.011
		–	–	–
0.01	200	0.046	0.049	0.011
		(0.001)	(0.002)	(0.000)
0.05	200	0.044	0.055	0.010
		(0.001)	(0.003)	(0.001)
0.50	200	0.037	0.066	0.009
		(0.001)	(0.001)	(0.000)
		'Balancing' properties		
$\#i > \#j$	5,418	0.012	0.258	0.019
$\#i \leq \#j$	5,547	0.017	0.374	0.017

Notes: We give the standard deviation of the distribution of estimated coefficients in parentheses. The 'balancing properties' panel gives the mean of the variables for industries that figure more often as i ($\#i \geq \#j$) and for industries that figure more often as j ($\#i \leq \#j$) in the dataset. All regressions and figures are computed by excluding industry pairs where one industry has less than 30 plants. Compare with column (2) in Table 2.

more often as i in the estimations ($\#i \geq \#j$, i.e., 3111, 3112 etc.) have both smaller input-output and labor pooling metrics than industries that feature more often as j ($\#i \leq \#j$, i.e., 3399, 3391 etc.). Yet, the gap is much larger for labor pooling than for the input-output metric. Since $\#i \geq \#j$ industries are overrepresented, whereas $\#i \leq \#j$ industries are underrepresent, the input-output coefficient is biased upward and the OES coefficient is biased downward.

S.3. Functional specialization in Canada

Table 15 breaks down employment types by rural and urban census divisions for Canada. As can be clearly see from that table, there is substantial functional specialization (see also Brunelle and Polèse, 2008). Indeed, in 2001 for example, column (1) shows that urban census divisions had on average 19.6% of manufacturing employment in management and research, compared to 10.4% in rural census divisions. For production, the corresponding figures are 58.2% and 76%, respectively. Note also that both clerical and retail and service employment are more strongly represented in urban areas, but the rural-urban gap is much smaller than for management and research or production. Similar patterns hold for 2003 and 2005.

Turning to columns (2) and (3), we split census divisions into rural and urban parts, based on the special census tabulations. We see that the patterns continue to hold. Column (3) is of particular interest to us. It summarizes results for census divisions that are split into rural and urban parts by the special census tabulations (118 census divisions out of 232). One might a priori expect that the rural-urban functional specialization gap could be smaller for those census divisions, since even their rural parts are 'close' to some urban core. Yet, as can be

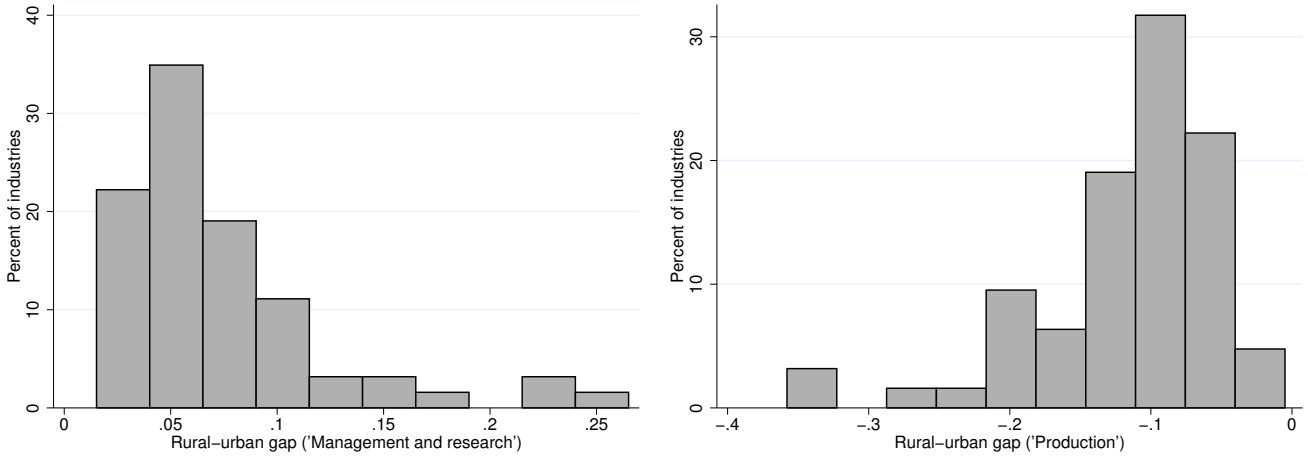
Table 15: Functional specialization of urban and rural census divisions in Canada.

