DISCUSSION PAPER SERIES

No. 10356

THE WORLD IS NOT YET FLAT: TRANSPORT COSTS MATTER!

Kristian Behrens, Théophile Bougna
and W. Mark Brown

INTERNATIONAL TRADE AND REGIONAL ECONOMICS

Centre for Economic Policy Research

ISSN 0265-8003

THE WORLD IS NOT YET FLAT: TRANSPORT COSTS MATTER!

Kristian Behrens, Théophile Bougna and W. Mark Brown

Discussion Paper No. 10356
January 2015
Submitted 07 January 2015

Centre for Economic Policy Research
77 Bastwick Street, London EC1V 3PZ, UK
Tel: (44 20) 7183 8801
www.cepr.org

THE WORLD IS NOT YET FLAT: TRANSPORT COSTS MATTER!

Abstract

We provide evidence for the effects of changes in transport costs, international trade exposure, and input-output linkages on the geographical concentration of Canadian manufacturing industries. Increasing transport costs, stronger import competition, and the spreading out of upstream suppliers and downstream customers are all strongly associated with declining geographical concentration of industries. The effects are large: changes in trucking rates, in import exposure, and in access to intermediate inputs explain between 20% and 60% of the observed decline in spatial concentration over the 1992–2008 period.

JEL Classification: C23, L60 and R12
Keywords: geographical concentration, input-output linkages, international trade exposure, transport costs and trucking rates

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

Théophile Bougna   bougna_lonla.theophile@uqam.ca
Université du Québec à Montréal

W. Mark Brown   Mark.Brown@statcan.gc.ca
Statistics Canada

1 Introduction

We provide evidence for the effects of changes in the costs of trading goods across space – as proxied by domestic trucking rates, international trade exposure, and customer-supplier linkages – on the geographical concentration of Canadian manufacturing industries. Using measures constructed from micro-geographic data, we find that increasing trucking rates, stronger import competition, and the spreading out of upstream suppliers and downstream customers are all strongly associated with declining geographical concentration of industries. The effects are large: holding all other variables fixed at their 1992 levels, changes in domestic trucking rates and in import exposure up to 2008 explain about 20% and 60% of the observed decline in spatial concentration, respectively. Hence, contrary to the widespread belief that the world has become 'flat' in the wake of the fall in transport, trade, and communication costs over the past two centuries, our key message is the opposite: even though the costs of trading goods across space may have hit their historical lows, changes in those costs still drive to a sizable extent changes in the economic geography of countries.[1] The world is not yet flat: transport costs matter! These results hold up to a variety of robustness checks and to instrumental variables estimations that deal with potential endogeneity concerns.

Assessing empirically the impact of transport costs on the spatial concentration of industries is important for several reasons. First, it is fair to say that, despite their fundamental theoretical role in spatial modeling, little is still known empirically on how transport costs drive the geographical concentration or dispersion of industries. Whereas many models tackle the questions of why and how spatial structure changes due to changes in the trading environment, much less is know empirically.[2] Second, assessing the direction of change in the geographical

[1]The fallacy of equating 'low' with 'unimportant' is reminiscent of the 'kaleidoscopic comparative advantage' debate in international trade: "[...] I was arguing that we now had "kaleidoscopic" comparative advantage – what we call in economic jargon, "knife-edge" specialization – so that specialization would shift among countries with small changes in cost conditions. The factors that had produced this situation were several, e.g. interest rates were less unequal across countries with integrated capital markets; technology used by multinationals located in different countries became more available across nations; the spread of technical education also meant that many in India and China read the same textbooks as Americans and Europeans; and so on. So, with kaleidoscopic (or "thin" or "knife-edge") comparative advantage in many activities, we were now confronted with volatility in, not the end of, comparative advantage." (Jagdish Bhagwati, "Why the world is not flat", 2010; available at `http://www.worldaffairsjournal.org/blog/jagdish-bhagwati/why-world-not-flat`).

[2]Even theory reaches different conclusions on the effects of changes in trade costs on the spatial structure of an economy. Krugman and Livas Elizondo (1996), Helpman (1998), and Behrens, Mion, Murata, and Südekum (2013) all find that decreasing trade costs are dispersive. However, Krugman (1991), Krugman and Venables (1995), and Fujita, Krugman, and Venables (1999) reach the opposite conclusion. Using a richer spatial structure involving two countries and four regions, Behrens, Gaigné, Ottaviano, and Thisse (2007) find that increasing international trade exposure is dispersive within countries, whereas falling domestic transport costs are agglomerative. The reasons underlying these diverging results are differences in the agglomeration and dispersion forces in the models, as well as in the modeling frameworks and the spatial structure used.

concentration of industries is important as there may be a tension between domestic policies that aim at growing clusters or at alleviating regional imbalances, and policies that aim at increasing international trade. Should trade be, for example, dispersive, pushing both domestic cluster policies and international trade agendas simultaneously may not deliver the expected results. Last, disentangling the effects of domestic shipping costs, international trade exposure, and access to both customers and suppliers on geographical concentration will also allow us to assess which components of transport costs are more likely to affect location patterns. Having an idea on this is important since all three components usually move simultaneously, thereby making assessments on the overall effects a rather complex endeavor.

Assessing empirically the impact of transport costs on the spatial concentration of industries is also a complicated task. First, we need fine measures of said spatial concentration across time to assess its changes. In this paper, we employ – for the first time to our knowledge – a long panel of continuous micro-geographic localization measures, computed from geo-coded plant-level data using the approach of Duranton and Overman (2005).[3] Using panel data allows us to go beyond existing studies that have mainly looked at the cross-sectional variation in the geographical concentration of industries. Instead, we look at the time-series variation over a nearly 20 year period to better understand what changes in covariates drive changes in the geographical concentration of industries. Dynamic analyses of agglomeration and changes therein are rare in the literature.[4] Yet, they are required if we want to control for unobserved heterogeneity and omitted variable bias in the estimations.

Secondly, we devote substantial effort to the construction of more sophisticated measures of transport costs – proxied by domestic trucking rates, international trade exposure, and input-output linkages among firms. We build trucking rates time series from the micro-data files on truck shipments within Canada. These measures capture time-changes in domestic transport costs and are invariant to the spatial structure of industry, thereby side-stepping the often endogenous nature of standard transportation measures (e.g, transportation margins from input-output accounts). Turning to trade exposure, we investigate in detail the impacts of international trade – broken down by imports and exports and by trading partners – on industry location. Last, concerning input-output linkages, we propose a novel and much more detailed micro-geographic measure than what has been used before in the literature. Loosely speaking, we construct plant-level measures that reflect the 'minimum distance' of a plant from a dollar of inputs, or the minimum distance it has to ship a dollar of outputs. Our proxies will allow us to derive more detailed evidence on the impacts of transport costs, international trade, and

[3]See Holmes and Stevens (2004) for an exhaustive survey of location patterns in North America. They do, however, not report results using continuous measures. Ellison, Glaeser, and Kerr (2010) use a 'lumpy approximation' of the Duranton and Overman (2005) measure and apply it to US manufacturing data.

[4]Dumais, Ellison, and Glaeser (2002) is one exception. They analyze the impact of entry, exit, and firm growth on the geographic distribution of manufacturing employment in the US between 1972 and 1992.

input-output linkages on the spatial structure of the economy.

Finally, as the analysis is at the industry level, we also need to deal with the possible endogeneity of our main covariates. For example, it is well documented that productivity rises as an industry concentrates geographically (see, e.g., Rosenthal and Strange, 2004; Combes and Gobillon, 2014). If the productivity gains from agglomeration are passed on to consumers and affect also trucking rates, the causality may actually run from agglomeration to transport costs and not the other way round. Furthermore, agglomeration may lead to imbalances in shipping patterns, and the latter may increase the cost of transportation due to standard logistics problems like 'backhaul' of empty trucks (e.g., Jonkeren, Demirel, van Ommeren, and Rietveld, 2009; Behrens and Picard, 2011). Turning to trade exposure, the spatial concentration of an industry may drive export partipation (via productivity gains) or may reduce import penetration (via lower prices), thus potentially biasing the estimated coefficient. To deal with endogeneity, we require some form of instrumental variables. Since we have a large number of industries and a fairly large time dimension, our setting lends itself well to the construction of internal instruments. We implement the method suggested by Lewbel (2012), which exploits heteroscedasticity and variance-covariance restrictions to obtain identification with 2SLS when some variables are endogenous and when external instruments are either weak or not available. We also follow Ellison, Glaeser, and Kerr (2010) and use US industry price indices – for the transportation sector and for manufacturing industries – to construct external instruments for the trucking rate series.

Our paper contributes to the growing literature that investigates how the geographical structure of national economies changes as trading goods – both within and across borders – becomes cheaper. Trade influences the spatial structure of economic activity via changes in market access (e.g., Redding and Sturm, 2008; Brülhart, Carrère, and Trionfetti, 2012; Brülhart, Carrère, and Robert-Nicoud, 2014), firm entry and exit (e.g., Dumais, Ellison, and Glaser, 2002; Behrens, 2014), tougher competition in product markets (e.g., D'Costa, 2010; Holmes and Stevens, 2014), infrastructure investments (e.g., Duranton and Turner, 2012; Duranton, Morrow, and Turner, 2014), cheaper access to foreign-sourced intermediates, changes in local labor market (e.g., Autor, Dorn, and Hanson, 2013; Dauth, Findeisen, and Suedekum, 2014), or any combination of these. See Brülhart (2011) for a review of the ambiguous theoretical and empirical effects of increased trade openness on the internal geography of countries.

The remainder of the paper is structured as follows. Section 2 briefly documents the evolutions of the geographical concentration of Canadian manufacturing industries. Section 3 describes our empirical strategy, constructs our key variables, and discusses the various identification issues we face. Section 4 presents our key results on the impacts of trade costs and measures related to customer and supplier access on the geographical concentration of Canadian manufacturing industries. We provide a large number of robustness checks and instrumental variables estimates. Section 5 concludes. Technical details are relegated to the appendix.

2 Trends in industrial localization from 1990 to 2009

As a prelude to the econometric analysis to follow, we first briefly describe the data and the measures of geographical concentration we use in this paper. We then provide a quick overview of the broad trends in the localization of Canadian manufacturing industries from 1990 to 2009.

2.1 Measuring localization

Our analysis is based on Statistics Canada's Annual Survey of Manufacturers (ASM) Longitudinal Microdata file from 1990 to 2009. This file contains between 32,000 and 53,000 plants per year, covering 257 NAICS 6-digit manufacturing industries. For every plant, we have information about: its primary NAICS industry; its employment; its sales; and its 6-digit postal code. The latter allows us to effectively geo-locate the plants using latitude and longitude coordinates of postal code centroids. A detailed description of the data is relegated to Appendix A.

We exploit the micro-geographic nature of our data and measure the geographical concentration of industries using the Duranton and Overman (2005, 2008; henceforth, DO) K-densities (see Appendix B for technical details). The DO K-densities look at how close plants are relative to each other by considering the kernel-smoothed distribution of bilateral distances between them. We explain in Section 3.2.1 why we use a kernel-smoothed distribution of bilateral distances and not on the raw distribution. The DO K-densities provide a very detailed micro-geographic description of location patterns, and allow for statistical testing of whether those patterns may be due to chance or not. We estimate the K-densities year-by-year for all industries at the NAICS 6-digit level. For each pair of plants, we compute the bilateral great circle distance between them using their geographical coordinates. Since the K-density is a distribution function, we can also compute its cumulative (CDF) up to some distance d. The CDF of the K-density at distance d tells us what share of plant pairs in an industry is located less than distance d from each other. Since we are not interested in identifying at which specific distances localization of firms occurs, the CDF of the K-density provides a better measure of the 'overall degree' of geographical concentration.

Table 1 summarizes the K-density CDF for the most localized industries in 1990, 1999, and 2009, respectively. To understand how to read that table, take 'Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing' (NAICS 315231) as an example. In 1990, 62 percent of the distances between plants in that industry are less than 50 kilometers. Put differently, if we draw two plants in that industry at random, the probability that these plants are less than 50 kilometers apart is 0.62. If we, however, draw two plants at random among *all* manufacturing plants, that same probability would only be about 0.08 (see Table 2 below). Clearly, this large difference suggests that the location patterns of plants in the 'Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing' industry are very

Table 1: Ten most localized NAICS 6-digit industries (based on plant counts).

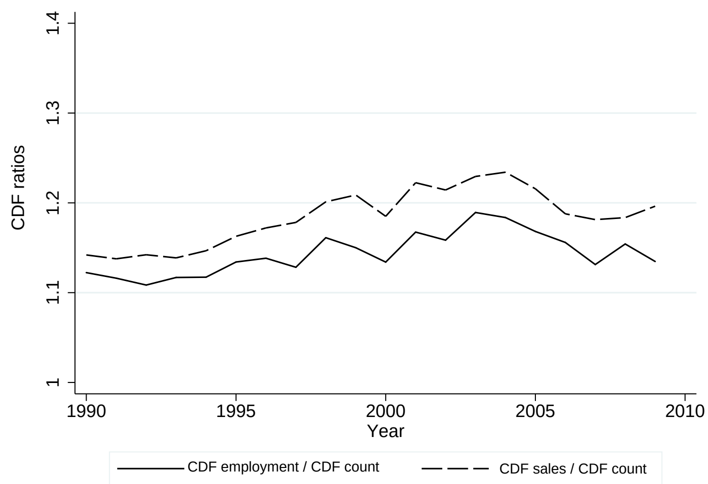
NAICS	Industry descripition	CDF
	1990	
315231	Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing	0.62
315233	Women's and Girls' Cut and Sew Dress Manufacturing	0.55
313240	Knit Fabric Mills	0.53
315292	Fur and Leather Clothing Manufacturing	0.42
315291	Infants' Cut and Sew Clothing Manufacturing	0.32
315210	Cut and Sew Clothing Contracting	0.30
337214	Office Furniture (except Wood) Manufacturing	0.21
332720	Turned Product and Screw, Nut and Bolt Manufacturing	0.21
313110	Fibre, Yarn and Thread Mills	0.19
333511	Industrial Mould Manufacturing	0.18
	1999	
315231	Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing	0.63
313240	Knit Fabric Mills	0.47
315210	Cut and Sew Clothing Contracting	0.22
333220	Rubber and Plastics Industry Machinery Manufacturing	0.20
336370	Motor Vehicle Metal Stamping	0.18
332720	Turned Product and Screw, Nut and Bolt Manufacturing	0.18
336330	Motor Vehicle Steering and Suspension Components (except Spring) Manufacturing	0.17
333519	Other Metalworking Machinery Manufacturing	0.16
337214	Office Furniture (except Wood) Manufacturing	0.15
315291	Infants' Cut and Sew Clothing Manufacturing	0.14
	2009	
315231	Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing	0.61
322299	All Other Converted Paper Product Manufacturing	0.29
337214	Office Furniture (except Wood) Manufacturing	0.17
336370	Motor Vehicle Metal Stamping	0.17
332720	Turned Product and Screw, Nut and Bolt Manufacturing	0.16
337215	Showcase, Partition, Shelving and Locker Manufacturing	0.15
321112	Shingle and Shake Mills	0.14
331420	Copper Rolling, Drawing, Extruding and Alloying	0.13
336360	Motor Vehicle Seating and Interior Trim Manufacturing	0.13
315110	Hosiery and Sock Mills	0.13

Notes: The CDF at distance d is the cumulative sum of the K-densities up to distance d. Results in this table are reported for a distance $d = 50$ kilometers.

different from those of manufacturing in general. Plants in that industry are much closer than they 'should be' if they were distributed like overall manufacturing.

Whereas the standard K-densities are computed based on plant counts, i.e., distances between pairs of plants without any weighting scheme, we can also compute weighted versions (see Duranton and Overman, 2005). In particular, we can weight pairs of plants by either plant-level employment or plant-level sales. For these weighted versions, the foregoing interpretations remain true, except that the unit of observation is now the employee or a dollar of sales. We generally report results for the weighted measures only as robustness checks, since the qualitative patterns are similar to the ones obtained from using the unweighted measures. However, comparing the unweighted to the employment- or sales-weighted K-densities reveals some interesting patterns. As can be seen from Figure 1, industries are on average

Figure 1: Year-on-year changes in the CDF ratios at 50 kilometers.



always more concentrated in terms of employment than in terms of plant counts, and even more concentrated in terms of sales than in terms of employment. This is a manifestation of agglomeration economies, and it is consistent with the findings of Holmes and Stevens (2002, 2014) and others that more localized plants tend to be larger and more productive than less localized plants. Note that the ratios are increasing until about 2004, and slightly decreasing afterwards. In 2009, within 50 kilometer distance, the concentration of employment exceeds that of plant counts by about 13%, whereas the concentration of sales exceeds that of plant counts by about 20%.

2.2 Decreasing localization

There is evidence that the geographical concentration of manufacturing industries has decreased over the first decade of the years 2000 in Canada (see Behrens and Bougna, 2013; Behrens, 2014). This de-concentration trend can clearly be seen in our data from Table 2. There has been a nearly monotonic decline in the mean value of the CDF across industries between 1990 and 2009. For example, the average CDF at 50 kilometers distance was 0.076 in 1990, 0.062 in 1999, and 0.056 in 2009, a 27.1% decrease over a twenty year period. Whereas concentration has decreased at all distances, the greatest declines, however, were at shorter distances: plants are dispersing, but less so at longer distances.[5] This finding suggests that the incentives for

[5]Whereas the CDF of the K-density is easily interpretable and provides a natural measure to track the changing concentration of industries, it cannot tell us anything about whether or not industries are statistically significantly concentrated or not. Table 9 in Appendix E summarizes location patterns by year, based on their statistical significance (see Duranton and Overman, 2005, and Appendix B for more information). As can be seen from Table 9, the share of statistically significantly localized industries has been decreasing over our study period, thus mimicking the downward trend in the K-density CDFs. In a nutshell, there is a clear trend towards less localization, and that trend is captured by both the CDF and the statistical tests for localization.

plants to locate in very close proximity to each other are lessening over time. It also likely reflects the fact that manufacturing industries have been 'bid out' of cities because of higher land and labor costs there, and that they are moving to smaller nearby urban, sub-urban, or rural areas as a consequence (see, e.g., Henderson, 1997). Still, the fact that the CDF continues to fall at 500 km suggests a broader geographic dispersion of manufacturing activity, which is likely driven by the rising manufacturing output in western Canada and the associated fundamental shifts in manufacturing location away from the 'traditional corridor' that runs through Quebec and Ontario.

Observe that the de-concentration trend also affects the employment-weighted and the sales-weighted measures of localization (see Table 2). Yet, as can be seen from Figure 2, although industries have in general become more geographically dispersed according to all three measures, the size of plant pairs in close proximity has tended to increase in relative terms regardless of whether size is measured by employment or by sales. Put differently, the process of dispersion is less pronounced when measured by either employment or sales, thus suggesting that smaller plants drive a substantial part of the dispersion process, either through entry and exit or through relocation.

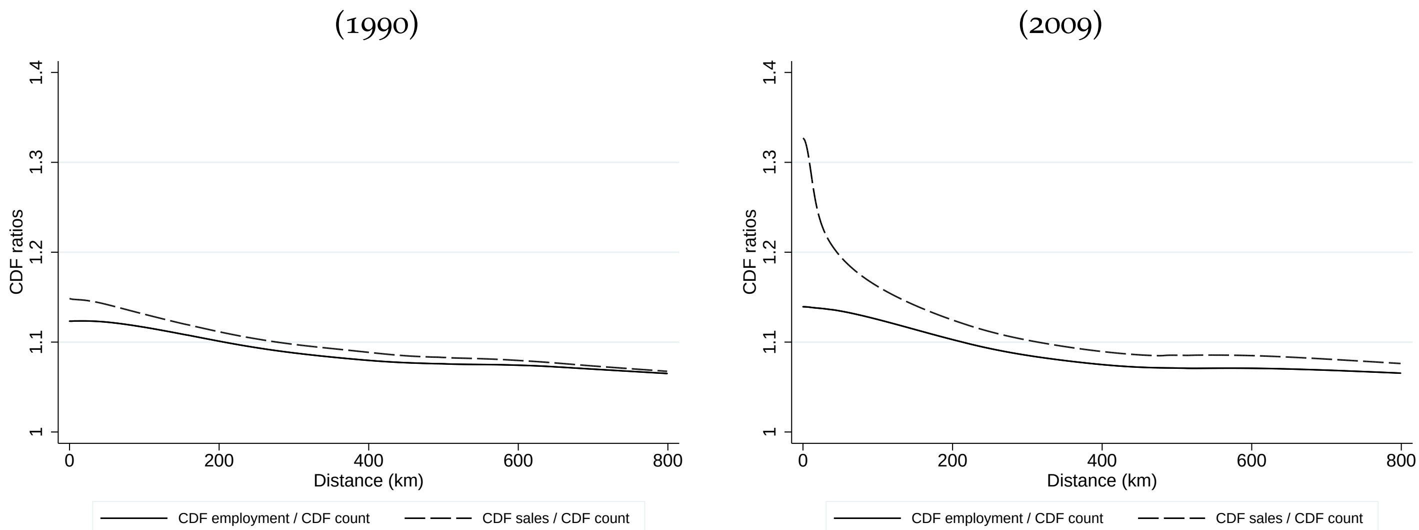
Table 2: Mean of the Duranton-Overman CDFs across industries, 1990 to 2009.