		(1) All of Canada		(2) Census divisions		(3) Census divisions	
Year	Employment type	rural areas	urban areas	fully rural	fully urban	rural parts	urban parts
2001	Management and research	0.104	0.196	0.096	0.142	0.114	0.197
	Clerical	0.077	0.124	0.074	0.082	0.085	0.125
	Retail and services	0.045	0.074	0.045	0.094	0.055	0.076
	Production	0.760	0.582	0.765	0.649	0.733	0.577
2003	Management and research	0.119	0.221	0.118	0.222	0.133	0.224
	Clerical	0.071	0.096	0.064	0.061	0.077	0.096
	Retail and services	0.041	0.059	0.041	0.062	0.046	0.061
	Production	0.757	0.599	0.767	0.625	0.726	0.593
2005	Management and research	0.119	0.221	0.121	0.202	0.130	0.224
	Clerical	0.069	0.096	0.066	0.057	0.075	0.097
	Retail and services	0.043	0.060	0.041	0.063	0.048	0.061
	Production	0.757	0.598	0.761	0.636	0.731	0.592

Notes: All figures are computed from the Scott's microgeographic data, using Statistics Canada's special tabulations to split plant-level employment into individual functions. In **(1)**, we compute shares by summing employment across all rural/urban areas in Canada. In **(2)**, we report average shares of employment by types in Census divisions that are either fully urban or fully rural. Finally, **(3)** reports average shares of employment by types in Census divisions that have both a rural and an urban part. The special census tabulations allow us to separate these rural from the urban parts.

seen, the difference with columns (1) and (2) is very small. Column (3) shows that 'mixed' rural-urban census divisions have even a slightly larger rural-urban gap. This highlights the fundamental importance of splitting census divisions into rural and urban parts when using geographic variation in functional specialization patterns.

Figure 7: Distribution of rural-urban gaps in employment types (3-digit NAICS).



There is also substantial heterogeneity across sectors in functional specialization. At the level of Canada, the 3-digit sector with the largest average rural-urban gap in management and research employment is NAICS 324 ('Petroleum and Coal Products Manufacturing'), with an average rural share of 0.154 and an average urban share of 0.395, whereas the sector with the smallest average rural-urban gap is NAICS 336 ('Transportation equipment manufacturing'), with an average rural share of 0.154 and an average urban share of 0.175. For production

employment, the configuration is naturally reversed. Figure 7 depicts the distribution of rural urban gaps across 3-digit NAICS industries for the years 2001, 2003, and 2005 pooled together.

S.4. Constructing locational advantage distributions

Locational advantage due to, e.g., infrastructure or access to the sea, is potentially important for the location of firms and, therefore, industries. If two industries rely on similar locational advantage, they may end up colocated even if there are no coagglomeration economies (Ellison and Glaeser, 1999; Ellison *et al.*, 2010). We now construct benchmark distributions of industries that use a large number of locational characteristics to assign plants to locations based on those characteristics. Our approach is similar in spirit to that of Klier and McMillen (2008), yet we exploit the microgeographic nature of our data to construct our benchmark distributions.

Consider location ℓ that hosts a manufacturing plant. We define a dummy variable $y_{\ell(i)}$ that takes value one if location ℓ hosts a firm in industry i, and zero otherwise. We then run, for each industry separately, a pooled probit regression of the following type:

$$y_{\ell(i),t} = \phi\left(\mathbf{X}_{\ell,t}\beta + \gamma_{G(\ell)} + \xi_t\right) + \varepsilon_{\ell,t} \tag{S.4}$$