	Unweighted				Employment weighted				Sales weighted			
	CDF at a distance of											
Year	10 km	50 km	100 km	500 km	10 km	50 km	100 km	500 km	10 km	50 km	100 km	500 km
1990	0.020	0.076	0.139	0.420	0.021	0.083	0.151	0.449	0.022	0.086	0.156	0.453
1991	0.019	0.076	0.139	0.423	0.022	0.083	0.152	0.447	0.023	0.087	0.156	0.453
1992	0.020	0.074	0.135	0.418	0.020	0.079	0.147	0.442	0.022	0.084	0.151	0.448
1993	0.019	0.072	0.132	0.416	0.020	0.079	0.145	0.440	0.021	0.082	0.148	0.446
1994	0.017	0.071	0.131	0.413	0.020	0.077	0.143	0.438	0.021	0.081	0.147	0.443
1995	0.017	0.068	0.126	0.402	0.019	0.076	0.141	0.432	0.020	0.080	0.145	0.438
1996	0.016	0.065	0.122	0.402	0.019	0.073	0.136	0.428	0.020	0.076	0.140	0.435
1997	0.016	0.066	0.123	0.401	0.017	0.072	0.135	0.427	0.019	0.077	0.140	0.433
1998	0.016	0.064	0.120	0.396	0.019	0.074	0.135	0.425	0.019	0.078	0.141	0.433
1999	0.015	0.062	0.118	0.398	0.017	0.072	0.134	0.426	0.018	0.076	0.139	0.434
2000	0.014	0.063	0.120	0.383	0.016	0.073	0.135	0.411	0.016	0.075	0.140	0.421
2001	0.013	0.061	0.118	0.383	0.015	0.072	0.136	0.412	0.016	0.076	0.142	0.421
2002	0.013	0.062	0.119	0.383	0.016	0.073	0.137	0.413	0.017	0.078	0.143	0.422
2003	0.013	0.060	0.117	0.384	0.015	0.072	0.137	0.416	0.016	0.075	0.141	0.422
2004	0.013	0.060	0.115	0.379	0.015	0.070	0.132	0.412	0.017	0.074	0.137	0.418
2005	0.012	0.059	0.113	0.379	0.014	0.068	0.130	0.409	0.016	0.072	0.134	0.415
2006	0.013	0.061	0.116	0.378	0.015	0.069	0.131	0.406	0.015	0.072	0.135	0.412
2007	0.012	0.057	0.110	0.374	0.015	0.064	0.122	0.399	0.017	0.069	0.127	0.406
2008	0.012	0.057	0.110	0.376	0.017	0.067	0.125	0.400	0.017	0.069	0.128	0.405
2009	0.013	0.056	0.107	0.373	0.015	0.063	0.121	0.397	0.017	0.068	0.126	0.403
Mean	0.015	0.064	0.121	0.394	0.017	0.073	0.136	0.422	0.019	0.077	0.141	0.428
Change	-36.0%	-27.1%	-22.6%	-11.3%	-28.7%	-23.3%	-20.3%	-11.4%	-21.5%	-21.2%	-19.3%	-11.0%

Notes: Authors' computations based on the Annual Survey of Manufacturers Longitudinal Microdata file, 1990–2009. The means of the CDF are based on 257 industries and are not weighted (but the CDFs for each industry are weighted by either employment in the middle columns, or by sales in the right columns; see Appendix B). 'Mean' refers to the mean of the K-densities over the 1990–2009 period. 'Change' is the percentage change between 1990 and 2009.

To conclude, the descriptive evidence points to a significant decrease in the geographical concentration of manufacturing industries in Canada over the last 20 years, no matter whether concentration is measured in terms of plant counts, employment, or sales. The pace of decline, however, differs across industries in systematic ways. Understanding which factors drive that decrease to what extent and for which industries, with a special focus on transportation costs, trade, and input-output linkages between plants, is the objective of the remainder of this paper.

Figure 2: Ratios of mean employment- and sales-based CDFs to count-based CDF by distance.



3 Empirical methodology

While the patterns highlighted in Section 2 show that there are clear trends in changes in the geographical concentration of industries, they do not allow us to isolate the factors that drive those changes. We therefore now turn to multivariate analysis to identify the sources of those changes and to measure their relative contribution. We first briefly spell out our empirical specification. We then explain the construction of our main variables and discuss the different identification problems.

3.1 Econometric specification

We work at the industry-year level and take advantage of the panel nature of our data. More precisely, we estimate the following baseline model:

$$\gamma_{m,t}(d) = \mathbf{T}_{m,t}\beta_T + \mathbf{C}_{m,t}\beta_C + \alpha_t + \mu_m + \varepsilon_{m,t} \quad (1)$$

where $\gamma_{m,t}(d)$ is the K-density CDF for industry m in year t at distance d; where $\mathbf{T}_{m,t}$ is a vector of 'trade cost' correlates that constitute our main variables of interest; where $\mathbf{C}_{m,t}$ is a vector of time-varying industry controls; where α_t and μ_m are time and industry fixed effects, respectively; and where $\varepsilon_{m,t}$ is the error term. The latter is assumed to be independently and identically distributed with the usual properties for consistency of OLS.

One may be worried by the fact that identification in (1) comes from the *within* variation in the data. The latter may be small given yearly data, especially for the spatial variables. This point has been raised in other studies (e.g., Ellison, Glaeser, and Kerr, 2010, p.1200), but those studies ususally use more aggregated measures of agglomeration like the Ellison and Glaeser (1997) index or similar discrete indices. Those measures change much more slowly over time than the K-densities, especially at short distances. The reason is that the micro-geographic measures are constructed from geo-coded data, and that there is a lot of churning at short distances that is not picked up by spatially more aggregated measures. This churning creates a tension. One the one hand, there is *substantial year-on-year variation*, which allows for identification using this within variation. On the other hand, there is also *a lot of noise* at a small geographical scale, which makes the estimates imprecise. As we argue in Section 3.2.1 below, the K-density CDF measures provide the right tools to balance these two conflicting points.

Table 3: Key variables and summary statistics.

Variable names and descriptions	Industry detail	Mean	Standard deviation Overall	Between	Within
$\mathbf{T}_{m,t}$: *Trade, transportation, and input-output variables*					
Share of industry imports from Asian countries (excluding OECD members)	NAICS6	0.12	0.23	0.17	0.06
Share of import s from OECD member countries (excluding U.S. and Mexico)	NAICS6	0.16	0.18	0.13	0.05
Share of impors from NAFTA countries (U.S. and Mexico)	NAICS6	0.66	0.33	0.26	0.07
Share of industry exports from Asian countries (excluding OECD members)	NAICS6	0.03	0.08	0.05	0.03
Share of export from OECD member countries (excluding U.S. and Mexico)	NAICS6	0.09	0.13	0.08	0.05
Share of exports from NAFTA countries (U.S. and Mexico)	NAICS6	0.83	0.26	0.19	0.07
Ad valorem trucking costs for an avg. load shipped 500km as a share of goods shipped	L-level	0.034	0.035	0.030	0.005
Industry mean of the avg. distance to a dollar of inputs from the 5 nearest plants (km)	NAICS6	242.99	152.33	95.94	56.39
Industry mean of the avg. distance to ship a dollar of output to the 5 nearest plants (km)	NAICS6	244.86	171.87	104.36	67.51
Minimum average distance to 5×257 closest plants	NAICS6	64.54	56.63	42.44	14.19
$\mathbf{C}_{m,t}$: *Industry-year control variables*					
Share of input from natural resource-based industries	L-level	0.11	0.2	0.17	0.03
Sectoral energy inputs as a share of total sector output L-level	L-level	0.03	0.057	0.044	0.013
Total industry employment	NAICS6	6938	9749.88	7744.11	2005.76
Herfindahl index of enterprise-level employment concentration	NAICS6	0.1	0.126	0.092	0.034
Mean plant size	NAICS6	74	181	139	42
Share of plants controlled by multi-plant firms	NAICS6	0.21	0.248	0.183	0.065
Share of foreign controlled plants	NAICS6	0.15	0.2	0.14	0.06
Share of hours worked by all workers with post-secondary education	NAICS6	0.4	0.115	0.07	0.045
Intramural research and development expenditures as a share of industry sales	L-level	0.0111	0.039	0.027	0.012

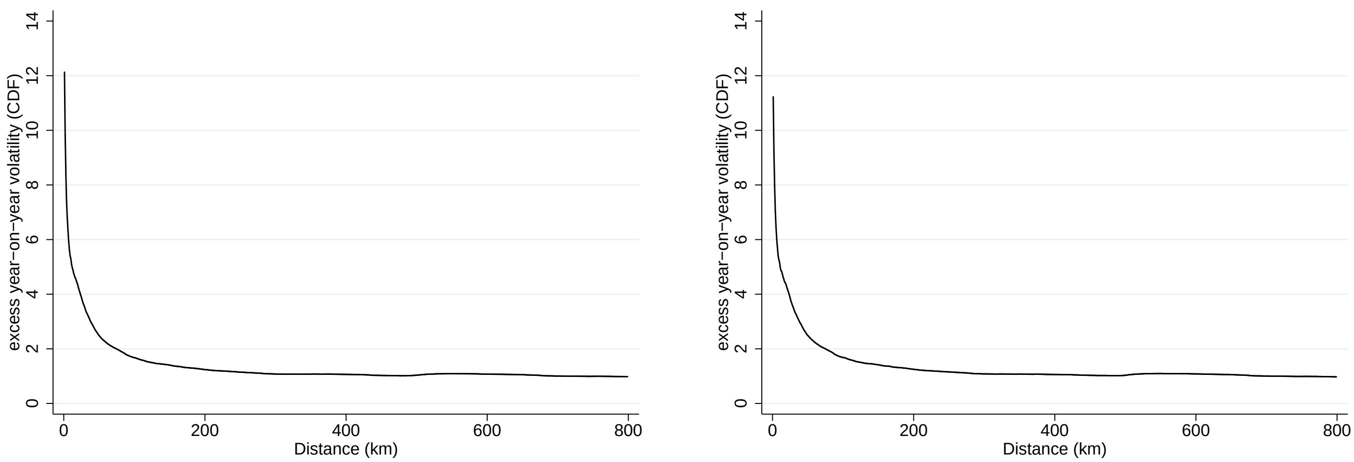
Notes: All descriptive statistics are based on the sample we use in the regression analysis, which includes 4,369 observations covering 257 industries and 17 years. The standard deviation is decomposed into between and within components, which measure the cross sectional and the time series variation, respectively. Some industry-level data are available at the L-level only, which is the finest level of data for public release in Canada (between the NAICS 3- and 4-digit levels of aggregation). Additional information regarding our data sources and the construction of our key variables is provided in Appendix A and in Section 3.2.