where $\mathbf{X}_{\ell,t}$ are location-specific variables, $\gamma_{G(\ell)}$ are geographic fixed-effects at either the province or the economic region level, and where ξ_t are year fixed effects. The error term $\varepsilon_{\ell,t}$ is assumed to have the usual properties. Note that this is not a panel regression, since each year contains a different set of locations (some locations being repeated across years). Table 16 summarizes the variables that we include into the regressions (S.4). These variables relate to: (i) the urban environment (CMA dummies, population totals); (ii) the production costs (wages, housing rents and prices); (iii) the labor composition (share of manufacturing workers, skill composition of the workforce); (iv) the international trading environment (cross-border industry links, distance to the nearest sea port, distance to nearest U.S. land border crossing); (v) major infrastructure variables (distance to highways, airports, railways); and (vi) natural resources in a stricter sens (distance from the coast, distance to major inland bodies of water, and a measure of proximity to resources like mineral deposits). We run (S.4) as a simple probit and not as a conditional probit. There are three reasons for this. First, we do not want to use plant-level characteristics (such as size or export status) because those characteristics are likely endogenous to location choices. For example, firms may start exporting if they are close to ports or airports. Second, we have only few characteristics that would be useful to include. Last, including plant characteristics requires to estimate for each of our approximately 150,000 plants the whole set of counterfactual probabilities, thus making the approach computationally difficult.

We include year (cohort) fixed effects and province fixed effects. The latter soak up a large number of provincial variations that may be important in determining the attractiveness of locations to industries. For example, electricity prices are lower in Quebec and in British

Table 16: Variables used to estimate the locational advantage probabilities.

Variable name	Type	Variable definition	Empirical implementation	Data sources
cma	Urban env.	Census metropolitan area dummy	As is	Census Geography Files
ln_pop	Urban env.	Population within 25km around the plant	log(pop. +1)	1996 and 2001 Census, DA level + GIS
ln_avg_income	Prod. costs	Average income in current C$, 25km around the plant	log(avg. income+1)	1996 and 2001 Census, DA level + GIS
ln_avg_value	Prod. costs	Average house values in current C$, 25km around the plant	log(avg. value+1)	1996 and 2001 Census, DA level + GIS
ln_avg_rent	Prod. costs	Average contract rent in current C$, 25km around the plant	log(avg. rent+1)	1996 and 2001 Census, DA level + GIS
ln_edu_share	Labor comp.	Share of highly educated within 25km around the plant	log((1+college)/(1+pop.))	1996 and 2001 Census, DA level + GIS
ln_mfg_ratio	Labor comp.	Share of mfg employment within 25km around the plant	log((1+mfg. employment)/(1+pop.))	1996 and 2001 Census, DA level + GIS
ln_naics3US_empl	Intern. trade	US employment in same 3-digit $\leq$ 800 km from plant	log(naics3US empl.+1)	2001, 2003, 2005 county business patterns
ln_dist_major_port	Intern. trade	Distance to the closest major seaport	log(dist. major port + 1)	GIS computation
ln_dist_border_crossing	Intern. trade	Great circle distance to the nearest U.S. land border crossing	log(dist. border crossing + 1)	Wikipedia page
ln_dist_major_airport	Infrastruct.	Distance to the closest top-10 freight airports in Canada	log(dist. majorairport + 1)	GIS computation
ln_dist_minor_airport	Infrastruct.	Distance to the closest minor airport	log(dist. minor airport + 1)	GIS computation
ln_dist_railway_station	Infrastruct.	Distance to the closest railway station	log(dist. railway station + 1)	National Railway Network files
ln_dist_major_road	Infrastruct.	Distance to the closest highway/expressway	log(dist. highway + 1)	United States Geological Survey (USGS)
ln_dist_railway_track	Infrastruct.	Distance to the closest railway track	log(dist. railway tracks + 1)	National Railway Network files
ln_dist_coastline	Resources	Distance to the coastline	log(dist. coastline + 1)	United States Geological Survey
ln_dist_major_water	Resources	Distance to the closest named body of water	log(dist. major water + 1)	Nat. Resources and Environment Canada
ln_nat_resource_count	Resources	Count of natural resource sites within 25km	log(natural resource count + 1)	USGS, Scotts, various

Notes: Variables used to estimate equation (S.4).

Columbia than in other provinces, thereby implying that aluminium tends to be concentrated there. British Columbia and Quebec have a hydroelectric share of more than 85%, and Quebec subsidizes the local aluminum industry via preferential pricing (Brunelle and Polèse, 2008). Ellison and Glaeser (1999) use U.S. data to control for this at the state level, but we can just include province fixed effects since we work at the plant level, so that all time-invariant tax and regulatory differences will be picked up by these variables.

Our procedure to construct the natural advantage distributions is the following:

1. We run a simple probit to estimate (S.4) using the variables summarized in Table 16. We run the regressions on an industry-by-industry basis for each industry i (86 4-digit industries), i.e., all coefficients that we estimate are industry specific.

2. We then construct for *each location* ℓ (whether chosen or not by industry i) the predicted probability that it will be chosen by a plant in industry i:

$$\widehat{y}_{\ell(i),t} = \phi\left(\mathbf{X}_{\ell,t}\widehat{\beta} + \widehat{\gamma}_{G(\ell)} + \widehat{\xi}_t\right). \tag{S.5}$$

3. To create the counterfactual distribution, we allocate the $n_{i,t}$ plants in industry i and year t to the $n_{i,t}$ sites with the largest values of $\widehat{y}_{\ell(i),t}$. This allocation is done in decreasing order of plants' age, i.e., under the assumption that older plants have choosen the best sites (we can also allocate plants randomly, and this makes little difference to the results).

4. We use the counterfactual distributions of plants thus obtained to compute the counterfactual K-densities and their associated confidence bands using 200 bootstrap replications.

The rightmost panel of Table 1 in the main text shows that about half of the 4-digit industry pairs are significantly coagglomerated based on locational advantage, whereas the other half is significantly dispersed. Surprisingly, there is very little randomness, which thus provides a strong justification to control for locational advantage in any empirical analysis.

Table 17: Cross-tabulation of observed and of locational advantage K-densities.

	2001			**2003**			**2005**		
	random	coagglomerated	codispersed	random	coagglomerated	codispersed	random	coagglomerated	codispersed
random	42	33	6	46	35	5	47	28	2
coagglomerated	548	1,164	166	670	1,051	166	732	955	166
codispersed	562	885	249	594	839	249	688	765	272

Notes: Cross tabulation of industry pairs by coagglomeration status for the 'locational advantage' distributions, and the observed distributions. The observed distribution is in columns, and the counterfactual distribution is in lines.

Table 17 cross-tabulates the observed coagglomeration patterns (in columns) with those predicted from the above procedure (in lines). As confirmed by the table, there are much less random coagglomeration patterns based on locational advantage alone: industries are either coagglomerated or codispersed, with little in between. The number of significantly

coagglomerated industry pairs is fairly similar under both the observed and the counterfactual distribution, and most of the industries that are significantly coagglomerated are so in both cases (the diagonal elements). This further confirms that locational advantage may be an important component of observed coagglomeration patterns.

Finally, the correlations between the strength of agglomeration as observed in the data and as constructed based on locational advantage only at 25km is relatively weak: it varies from 0.07 to 0.12, depending on the year. The correlations between the DO K-density cumulative distributions as observed in the data and as constructed based on locational advantage are larger, varying from about 0.24 to 0.31 depending on the year. These correlations suggest that about 1-5% of the observed coagglomeration would be explained in a univariate regression of observed patterns on our measures of locational advantage. While this is not negligible, it is also not very large. Including the industries' input and output shares with primary industries increases the explanatory power of the regression to about 30%. Yet, we doubt that "at least half of observed geographic concentration is due to natural advantages" (Ellison and Glaeser (1999, p.316). At least for Canada, about 20–30% seems more likely.

S.5. Additional tables and results

This appendix contains additional tables and results.

Table 18: Determinants of coagglomeration patterns, excluding Marshallian covariates.