Table 3 summarizes our main variables, provides descriptive statistics, and reports the within and between components of the variance. As can be seen, there is substantial time variation in our data, although the bulk of the variation remains cross-sectional, as expected.

3.2 Construction of the key variables

We now describe in detail the construction of our key variables: (i) our K-density geographical concentration measures; (ii) our industry measures of transportation costs; (iii) our micro-geographic input-output linkages; and (iv) our measures of industries' international trade exposure. We also discuss a number of methodological issues related to their construction.

Figure 3: 'Excess volatility' of the raw CDFs, linear trend (left) and autoregressive (right).



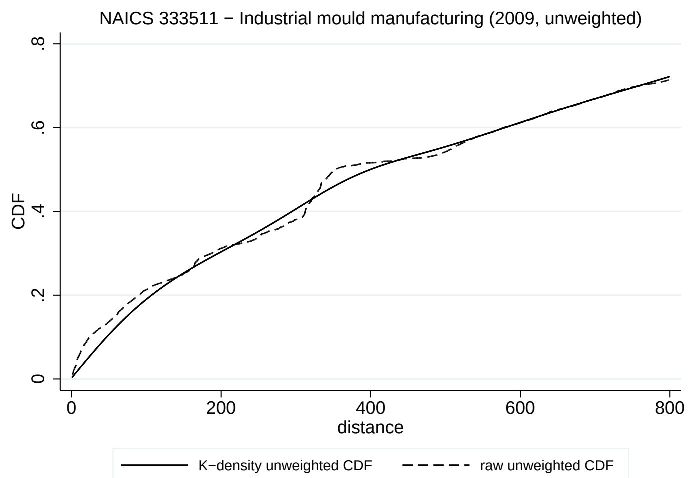
3.2.1 K-density CDFs

The technical details concerning the construction of the K-density CDFs are given in Section 2.1 and in Appendix B. Here, we discuss a number of issues linked to the time variability and the smoothing that we mentioned above. Starting with the former point, Figure 3 depicts the year-on-year 'excess volatility' at each distance d between 1 kilometer and 800 kilometers. The excess volatility is defined as the ratio of the year-on-year volatility of the raw distribution and that of the kernel-smoothed distribution.[6] As can be seen from Figure 3, the raw distribution is always more volatile than the smoothed distribution, and *especially so at short distances*. Whereas for distances greater than about 200 kilometers the volatility of the raw and the smoothed CDFs are roughly identical, the raw distribution is up to 11 or 12 times more volatile at short distances. In other words, due to substantial churning at the plant level, the micro-geographic measures contain a lot of noise in the time-series at short distances, though it is at those distances that

[6]See Appendix B for the formal definition of the 'raw' distribution. We use standard measures of volatility based on the year-on-year variance, the fitting of a linear trend, or an autoregressive AR(1) model.

the effects of transport costs and trade that we intend to identify are most likely to operate. Thus, smoothing is important to reduce the noise in the time series.[7]

Figure 4: Example of raw vs kernel-smoothed CDFs for plant counts.



Smoothing has, however, the drawback to alter the raw distribution. Figure 4 depicts the 'raw' (unsmoothed) CDF of the bilateral distances as a dashed line, and the K-density CDF (smoothed) as a solid line for a representative industry – 'Industrial mould manufacturing'. Two comments are in order. First, as can be seen, the smoothed CDFs are less volatile and more regular than the unsmoothed CDFs, though the two become very similar at longer distances starting at about 200 kilometers. As can also be seen from Figure 4, the smoothed CDFs tend to underestimate the degree of geographical concentration at short distances. This point has been recently made by Murata, Nakajima, and Tamura (2014), who show that there is a downward bias in the Duranton-Overman K-density estimates at short distances due to 'reflection' and the use of a differentiable kernel function.

To summarize, there are costs and benefits of using the smoothed CDFs compared to the unsmoothed CDFs. The benefit is that the smoothed densities exhibit substantially less year-on-year variability at short distances, thus reducing the noise due to plant-level churning that shows up in the data and that affects the micro-greographic concentration measures. The cost is that the smoothed densities underestimate the degree of geographical concentration at short distances, thus potentially biasing the estimated coefficients on the trade cost covariates towards zero. Since identification stems from the time-series variation in our approach, we believe that the benefits of using the smoothed CDFs outweight the costs.[8]

[7] We ran our analysis using the raw CDFs as dependent variables, but the results for short distances become very imprecise. Most coefficients are not statistically significant due to their large standard errors.

[8] In a cross-sectional analysis, we would rather use the raw CDFs since there is no need to smooth out any time-series volatility. However, Duranton and Overman (2005) argue that even in a cross section smoothing may be required to cope with unobserved variation in, e.g., the density of the road network.

3.2.2 Transportation costs

Transportation costs loom large in the theoretical literature on industry location and geographical concentration. Industries with high transportation costs – either for their inputs, for their outputs, or for both – should agglomerate production in locations close to their suppliers or customers to minimize those costs. Despite their dominant theoretical role, it is fair to say that limited work has gone thus far into the elaboration of good measures of transportation costs, and even less into their application to the analysis of changes in agglomeration. Rosenthal and Strange (2001), for example, use the ratio of inventories to sales at the end of the year as a proxy for 'perishability of output', itself a proxy for transportation costs. Lu and Tao (2009) use a similar proxy, namely the finished goods to output ratio, where finished goods are inventories not yet sold. Ellison, Glaeser, and Kerr (2010) do not even talk about the possible role of transportation costs in their analysis, the reason being that these costs are assumed to have become 'negligible'. While this may be the case in a cross-section of industries – with transport costs on average around 3–4% of the value of the shipment according to our estimates – our results show that their time-series variation is a major driver of the changes in the location patterns of industries. In other words, transport costs matter!

Our work aims to improve our understanding of how changes in transportation costs influence changes in the geographical concentration of industries. To this end, we use *direct measures* of transportation costs constructed from detailed micro-data files on shipments within Canada. To estimate ad valorem rates, we first use a pricing model to predicted trucking firm revenues for a 500 kilometers trip by commodity for the average tonnage using shipment (waybill) data from Statistics Canada's Trucking Commodity Origin-Destination Survey (see Brown and Anderson, 2015, for details). We estimate the 'prices' charged by trucking firms as a function of distance shipped, tonnage, and a set of commodity and firm fixed effects.[9] The prices are then converted into ad valorem trucking costs by estimating the value of each shipment. This value is derived by multiplying the tonnage of the average shipment on a commodity basis by their respective value per tonne derived from an 'experiment export trade file' produced only in 2008. The ad valorem estimates at the commodity level in 2008, in turn, are used to estimate ad valorem rates $\tau_{m,2008}$ for L-level industries in 2008 using a set of industry-commodity concordances. Yearly trucking industry price indices $p_{\text{trans},t}$ and manufacturing industry price indices $p_{m,t}$ from Statistics Canada's KLEMS database are then used to project the ad valorem rates backwards and forwards in time, thereby creating an industry-specific ad valorem transportation rate time series $\tau_{m,t}$:

$$\tau_{m,t} = \frac{p_{\text{trans},t}}{p_{m,t}} \tau_{m,2008}. \tag{2}$$

Although our measures of transport costs are much more direct and detailed than those

[9]While we do not directly control for the time costs of transportation they will be, at least partially, embedded in the transportation prices (which would capture quality of service for time-dependent trips).

used before in the agglomeration literature, they are by construction unlikely to be fully exogenous to industrial location patterns since they depend on price indices. We come back to this point in Section 3.3 below when we discuss the different identification issues. Note, however, that we estimate transportation costs for a 'representative shipment' by truck, holding distance fixed at 500 kilometers. Hence, variable shipping distances that result from optimal location choices of plants in an industry have a priori no direct influence on our measures.

Figure 5: Changes in average transportation costs, 1990–2009.

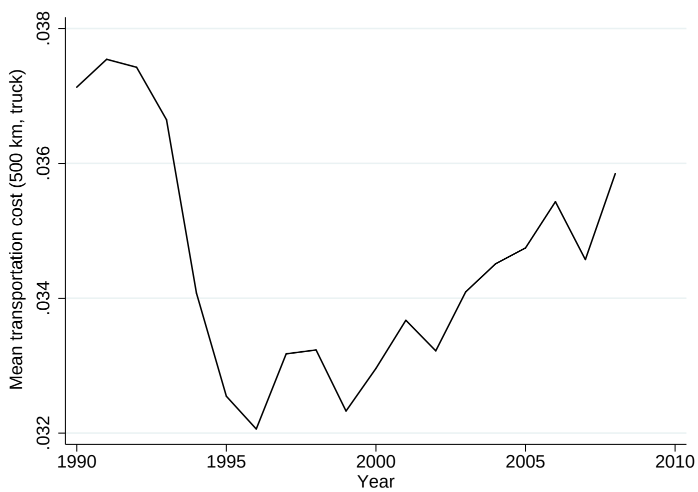


Figure 5 depicts the year-on-year changes in the (unweighted) cross-industry average transportation costs for a 500 kilometers shipment. As can be seen, transport costs are first decreasing – due, essentially, to reductions in labor costs at constant fuel prices – and then increasing – due, essentially, to increasing fuel prices at constant labor costs. They range from about 3.8% of the value of the shipment in the early nineties, to about 3.2% in the mid-nineties. Since industries tend to localize when their shipping costs are either high (market access) or low (to exploit other sources of agglomeration economies), we expect transportation costs to have a non-linear and negative effect on the degree of industrial agglomeration, especially for industries characterized by intermediate values of transport costs. Since there is significant time- and cross-industry variation in transportation costs in our data (see Table 3), we will be able to estimate precisely the effect of transportation costs on the geographical patterns of industries.