	(1)	(2)	(3)	(4)	(5)	(6)	(7) (Excl3)	(8)	(9) (Excl3)
Dependent variable	**CDF of coagglomeration at 25 kilometers**								
Locational advantage (CDF, 25km)	0.087^a	0.079^a	0.077^a	0.078^a	0.077^a	0.060^a	0.070^a	0.053^a	0.060^a
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.004)	(0.005)	(0.005)	(0.006)
Multiplant share		-0.060^a	-0.036^a	-0.055^a	-0.035^a	-0.035^a	-0.034^a	-0.010	-0.009
		(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)	(0.006)	(0.006)
Within-plant coagglomeration		0.040^a	0.028^a	0.035^a	0.027^a	0.009^a	0.026^a	0.026^a	0.047^a
		(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.005)	(0.004)	(0.006)
Input share (primary)			-0.236^a		-0.217^a	-0.189^a	-0.194^a	-0.081^a	-0.084^a
			(0.014)		(0.015)	(0.014)	(0.015)	(0.011)	(0.011)
Output share (primary)			-0.086^a		-0.065^b	-0.059^c	-0.116^a	-0.052^b	-0.040
			(0.031)		(0.032)	(0.032)	(0.034)	(0.021)	(0.025)
Input share (business services)				-0.033^a	-0.014	-0.015	-0.023^b	0.014	0.011
				(0.011)	(0.011)	(0.009)	(0.010)	(0.009)	(0.010)
Output share (business services)				-0.103^a	-0.054^a	-0.039^a	-0.046^a	0.023^c	0.025^b
				(0.013)	(0.014)	(0.012)	(0.012)	(0.012)	(0.013)
Locational correlations						0.518^a	0.511^a		
						(0.017)	(0.018)		
Bonacich centrality								-0.392^a	-0.391^a
								(0.008)	(0.008)
Observations	10,292	10,292	10,292	10,292	10,292	10,292	9,729	10,292	9,729
R-squared	0.848	0.851	0.855	0.852	0.855	0.895	0.895	0.890	0.891
Dependent variable	**Strength of coagglomeration at 25 kilometers**								
Locational advantage (Strength, 25km)	0.082^a	0.078^a	0.077^a	0.078^a	0.077^a	0.074^a	0.077^a	0.055^a	0.058^a
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.013)	(0.012)	(0.013)
Multiplant share		-0.027^c	-0.023	-0.028^c	-0.023	-0.024	-0.013	-0.007	0.004
		(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.016)	(0.015)	(0.016)
Within-plant coagglomeration		0.026^b	0.028^b	0.027^b	0.028^b	0.019^c	0.053^a	0.028^b	0.061^a
		(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.017)	(0.011)	(0.017)
Input share (primary)			-0.045^c		-0.052^c	-0.038	-0.047^c	0.039	0.032
			(0.026)		(0.027)	(0.026)	(0.027)	(0.025)	(0.027)
Output share (primary)			0.219^b		0.210^b	0.214^b	0.259^b	0.224^a	0.310^a
			(0.086)		(0.088)	(0.089)	(0.106)	(0.081)	(0.101)
Input share (business services)				-0.009	-0.004	-0.003	-0.031	0.018	-0.007
				(0.023)	(0.023)	(0.023)	(0.026)	(0.023)	(0.026)
Output share (business services)				0.027	0.025	0.032	0.031	0.076^a	0.081^a
				(0.027)	(0.029)	(0.029)	(0.029)	(0.028)	(0.028)
Locational correlations						0.258^a	0.256^a		
						(0.020)	(0.021)		
Bonacich centrality								-0.267^a	-0.268^a
								(0.019)	(0.020)
Observations	10,292	10,292	10,292	10,292	10,292	10,292	9,729	10,292	9,729
R-squared	0.396	0.396	0.397	0.396	0.397	0.407	0.406	0.414	0.412

Notes: Results for all $3 \times 3,655$ unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All computations for 0-800 kilometers, with 1 kilometer steps for the K-density CDF, and 5 kilometers steps and 200 bootstrap replications for the (global) confidence bands for the strength of agglomeration. The dependent variables are computed at 25 kilometers distance in both panels. All regressions include industry and year fixed effects. In specifications **(7)** and **(9)**, we exclude all industry pairs within the same 3-digit NAICS industry. The dependent variable is the cumulative distribution of the DO K-density at 25km distance in the top panel, and the strength of coagglomeration at 25km distance in the bottom panel. See Appendix B.1 for details. Huber-White robust standard errors in parentheses.

Table 19: Coagglomeration patterns of manufacturing industries in Canada, total employment (with controls).

Dependent variable	CDF of coagglomeration								Strength of coagglomeration							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8) (IV)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16) (IV)
Input-output links	0.044^{a}	0.022^{b}	0.028^{a}	0.026^{a}	0.083^{a}	0.031^{a}	0.025^{a}	0.044^{b}	0.028^{c}	0.015	0.021	0.035^{b}	0.040^{b}	0.037^{b}	0.032^{b}	-0.004
	(0.011)	(0.010)	(0.010)	(0.006)	(0.013)	(0.007)	(0.006)	(0.018)	(0.016)	(0.017)	(0.017)	(0.015)	(0.017)	(0.015)	(0.016)	(0.026)
Labor similarity	0.127^{a}	0.038^{a}	0.046^{a}	0.031^{a}	0.077^{a}	0.048^{a}	0.024^{a}	0.018^{b}	0.082^{a}	0.053^{a}	0.057^{a}	0.015	0.063^{a}	0.009	0.007	0.018
	(0.011)	(0.009)	(0.009)	(0.007)	(0.011)	(0.007)	(0.007)	(0.009)	(0.011)	(0.011)	(0.011)	(0.013)	(0.011)	(0.014)	(0.014)	(0.016)
Patent citations	-0.015	0.004	0.001	0.003	-0.002	0.008	0.003	0.002	-0.025^{b}	-0.021^{b}	-0.023^{b}	-0.005	-0.025^{b}	-0.004	-0.005	-0.003
	(0.010)	(0.008)	(0.008)	(0.004)	(0.011)	(0.005)	(0.004)	(0.004)	(0.010)	(0.010)	(0.010)	(0.007)	(0.011)	(0.008)	(0.007)	(0.007)
Locational advantage (CDF, 25km)	0.217^{a}	0.166^{a}	0.161^{a}	0.072^{a}			0.072^{a}	0.070^{a}								
	(0.011)	(0.010)	(0.010)	(0.005)			(0.005)	(0.005)								
Locational advantage (Strength, 25km)									0.094^{a}	0.084^{a}	0.082^{a}	0.076^{a}			0.075^{a}	0.077^{a}
									(0.015)	(0.014)	(0.014)	(0.012)			(0.012)	(0.012)
Input share (primary)		-0.475^{a}	-0.418^{a}	-0.221^{a}			-0.210^{a}	-0.212^{a}		-0.154^{a}	-0.132^{a}	-0.060^{b}			-0.051^{c}	-0.047^{c}
		(0.008)	(0.008)	(0.015)			(0.015)	(0.015)		(0.010)	(0.010)	(0.027)			(0.027)	(0.027)
Output share (primary)		-0.182^{a}	-0.183^{a}	-0.071^{b}			-0.068^{b}	-0.068^{b}		-0.101^{a}	-0.097^{a}	0.202^{b}			0.209^{b}	0.210^{b}
		(0.005)	(0.005)	(0.033)			(0.033)	(0.033)		(0.009)	(0.009)	(0.090)			(0.090)	(0.088)
Input share (business services)			0.044^{a}	-0.010			-0.010	-0.007			0.057^{a}	0.001			0.001	-0.004
			(0.008)	(0.011)			(0.011)	(0.011)			(0.010)	(0.023)			(0.023)	(0.024)
Output share (business services)			0.168^{a}	-0.053^{a}			-0.052^{a}	-0.053^{a}			0.027^{b}	0.023			0.025	0.027
			(0.009)	(0.014)			(0.014)	(0.014)			(0.012)	(0.029)			(0.029)	(0.029)
Multiplant share					-0.285^{a}	-0.047^{a}	-0.031^{a}	-0.032^{a}					-0.058^{a}	-0.026^{c}	-0.022	-0.020
					(0.008)	(0.008)	(0.007)	(0.007)					(0.009)	(0.015)	(0.016)	(0.016)
Within-plant coagglomeration					0.032^{b}	0.023^{a}	0.008^{c}	0.002					0.005	0.016	0.012	0.024^{c}
					(0.014)	(0.006)	(0.005)	(0.007)					(0.013)	(0.012)	(0.012)	(0.014)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	No	No	No	Yes	No	Yes	Yes	Yes	No	No	No	Yes	No	Yes	Yes	Yes
Observations	10,292	10,292	10,292	10,292	10,292	10,292	10,292	10,292	10,292	10,292	10,292	10,292	10,292	10,292	10,292	10,292
R-squared	0.078	0.339	0.368	0.856	0.113	0.847	0.856	—	0.021	0.056	0.060	0.398	0.015	0.392	0.398	—

Notes: Results for all $3 \times 3,655$ unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All computations for 0-800 kilometers, with 1 kilometer steps for the K-density CDF, and 5 kilometers steps and 200 bootstrap replications for the (global) confidence bands for the strength of agglomeration. The dependent variables are computed at 25 kilometers distance in both panels. All regressions include industry and year fixed effects. In specifications **(7)** and **(9)**, we exclude all industry pairs within the same 3-digit NAICS industry. The dependent variable is the cumulative distribution of the DO K-density at 25km distance in the top panel, and the strength of coagglomeration at 25km distance in the bottom panel. See Appendix B.1 for details. In specifications (8) and (16), we report 2SLS results instrumenting the input-output links with their U.S. counterparts. Huber-White robust standard errors in parentheses.