3.2.3 International trade exposure

While transportation costs capture the 'domestic' part of trade in our model, we also control finely for the role of international trade in the location of industries. It is indeed well known theoretically – though less so empirically – that trade influences the spatial structure of economic activity via firm entry and exit, tougher competition in product markets, cheaper access to foreign-sourced intermediates, and changes in local labor markets (e.g., D'Costa, 2010; Brülhart, Carrère, and Trionfetti, 2012; Autor, Dorn, and Hanson, 2013; Behrens, 2014; Brülhart, Carrère, and Robert-Nicoud, 2014; Holmes and Stevens, 2014). We use detailed yearly data on imports and exports by industry and country of origin and destination to control for industries' import and export exposure (the ratio of industry imports or exports to industry sales). To disentangle the different effects that depend on whether trade is in intermediates or final goods (on which we have unfortunately no information in our data), and on whether trade is 'North-North' or 'North-South', we break these measures down by countries of origin: low-cost Asian countries; OECD countries; and NAFTA countries.

Figure 6: Changes in import- and export trade values (left), and import shares (right).

The left panel of Figure 6 depicts the changes in the average import and export values by industry over our study period. The right panel provides a snapshot of how import and export shares change across broad groups of trading partners. As one can see, the importance of international trade has dramatically increased – at least up to the trade collapse starting 2008 – and there has been a progressively increasing re-orientation of trade towards Asian countries (especially for imports).

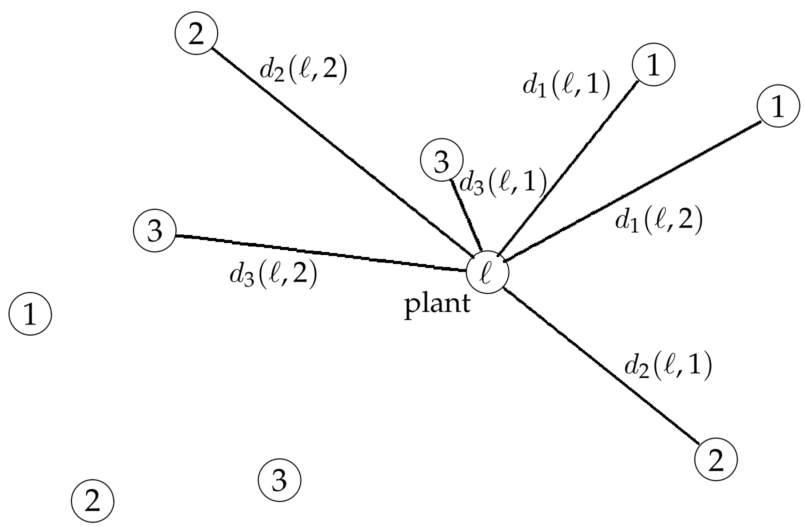
3.2.4 Input-output linkages

Another important trade-related source of agglomeration are input and output linkages. Many studies find that customer-supplier relationships is the most important mechanism to explain the co-location of industries, which is suggestive of their importance for geographical concentration.[10] Despite their importance, the empirical treatment of input-output linkages has

[10]Holmes (1999) documents that plants in US manufacturing industries that are geographically more concentrated are more vertically disintegrated. Their purchased inputs as a percent of the value of outputs is higher in areas where the industry concentrates, thus suggesting that input-output linkages may drive industry localization. Note, however, that he cannot rule out reverse causality: plants in industries that concentrate geographically

been rather limited until now. Rosenthal and Strange (2001) use manufacturing and non-manufacturing inputs purchased by the industry per dollar of output. Lu and Tao (2009) use the export-intensity of a sector as a proxy for input sharing.[11] Another approach to modelling input sharing – the most widely adopted in the literature – is to use input-output accounts to measure the extent that industries buy and sell from one another (e.g., Duranton and Overman, 2005, 2008; Ellison, Glaeser, and Kerr, 2010). The drawbacks of all these approaches is that the input-output measure is potentially endogenous, and that it does not take into account any geographical information.

Figure 7: Constructing input-output distances and 'minimum distance' measures.



Our measures of input and output linkages are very different and make use of the micro-geographic nature of our data. Consider a plant ℓ active in sector $\Omega(\ell)$. Let Ω denote the set of sectors and Ω_s the set of plants in sector s. Let $k_s(i,\ell)$ denote the ith closest sector-s plant to plant ℓ. Our micro-geographic measures of input- and output linkages are constructed as weighted averages as follows:

$$\mathcal{I}\text{dist}(\ell) = \sum_{s\in\Omega\setminus\Omega(\ell)} \omega^{\text{in}}_{\Omega(\ell),s} \times \frac{1}{N}\sum_{i=1}^{N} d(\ell, k_s(i,\ell)), \tag{3}$$

for inputs, and

$$\mathcal{O}\text{dist}(\ell) = \sum_{s\in\Omega\setminus\Omega(\ell)} \omega^{\text{out}}_{\Omega(\ell),s} \times \frac{1}{N}\sum_{i=1}^{N} d(\ell, k_s(i,\ell)), \tag{4}$$

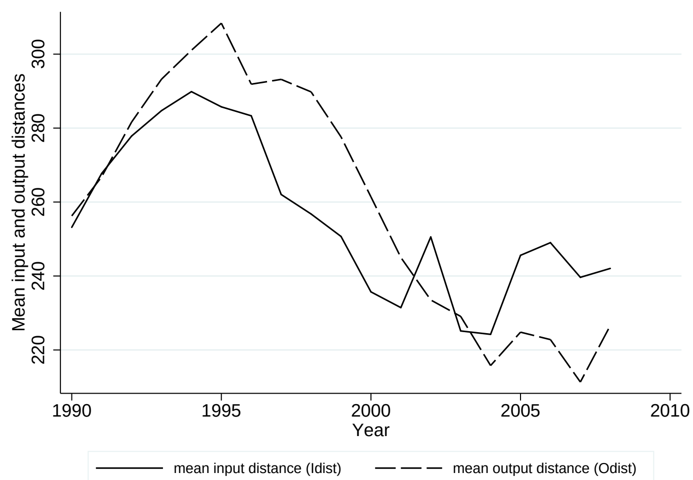
for some unobserved reason may vertically disintegrate more because of that concentration.

[11]The rationale for this proxy is that, when compared to other industries, export industries strongly rely on inputs and information sharing like the information on procedures and international markets where they sell their products. This measure thus cannot disentangle information externalities from input sharing.

for outputs, where $d(\cdot,\cdot)$ is the great circle distance between the plants' postal code centroids, and where $\omega^{\text{in}}_{\Omega(\ell),s}$ and $\omega^{\text{out}}_{\Omega(\ell),s}$ are sectoral input- and output shares.[12] Figure 7 illustrates the construction for the case where $N = 2$ and with three industries.

Since by construction $\sum_s \omega^{\text{in}}_{\Omega(\ell),s} = \sum_s \omega^{\text{out}}_{\Omega(\ell),s} = 1$, we can interpret $\mathcal{I}\text{dist}(\ell)$ as the minimum average distance of plant ℓ to a dollar of inputs from its N closest suppliers. Analogously, $\mathcal{O}\text{dist}(\ell)$ is the minimum average distance plant ℓ has to ship a dollar of outputs to its N closest (industrial) customers.[13] The larger are $\mathcal{I}\text{dist}(\ell)$ or $\mathcal{O}\text{dist}(\ell)$, the worse are plant ℓ's input or output linkages – it is, on average, further away from a dollar of intermediate inputs or a dollar of demand emanating from the other industries.

Figure 8: Changes in average input-output distances, 1990–2009.



Note that our input and output linkages make use of plant-level location information, but only of *national* input and output shares. The latter is due to the fact that we do not directly observe input-output linkages at the plant level. Yet, given this, our procedure has the advantage to sidestep problems of endogeneity of those measures. Note also that our input-output measures are computed across all industries except the one the plant belongs to. Thus, our measures capture finely the whole cross-industry location patterns, but do not pick up industrial localization of the sector itself since it is excluded from the computation. This is important to not confound input-output linkages with other drivers of geographical concentration.

We compute the measures (3) and (4) for all years and for all plants, using the $N = 3, 5, 7, 10$ nearest plants in each industry. We then average them across plants in each industry and each

[12] Appendix C provides additional details on the input and output shares.

[13] Unfortunately, we have no micro-geographic information on final demand and thus cannot include it in our output linkage measures. Using a population-weighted market potential measure as a proxy is infeasible because of the very strong persistence in time. However, our industry fixed effects are likely to control for slow-changing final demand due to changes in the population distribution.

year to get an industry-year specific measure of both input and output distances:

$$\mathcal{O}\text{dist}_s = \frac{1}{|\Omega_s|}\sum_{\ell=1}^{|\Omega_s|}\mathcal{O}\text{dist}(\ell) \quad \text{and} \quad \mathcal{I}\text{dist}_s = \frac{1}{|\Omega_s|}\sum_{\ell=1}^{|\Omega_s|}\mathcal{I}\text{dist}(\ell), \tag{5}$$

where $|\Omega_s|$ denotes the number of plants in industry s. As expected, these measures are strongly correlated. Yet, despite that correlation we can include them simultaneously into our regressions and still identify their effect on industrial localization.

Figure 8 depicts the time-series changes in the (unweighted) average input and output measure across all industries. As one can see, in 2000 for example, plants were on average located about 235 kilometers from a dollar of inputs, and had to ship a dollar of their output on average over a distance of 260 kilometers.[14]

One potential problem with the measures (3) and (4) is that they tend to be mechanically smaller in denser areas. To control for this fact, we also compute a 'minimum distance measure', i.e., the distance of plant ℓ from the $M = N \times 257$ closest plants regardless of their industry. Including that measure into our regressions then controls for the overall plant density in a location, which implies that our input-output linkage measures pick up the effect of being closer to a dollar of inputs or outputs conditional on the overall density of the area the plant is located in. Formally, we compute for each plant ℓ the following measure:

$$\mathcal{M}\text{dist}(\ell) = \frac{1}{M}\sum_{i=1}^{M} d(\ell, k_{\setminus\Omega(\ell)}(i,\ell)), \tag{6}$$

where $d(\ell, k_{\setminus\Omega(\ell)}(i,\ell))$ denotes the distance to the ith closest plant in any industry but $\Omega(\ell)$. We then average this measure across all plants in the same industry as before.

3.2.5 Industry-level controls

The literature on industrial localization has identified many important sources of externalities that cause the spatial concentration of industries and changes therein (see Duranton and Puga, 2004, for a review). Knowledge spillovers and labor market pooling are among the most important 'Marshallian' factors, but various other structural characteristics like industry size, an industry's dependence on raw materials, the presence of multi-unit firms, or foreign ownership also affect their spatial structure.