Table 20: Coagglomeration patterns of manufacturing industries in Canada, by functional types (with controls).

Dependent variable	CDF of coagglomeration				Strength of coagglomeration			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Management and research	Clerical	Retail and services	Production	Management and research	Clerical	Retail and services	Production
Baseline shares								
Input-output links	0.017^{a}	0.036^{a}	0.021^{a}	0.025^{a}	0.011	0.042^{a}	0.014	0.032^{b}
	(0.006)	(0.007)	(0.006)	(0.006)	(0.011)	(0.013)	(0.011)	(0.015)
Labor similarity	-0.038^{a}	0.050^{a}	0.033^{a}	0.022^{a}	-0.058^{a}	-0.051^{a}	0.006	-0.003
	(0.009)	(0.007)	(0.008)	(0.008)	(0.016)	(0.012)	(0.014)	(0.020)
Patent citations	0.008^{c}	0.009^{b}	0.003	0.001	0.009	0.015^{c}	-0.010	-0.017^{b}
	(0.005)	(0.005)	(0.005)	(0.005)	(0.008)	(0.007)	(0.008)	(0.007)
Observations	10,292	10,292	10,292	10,292	10,292	10,292	10,292	10,292
R-squared	0.851	0.838	0.817	0.829	0.477	0.448	0.400	0.320
Smoothed shares								
Input-output links	0.019^{a}	0.037^{a}	0.021^{a}	0.025^{a}	0.016^{c}	0.054^{a}	0.010	0.023^{b}
	(0.006)	(0.007)	(0.006)	(0.006)	(0.009)	(0.012)	(0.011)	(0.012)
Labor similarity	-0.034^{a}	0.054^{a}	0.025^{a}	0.021^{a}	-0.053^{a}	-0.041^{a}	0.010	-0.005
	(0.009)	(0.007)	(0.008)	(0.008)	(0.016)	(0.012)	(0.014)	(0.019)
Patent citations	0.006	0.010^{b}	0.004	0.002	0.003	0.018^{b}	-0.015^{c}	-0.013^{c}
	(0.004)	(0.005)	(0.005)	(0.005)	(0.008)	(0.007)	(0.009)	(0.007)
Observations	10,292	10,292	10,292	10,292	10,292	10,292	10,292	10,292
R-squared	0.856	0.839	0.823	0.828	0.475	0.459	0.411	0.315
Size-adjusted shares								
Input-output links	0.021^{a}	0.034^{a}	0.025^{a}	0.027^{a}	0.024^{b}	0.049^{a}	0.014	0.022^{c}
	(0.006)	(0.006)	(0.006)	(0.007)	(0.012)	(0.014)	(0.013)	(0.012)
Labor similarity	-0.013	0.057^{a}	0.037^{a}	0.018^{b}	-0.067^{a}	-0.038^{a}	-0.006	-0.013
	(0.010)	(0.007)	(0.007)	(0.007)	(0.017)	(0.012)	(0.013)	(0.016)
Patent citations	0.004	0.007	0.004	-0.000	0.002	0.011	-0.006	-0.018^{b}
	(0.004)	(0.004)	(0.005)	(0.005)	(0.007)	(0.007)	(0.008)	(0.007)
Observations	10,292	10,292	10,292	10,292	10,292	10,292	10,292	10,292
R-squared	0.848	0.844	0.850	0.827	0.454	0.440	0.426	0.319

Notes: Results for all $3 \times 3,655$ unique industry pairs obtained from 86 NAICS 4-digit industries in 2001, 2003, and 2005. We exclude pairs where at least one industry has less than 30 plants. All variables are standardized. All computations for 0-800 kilometers, with 1 kilometer steps for the K-density CDF, and 5 kilometers steps and 200 bootstrap replications for the (global) confidence bands for the strength of agglomeration. The dependent variables are computed at 25 kilometers distance in both panels. All regressions include industry and year fixed effects, but no other controls. See Appendix B for details on how we construct the different shares. The 'Labor similarity' variable is specific to each employment type. The included controls are: Locational advantage (CDF, 25km) *or* Locational advantage (Strength, 25km); input share (primary); output share (primary); input share (business services); output share (business services); multiplant share; and within-plant coagglomeration. Huber-White robust standard errors in parentheses.