In the subsequent analysis, we control for these confounding time-varying agglomeration factors as follows. First, we control for knowledge spillovers using as a proxy an industry's research and development (R&D) intensity, i.e., the ratio of R&D expenditure to total output of

[14]Time-series changes in the input- and output-distance measures may reflect three things: (i) entry or exit of potential suppliers; (ii) changes in the geographical location of input suppliers and/or clients; and (iii) changes in the input-output coefficients, i.e., the technological relationships. We cannot dissociate the sources (i) and (ii) in our analysis, but entry and exit are vastly more important than relocation when looking at plant-level data.

that industry. By their very nature, knowledge spillovers are very hard to measure directly. The literature has often proxied them using patent citation data, i.e., patents originating in industry i that are cited by patents of industry j. While useful in a cross-sectional context, our twenty year panel does not allow us to exploit patent citation data. Second, along with knowledge spillovers, labor market pooling is another important source of agglomeration. To construct good proxies for labor market pooling, it is important to identify industry characteristics that are related to the specialization of the industry's labor force (see Rosenthal and Strange, 2001; and Lu and Tao, 2009). The literature suggests that agglomeration occurs because workers are able to move across firms and industries, thus improving the average quality of firm-worker matches. Furthermore, idiosyncratic productivity shocks at the firm level can be better hedged in locations where firms using similar workers concentrate. Firms also agglomerate to take advantage of scale economies associated with a large labor pool that allows industries to use the same type of workers. Since it is difficult to identify these characteristics, we employ a proxy related to workers' occupations. More specifically, we use the share of hours worked by all workers with post-secondary education in the total number of hours worked.[15]

We finally construct numerous time-varying controls that proxy for the remaining agglomeration factors in our econometric analysis. We firstly control for the importance of natural advantage in the agglomeration process. The importance of doing so has been pointed out, among others, by Kim (1995) and by Ellison and Glaeser (1999). We use the share of inputs from natural resource-based industries, and the sectoral energy inputs as a share of total sector output, as proxies for natural advantage. We secondly control for basic industry structure and scale effects by including the following controls: total industry employment; mean plant size; the Herfindahl index of firm-level concentration (employment based);[16] the share of plants controlled by multi-unit firms; and the share of plants controlled by foreign firms (see Table 3). These controls proxy for sectoral differences in the size distribution of firms and plants, for potential differences in the location patterns of multinationals and multi-unit firms, as well as for differences in 'business culture' (Rosenthal and Strange, 2003).

Note that all these controls are time varying and industry specific. When combined with both time and industry fixed effects, they will control for a wide range of factors that may drive changes in the degree of geographical concentration of industries that are unrelated to changes in transportation costs, trade, or input and output linkages. This will provide better identification. We now discuss remaining identification issues.

[15]We also tried to construct proxies for labor market conditions using the non-production to production worker ratio and others educational characteristics of the workforce. The latter are available at a more aggregated industry level (L-level) from Statistics Canada's KLEMS database (e.g., the share of hours worked by all workers with a university degree, and the labor productivity index). These measures, however, proved to not give significant results in the time series because they change quite slowly over time.

[16]Estimates using a Herfindahl index of plant-level concentration are qualitatively similar.

3.3 Identification issues

The three main problems that plague the identification of agglomeration effects are unobserved heterogeneity, omitted variable bias, and simultaneity bias. All studies based on cross-sectional data at the industry level (e.g., Rosenthal and Strange, 2001; Ellison, Glaeser, and Kerr, 2010) are potentially prone to these identification problems and use different strategies to overcome them. The panel nature of our data allows us to control for industry-specific time-invariant factors and general macroeconomic trends. Furthermore, the inclusion of a large set of time-varying industry controls for natural advantage, industry structure, ownership structure, and proxies for labor demand conditions and knowledge spillovers (see Section 3.2.5) substantially reduces the risk of omitted variable bias when estimating our key coefficients β_T for the trade cost correlates. However, neither the panel structure nor the controls will help with potential problems of reverse causality. These may affect our three variables of interest, namely transportation costs, trade exposure, and input-output linkages.

Transportation costs. It is well documented that productivity rises as an industry concentrates geographically (see, e.g., Rosenthal and Strange, 2004; Combes and Gobillon, 2014). Because our measure of transportation costs is computed on an ad valorem basis and includes the industry price index, the causality may run from agglomeration to lower prices and, therefore, lower ad valorem transportation costs. At the same time, agglomeration may lead to imbalances in shipping patterns, and the latter may increase the cost of transportation due to standard logistics problems like 'backhaul' of empty trucks (e.g., Jonkeren, Demirel, van Ommeren, and Rietveld, 2009; Behrens and Picard, 2011). Agglomeration would thus increase the transportation price index and affect our estimates. In a nutshell, $p_{\text{trans},t}/p_{m,t}$ in expression (2) is likely to be endogenous to the degree of geographical concentration of an industry, with stronger concentration increasing that ratio due to a combination of rising freight prices and lower output prices. Thus, the estimated OLS coefficient for transportation costs is likely to be upward biased in our model.[17]

To deal with that problem, we adopt three different strategies. First, we clear out the effect of productivity growth on prices (the presumed source of endogeneity) by regressing our transportation cost series on industry multi-factor productivity indices (from the KLEMS database), as well as industry and year fixed effects. We then use the residual from that regression as a proxy for the transportation cost series. By definition, that residual is orthogonal to any productivity driven price changes that could stem from the changing geographic concentration of industries. This strategy does, however, not deal directly with the transportation price index.

[17] Industries that agglomerate are also likely to ship their output over different distances than industries that are less concentrated because of their location choices. This problem does not affect our estimates since our measure of transportation costs is constructed for a representative shipment over a fixed distance of 500 kilometers.

Second, as we have a large number of industries and a fairly large time dimension, our setting lends itself well to the construction of internal instruments. We implement the method suggested by Lewbel (2012), which exploits heteroscedasticity and variance-covariance restrictions to obtain identification with 2SLS when some variables are endogenous and when external instruments are either weak or not available.

Third, we use US manufacturing industry price indices as external instruments for the transportation cost series. The instrumentation strategy is similar to that of Ellison, Glaeser, and Kerr (2010), who instrument the US input-output matrix and the US industry labor requirements with those of the UK. The underlying idea is the following. Assume that the geographical concentration of an industry increases over time because of unobserved factors that we cannot control for in our analysis. The increasing geographical concentration then raises ad valorem transportation costs via price decreases of the industry's output. Provided that the changes for the US are not driven by the same unobserved factors that affect the spatial concentration of the industry in Canada, but that the US price series are correlated with the changes in $p_{\text{trans},t}/p_{m,t}$, they will provide valid instruments for the Canadian transportation cost series. Two potential limitations of these instruments are the following: (i) there may be common underlying unobserved factors that drive changes in the concentration of the same industries in Canada and the US; or (ii) the geographical concentration of an industry in Canada affects directly the productivity – and, therefore, the price indices – in the US. While we cannot completely rule out those possibilities, neither strikes us as extremely plausible. First, the panel nature and the extensive set of time-varying controls should pick up most of the unobserved factors that may drive the increasing concentration of the industry; and second, the Canadian economy is small compared to the US economy, so that changes in the degree of concentration in Canada are very unlikely to have substantial productivity impacts in the US.[18]

Trade exposure. As argued above, the geographical concentration of plants increases productivity and, therefore, may increase the propensity of an industry to export and to import. For example, the agglomeration of an industry may reduce prices, which makes import penetration harder. In that case, the dispersion of an industry may be associated with increasing imports since productivity falls. Also, the agglomeration of an industry may be associated with rising exports due to productivity gains – although the productivity gains reduce unit export values, the total value of exports may increase. We deal with the potential endogeneity of trade flows using the Lewbel (2012) estimator with internal instruments.

[18]The empirical elasticity of productivity to the density or size of economic activity is usually in the 3–8 percent range, and thus huge changes in the geographical structure would be required to obtain large productivity changes. Furthermore, empirical work has documented that the effects of shocks to Canadian productivity have very limited effects on the US, safe for a couple of states relatively close to the border or a couple of border-spanning industry networks (like the automotive industry).

Input-output linkages. Our measures of input-output linkages are, by construction, reasonably exogenous to the spatial stucture of the economy. First, observe that we compute those measures using national input-output shares instead of plant-level input-output shares. Hence, we do not pick up spuriously large values for inputs or outputs – due to substitution effects – when plants are located in close proximity to plants in related industries. Second, we exclude the own industry from the computation, so that the measure only picks up cross-industry links and not the geographical concentration of the industry itself (which is on the left-hand side of our regressions). Last, for each plant, the input and output distance is computed using *all other 256 industries in Canadian manufacturing*. For the geographical concentration of one industry to drive the input-output linkage measure, that industry would need to substantially affect the whole location patterns of most other related industries, which strikes us as fairly unlikely (though we cannot completely rule out this possibility). Although the input- and output-measures should be reasonably exogenous, we will also instrument them following Lewbel (2012) in the subsequent regressions. As we will see, our results are very stable across specifications.

As should be clear from the foregoing discussion, it is virtually impossible to fully solve all endogeneity issues given the level of aggregation at which we carry out our analysis. Yet, the panel nature of our data, our extensive set of time-varying controls, as well as the construction and instrumentation strategies for our main variables of interest – transportation costs, trade exposure, and input-output linkages – all help us to be reasonably confident that we identify causal effects of changes in those covariates $\mathbf{T}_{m,t}$ on our measure $\gamma_{m,t}(d)$ of geographical concentration.

4 Empirical results

We estimate four specifications based on equation (1), which differ by the set of industry characteristics and controls that they include.[19] **Model 1** includes a measure of industry size, proxies for industry structure (the Herfindahl index of the firm-size distribution, mean plant size, the share of plants controlled by multiplant firms, and the share of plants controlled by foreign-owned firms), and proxies for natural advantages (the share of inputs from natural resource-based industries, and the share of energy inputs in total output). It also includes the 'Marshallian covariates', namely the proxies for the skill composition of the workforce and for knowledge spillovers. **Model 2** adds our trade variables (import and export shares by

[19]We performed the Hausman test for (1) to confirm that the appropriate estimator is a fixed-effects estimator and not a random-effects estimator. The result of the test strongly confirms (at the 1% level) that the fixed-effects estimator is the preferred specification. Note also that we work with the universe of manufacturing industries, so that there is no sampling variability with respect to industries.

broad trading partner groups) to the baseline case. **Model 3** includes transportation costs and our input-output distances – the industry mean of the average minimum distance to a dollar of inputs or outputs computed using the five nearest plants in each industry – as well as our minimum distance (density) control.[20] Finally, **Model 4** – our preferred specification – includes all the variables and uses the residual transport cost obtained from a first-stage regression of that cost on industry multi-factor productivities and a set of industry and year fixed effects (see Section 3.3 for details).[21]

4.1 Baseline results

Our baseline results are presented in Table 4, which uses the unweighted (plant count) CDF at 50 kilometers distance as the dependent variable. Robustness checks with respect to that distance are provided in the next section, whereas robustness checks using the employment- and sales-weighted CDFS are relegated to Appendix E (see Table 10). All variables except the trade shares and the shares of plants controlled by multiplant and by foreign firms enter as natural logarithms into the regressions, so that their coefficients can be interpreted as elasticities.

As can be seen from Table 4, in Model 1, which includes only control variables, only total industry employment and the share of plants controlled by foreign firms are statistically significant. Put differently, growing industries and industries with an increasing share of foreign-controlled plants tend to become more localized. The first finding is at odds with results by Dumais, Ellison, and Glaeser (2002), who document that growing US manufacturing industries tend to disperse, whereas shrinking ones concentrate (see also Behrens, 2014, for the case of textiles in Canada). The second finding is in line with previous evidence which documents that foreign firms tend to locate within existing clusters (see, e.g., Head, Ries, and Swenson, 1995; and Guimaraes, Figueiredo, and Woodward, 2000). The natural resource share of inputs variables are basically never significant across all four models, i.e., changes in natural advantage is not strongly associated with changes in localization. One of the reasons for this is that their time variation is small. The same holds true for the 'Marshallian covariates', which are not significant either. Again, lack of time-series variation may explain that result.

Turning to Model 2, rising shares of imports are across the board associated with falling localization. The (non-OECD) Asian share of imports, which we use as a proxy for low-wage countries, has the largest estimated coefficient in absolute value and is the most statistically significant. One explanation for the dispersive effect of import competition is that firms become more footlose as they source a larger share of their intermediates from abroad and no

[20] Using $N = 3, 5, 10$ yields qualitatively very similar results.

[21] When using the 'ad valorem trucking cost residual' from the first-stage regression, we need to bootstrap the standard errors to control for the presence of an estimated regressor. We did this for the baseline specification (see Model 4 in Table 8), and it makes virtually no difference. We hence report non-bootstrapped standard errors in most specifications.

Table 4: Baseline estimation results for specification (1).

Variables	Dependent variable is the CDF at 50 kilometers (Model 1)	(Model 2)	(Model 3)	(Model 4)
Total industry employment	0.179[b]	0.150[b]	0.288[a]	0.289[a]
	(0.070)	(0.067)	(0.039)	(0.039)
Firm Herfindahl index (employment based)	-0.028	-0.038	0.002	0.001
	(0.036)	(0.035)	(0.021)	(0.021)
Mean plant size	-0.026	-0.029	-0.280[a]	-0.282[a]
	(0.078)	(0.077)	(0.045)	(0.044)
Share of plants affiliated with multiplant firms	-0.301	-0.203	-0.006	-0.005
	(0.191)	(0.164)	(0.100)	(0.099)
Share of plants controlled by foreign firm	0.584[a]	0.660[a]	0.338[a]	0.340[a]
	(0.216)	(0.214)	(0.125)	(0.124)
Natural resource share of inputs	0.024	0.034[c]	0.008	0.008
	(0.023)	(0.020)	(0.014)	(0.014)
Energy share of inputs	-0.037	-0.024	0.054	0.037
	(0.052)	(0.026)	(0.040)	(0.040)
Share of hours worked by all workers with post-secondary education	0.032	0.013	0.036	0.032
	(0.078)	(0.069)	(0.045)	(0.045)
In-house R&D share of sales	-0.031	0.006	0.011	0.014
	(0.020)	(0.022)	(0.015)	(0.015)
Asian share of imports		-1.570[a]	-1.132[a]	-1.119[a]
		(0.456)	(0.380)	(0.383)
OECD share of imports		-1.032[b]	-0.491	-0.476
		(0.412)	(0.344)	(0.345)
NAFTA share of imports		-1.114[a]	-0.562[c]	-0.549[c]
		(0.382)	(0.327)	(0.327)
Asian share of exports		0.473	0.482	0.482
		(0.500)	(0.405)	(0.412)
OECD share of exports		0.412[c]	0.440[b]	0.443[b]
		(0.237)	(0.189)	(0.193)
NAFTA share of exports		0.353	0.319	0.318
		(0.267)	(0.196)	(0.201)
Ad valorem trucking costs		-0.291[b]	-0.208[b]	
		(0.135)	(0.088)	
Ad valorem trucking costs (residual)				-0.260[a]
				(0.079)
Input distance			-0.361[a]	-0.358[a]
			(0.055)	(0.055)
Output distance			-0.313[a]	-0.318[a]
			(0.042)	(0.043)
Average minimum distance			-0.296[a]	-0.294[a]
			(0.039)	(0.039)
Number of NAICS industries	257	257	257	257
Number of years	17	17	17	17
Year dummies	yes	yes	yes	yes
Industry dummies	yes	yes	yes	yes
Observations (NAICS$\times$ years)	4,369	4,369	4,369	4,369
R^2	0.089	0.137	0.516	0.518

Notes: The dependent variable is the unweighted (count based) Duranton-Overman K-density CDF. ^{a}, ^{b} and ^{c} denote coefficients significant at the 1%, 5% and 10% levels, respectively. We use simple OLS. Standard errors are clustered at the industry level and given in parentheses. Our measures of input and output distances, as well as average minimum distance, are computed using $N = 5$. 'Ad valorem trucking costs (residual)' denotes the residual of the regression of 'Ad valorem trucking costs' on industry multi factor productivity. A constant term is included but not reported.

longer rely on (localized) domestic suppliers. Another explanation, for which Holmes and Stevens (2014) provide empirical evidence, is that import competition from low-wage countries leads to significant exit of large plants that produce standardized 'main segment' goods.[22] If those plants are the ones that are predominantly clustered at short distances, their exit will significantly reduce the extent of measured localization.[23] As can be also seen from Model 2 in Table 4, rising export shares are across the board associated with increasing localization, though the effect is only significant for the share of exports to OECD countries. This pattern may be driven by the fact that more isolated non-exporting plants have a higher chance to exit the market, or that localization increases the export participation and performance of plants (e.g., Koenig, Mayneris, and Poncet, 2010).

Regarding transportation costs, we have no clear prior as to their impact, as stated before. In theory, the effects of changes in transportation costs on the geographical concentration of economic activity depend on the underlying dispersion forces in the economy. If, on the one hand, firms tend to serve a predominantly dispersed immobile demand, lower transportation costs would tend to be agglomerative, as in Krugman (1991). If, on the other hand, all demand is a priori mobile and dispersion stems from urban costs due to agglomeration, lower transportation costs would tend to be dispersive (Helpman, 1998; Behrens, Mion, Murata, and Suedekum, 2012). As can be seen from Model 2 in Table 4, lower transportation costs are associated with more geographical concentration in our estimations.[24]

Model 3 adds our input- and output-linkage measures, whereas Model 4 uses the residual transport cost instead of the original variable. The input-output coefficients are highly significant and negative in all specifications, and they tend to be of similar magnitude (as in Ellison, Glaeser, and Kerr, 2010): industries tend to follow their suppliers and customers. If supplier industries tend to become more dispersed (in the sense of being, on average, further away from

[22]We cannot disentangle the impact of exit vs relocation on the spatial structure. However, we control for the size of the industry, which at least partly picks up entry and exit dynamics. Note that relocations are quite rare and should have little impact on our results. The bulk of the variation is driven by entry and exit.

[23]This is a somewhat surprising result, because we would expect the productivity enhancing effects of localization to shelter firms from low-wage competition. Yet, one should keep in mind that clustering provides firms with benefits as long as clusters grow (positive shocks), but that the unravelling of clusters (negative shocks) may lead to a domino effect as the agglomeration benefits dissipate with the exit of firms. Also, as shown by Holmes and Stevens (2014), plants in clusters operate on different market segments than non-clustered plants, and they are more vulnerable to import competition.

[24]We also experimented with different non-linear transportation cost specifications. More precisely, we estimated the effect of transportation costs with a spline, allowing the coefficients to vary between ad valorem rates of 0 to 0.05% (low), 0.05 to 15% (moderate), and 15% or greater (high). These are admittedly arbitrary categories, but ones that we believe make intuitive sense. The results are, by and large, consistent with the simpler specification that we use. Yet, we find that at low levels, the effect of transportation costs is positive or insignificant. At moderate levels, the coefficient is negative and always significant, and at high levels the coefficient is negative and insignificant.

plants in the downstream industry), the downstream industry becomes less concentrated too. This result suggests that the geographic concentration of upstream supply and downstream demand goes hand-in-hand with increasing localization of an industry. Note that this effect is not driven by changes in overall density, since we control for this (and the associated variable is highly significant). The coefficient for transportation costs remains fairly stable when introducing the input-output linkages, as can be seen from Model 3, albeit it slightly decreases in absolute value, as expected. Last, as can be seen from Model 4, the coefficient on transportation costs becomes larger in absolute value when using the productivity-purged residual. This is in line with our expectations discussed in Section 3.3, where we have argued that endogeneity concerns due to reverse causality are likely to bias the coefficient upwards (towards zero in this case). Observe that the endogeneity bias does not seem to be too severe, which is in line with findings related to the endogeneity of wages in standard 'wage-density' regressions (see, e.g., Combes, Duranton, and Gobillon, 2011, for a discussion). Last, as can be seen from our prefered specification (Model 4 in Table 4), about half of the time-series variation in localization is explained by the model.

As shown in Section 2.2, the degree of localization of manufacturing industries has significantly fallen in Canada between 1990 and 2009. How much of that change is explained by changes in transportation or trade costs? To see how much of the observed change can be attributed to changes in those variables, we compute the predicted change in the CDFs by holding, one-by-one, the: (i) ad valorem trucking costs; (ii) different import shares; and (iii) the input or output distances to their 1992 values, while still allowing the other variables to change through time. The results are summarized in Table 5.

Table 5: Predicted contributions to changes in geographical concentration.

Observed avg. CDF changes 1992–2008	Counterfactual avg. CDF changes 1992–2008 for changes in			
	Ad valorem trucking costs	Import shares	Input distances	Output distances
-23.37%	-28.36%	-14.63%	-30.32%	-31.86%

Notes: Observed and predicted changes in the unweighted cross-industry average CDFs at 50 kilometer distance.