Table 21: Management and research versus production employment (robustness).

	(1)	(2)	(3) (Excl3)	(4)	(5)	(6)	(7)	(8)	(9) (IV)
Management and research employment									
Input-output links	0.001	0.017^{a}	0.028^{a}			0.018^{a}	0.015^{a}	0.013^{b}	0.028^{b}
	(0.010)	(0.006)	(0.008)			(0.006)	(0.006)	(0.006)	(0.013)
Labor similarity	0.071^{a}	-0.038^{a}	-0.007	-0.035^{a}	-0.037^{a}	-0.037^{a}			-0.040^{a}
	(0.008)	(0.009)	(0.011)	(0.009)	(0.009)	(0.009)			(0.010)
Patent citations	0.018^{c}	0.008^{c}	0.009^{c}	0.009^{c}	0.008^{c}		0.007	0.007	0.007
	(0.009)	(0.005)	(0.005)	(0.005)	(0.005)		(0.005)	(0.005)	(0.005)
Input links				0.004					
				(0.005)					
Output links					0.024^{a}				
					(0.005)				
Patent citations (make based)						0.006			
						(0.005)			
Labor movement							0.001		
							(0.006)		
Labor similarity (total employment)								0.011^{c}	
								(0.007)	
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls include	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	10,292	10,292	9,729	10,292	10,292	10,292	10,292	10,292	10,292
R-squared	0.333	0.851	0.852	0.851	0.851	0.851	0.851	0.851	—
Production employment									
Input-output links	0.028^{a}	0.025^{a}	0.046^{a}			0.026^{a}	0.026^{a}	0.024^{a}	0.047^{b}
	(0.010)	(0.006)	(0.009)			(0.006)	(0.006)	(0.006)	(0.019)
Labor similarity	0.039^{a}	0.022^{a}	0.031^{a}	0.026^{a}	0.023^{a}	0.022^{a}			0.016^{c}
	(0.010)	(0.008)	(0.007)	(0.008)	(0.008)	(0.008)			(0.010)
Patent citations	0.002	0.001	0.003	0.002	0.002		0.002	0.001	0.000
	(0.008)	(0.005)	(0.005)	(0.005)	(0.005)		(0.004)	(0.005)	(0.005)
Input links				0.011^{c}					
				(0.006)					
Output links					0.029^{a}				
					(0.006)				
Patent citations (make based)						0.001			
						(0.005)			
Labor movement							0.015^{b}		
							(0.008)		
Labor similarity (total employment)								0.028^{a}	
								(0.009)	
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls included	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	10,292	10,292	9,729	10,292	10,292	10,292	10,292	10,292	10,292
R-squared	0.366	0.829	0.832	0.829	0.829	0.829	0.829	0.829	—

Notes: See the notes of Table 6. Additionally, the following controls are included: Locational advantage (CDF, 25km); input share (primary); output share (primary); input share (business services); output share (business services); multiplant share; and within-plant coagglomeration. Huber-White robust standard errors in parentheses.

Additional references

[1] Benassy, Jean-Pascal. 1996. "Taste for variety and optimum production patterns in monopolistic competition." *Economics Letters* 52**(1)**: 41–47.

[2] Brunelle, Cédric, and Mario Polèse. 2008. "Functional specialization across space: a case study of the Canadian Electricity Industry, 1971–2001." *The Canadian Geographer* 52**(4)**: 486–504.

[3] Ellison, Glenn D., Edward L. Glaeser, and William R. Kerr. 2009. "Mathematical Appendix to 'What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns'." Available online at `https://www.aeaweb.org/aer/data/june2010/20070331_app.pdf`.

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