As can be seen, the observed change in the cross-industry average CDF between 1992 and 2008 at a distance of 50 kilometers is -23.37%. Holding the ad valorem trucking rate fixed at its 1992 level, the change would have been -28.36%. Thus, had transportation costs not decreased, the geographical concentration would have fallen by about 5 percentage points more (about 20% of the overall change). Turning to imports, holding all import shares constant at their 1992 level, the change in the CDF would have been -14.63%. In words, had imports remained at their 1992 levels, the geographical concentration would have fallen by about 9 percentage points (i.e., 60%) less than what we observed. Clearly, these are large effects, thus showing that *transportation costs and trade exposure have sizable effects on the spatial structure of economic activity*. Last, turning to input and output distances, in the former case the change would have been

-30.32% (about 7 percentage points more) and in the latter case the change would have been -31.86% (about 8.5 percentage points more). Had supplier and customer access not changed – these distances fell through time, as can be seen from Figure 8 – the dispersion of industries would have been even greater than the one we observed.

4.2 Robustness checks

We now provide evidence on the robustness of our key findings. To this end, we run five main types of robustness checks. First, we investigate the robustness of our results to the choice of the dependent variable. Table 10 in Appendix E shows that the effect of transportation costs on localization is weaker – and the explanatory power of the model lower – when the latter is measured using either employment- or sales-weighted CDFs. Although the key qualitative flavor of the results and the sign and significance of our key coefficients remain largely unchanged, the estimates using employment- or sales-weighted K-densities are less sharp. Furthermore, the effect of import competition tends to be more limited to imports from Asia, and the coefficient tends to be smaller. This suggests that much of the adaptation to import competition, particularly from low wage countries which are responsible for the bulk of exit in Canadian manufacturing (Behrens, 2014), occurs for smaller plants and firms. Turning to the residual transportation cost variable, it remains significantly negative in all specifications that we estimate, irrespective of how we construct the dependent variable. The same holds true for the input-output distances and the overall density control. In a nutshell, changes in transportation costs and in input-output linkages have a significant effect on the spatial concentration of economic activity, no matter whether we consider plants, employment, or sales to measure that concentration.

Second, we check the robustness of our results to the choice of the distance d at which the K-density CDF is evaluated. Doing so allows us to highlight how our key covariates influence the localization of industries at different geographical scales. Furthermore, we can provide plots of the marginal effects of our variables of interest over the whole distance range, thus allowing for a fine analysis of the spatial dimensions of the changes in agglomeration due to changes in the trading environment. The left half of Table 6 summarizes our results for different distances. To save space, we only report results for Model 4 at three selected distances: 10, 100, and 500 kilometers. As can be seen, the qualitative results do not depend on the distance threshold d. This holds true for all our key variables, thus showing that transportation costs, trade, and input-output linkages matter at most spatial scales. Furthermore, there is a general tendency for the values and significance of the covariates to attenuate as the CDF increases in distance. This can be seen from the right half of Table 6, where we define the incremental distance of the CDF between distance d_1 and distance $d_2 > d_1$ as follows: $\Delta\gamma_m(d_1, d_2) = \gamma_m(d_1) - \gamma_m(d_2)$. We estimate the marginal effects of our variables by 'distance bands'. As one can see, there

Table 6: Estimation results for specification (1) by distance and by incremental change in the CDF.

	Model (4), by distance			Model (4), by incremental CDF			
Variables	CDF 10km	CDF 100km	CDF 500km	$\Delta\gamma_m(10,25)$	$\Delta\gamma_m(25,50)$	$\Delta\gamma_m(50,100)$	$\Delta\gamma_m(100,500)$
Asian share of imports	-1.359^a	-0.923^a	-0.307^b	-1.029^b	-0.724^b	-0.352	0.583
	(0.467)	(0.299)	(0.139)	(0.433)	(0.337)	(0.235)	(0.429)
OECD share of imports	-0.666	-0.334	0.018	-0.451	-0.174	0.102	0.721
	(0.425)	(0.271)	(0.158)	(0.374)	(0.285)	(0.211)	(0.455)
NAFTA share of imports	-0.710^c	-0.411	-0.037	-0.527	-0.284	0.007	0.587
	(0.396)	(0.254)	(0.135)	(0.359)	(0.268)	(0.190)	(0.372)
Asian share of exports	0.399	0.415	0.096	0.630	0.658	0.421	-0.782
	(0.439)	(0.345)	(0.123)	(0.426)	(0.404)	(0.264)	(0.714)
OECD share of exports	0.366^c	0.419^b	0.265^a	0.545^a	0.662^a	0.470^a	-0.112
	(0.219)	(0.166)	(0.094)	(0.197)	(0.224)	(0.156)	(0.304)
NAFTA share of exports	0.217	0.314^c	0.139^c	0.440^b	0.541^b	0.431^a	-0.191
	(0.231)	(0.174)	(0.080)	(0.211)	(0.215)	(0.162)	(0.274)
Ad valorem trucking costs (residual)	-0.269^a	-0.250^a	-0.212^a	-0.253^a	-0.238^a	-0.229^a	-0.105
	(0.080)	(0.073)	(0.048)	(0.079)	(0.080)	(0.069)	(0.090)
Input distance	-0.382^a	-0.340^a	-0.242^a	-0.332^a	-0.322^a	-0.315^a	-0.193^a
	(0.063)	(0.049)	(0.033)	(0.061)	(0.055)	(0.054)	(0.041)
Output distance	-0.307^a	-0.307^a	-0.197^a	-0.341^a	-0.340^a	-0.302^a	-0.122^a
	(0.046)	(0.040)	(0.027)	(0.045)	(0.045)	(0.045)	(0.039)
Average minimum distance	-0.322^a	-0.268^a	-0.137^a	-0.298^a	-0.243^a	-0.204^a	-0.038
	(0.046)	(0.035)	(0.024)	(0.041)	(0.043)	(0.038)	(0.036)
R^2	0.473	0.540	0 .545	0.481	0 .417	0.436	0.168

Notes: All estimations for 257 industries and 17 years (4,369 observations). The dependent variable is the unweighted (count based) Duranton-Overman K-density CDF at the reported distance. ^{a}, ^{b} and ^{c} denote coefficients significant at the 1%, 5% and 10% levels, respectively. We use simple OLS. All specifications include industry and year fixed effects. Standard errors, given in parentheses, are clustered at the industry level. Our measures of input and output distances are computed using $N = 5$. 'Ad valorem trucking costs (residual)' denotes the residual of the regression of 'Ad valorem trucking costs' on industry multi factor productivity. A constant term is included but not reported. All industry controls (Total industry employment; Firm Herfindahl index (employment based); Mean plant size; Share of plants affiliated with multiplant firms; Share of plants controlled by foreign firms; Natural resource share of inputs; Energy share of inputs; Share of hours worked by all workers with post-secondary education; In-house R&D share of sales) are included but not reported.

is basically no more additional effect of our covariates on the degree of localization beyond about 100 kilometers, except for our input-output measures. Furthermore, the largest (and statistically most significant results) occur in the distance bands between either 10 and 25 kilometers, or between 25 and 50 kilometers. This result suggests that many of the agglomeration mechanisms linked to transportation, trade, and input-output linkages operate at the scale of metropolitan areas.[25] At longer distances – beyond about 200 kilometers – other factors that do not figure in our model drive the clustering of firms, or incremental clustering becomes weak and fairly unimportant.[26] The decrease in the marginal effects can be clearly seen from

[25]For example, the island of Montreal is about 50 kilometers long.

[26]This result is not really surprising. There are two possible explanations. First, the determinants of localization may operate at 'small' spatial scales, whereas they are no longer very relevant at longer distances. Second, the CDFs across industries tend to display less variation the longer is the distance d. The reason is that they are bounded from above by unity, and we converge by construction to that value for all industries if we compute them over sufficiently large distances. This problem is similar to the spatial scale of aggregation issue when using different spatial scales to compute discrete measures like the Ellison and Glaeser (1997) index used by Rosenthal

Figure 9, which depicts the incremental change in coefficients of our key variables by 10 kilometers steps increases in distances (since all marginal coefficient changes are statistically zero after 200 kilometers, we limit the plots to that range).

Table 7: Estimation of specification (1) excluding textile and high-tech industries.

	Excluding textiles industries			**Excluding high-tech industries**		
Variables	CDF 10km	CDF 100km	CDF 500km	CDF 10km	CDF 100km	CDF 500km
Asian share of imports	-0.568^c	-0.508^c	-0.211	-1.517^a	-1.035^a	-0.380^b
	(0.322)	(0.282)	(0.174)	(0.554)	(0.350)	(0.155)
OECD share of imports	-0.035	0.007	0.137	-0.860	-0.474	-0.084
	(0.275)	(0.241)	(0.181)	(0.530)	(0.333)	(0.177)
NAFTA share of imports	-0.097	-0.062	0.076	-0.878^c	-0.531^c	-0.133
	(0.251)	(0.221)	(0.156)	(0.499)	(0.317)	(0.157)
Asian share of exports	0.627	0.505	0.096	0.468	0.469	0.111
	(0.440)	(0.358)	(0.130)	(0.490)	(0.378)	(0.121)
OECD share of exports	0.471^b	0.413^b	0.249^b	0.346	0.424^b	0.271^a
	(0.186)	(0.161)	(0.097)	(0.236)	(0.170)	(0.098)
NAFTA share of exports	0.400^b	0.348^b	0.128	0.149	0.275	0.124
	(0.196)	(0.170)	(0.080)	(0.246)	(0.179)	(0.085)
Ad valorem trucking costs (residual)	-0.213^a	-0.210^a	-0.193^a	-0.396^a	-0.324^b	-0.205^a
	(0.077)	(0.072)	(0.049)	(0.145)	(0.128)	(0.068)
Input distance	-0.458^a	-0.439^a	-0.315^a	-0.387^a	-0.346^a	-0.245^a
	(0.051)	(0.049)	(0.036)	(0.075)	(0.057)	(0.038)
Output distance	-0.265^a	-0.245^a	-0.155^a	-0.333^a	-0.336^a	-0.216^a
	(0.043)	(0.040)	(0.029)	(0.051)	(0.044)	(0.030)
Average minimum distance	-0.289^a	-0.265^a	-0.142^a	-0.321^a	-0.257^a	-0.128^a
	(0.041)	(0.038)	(0.026)	(0.053)	(0.038)	(0.026)
R^2	0.516	0.532	0.539	0.481	0.556	0.553

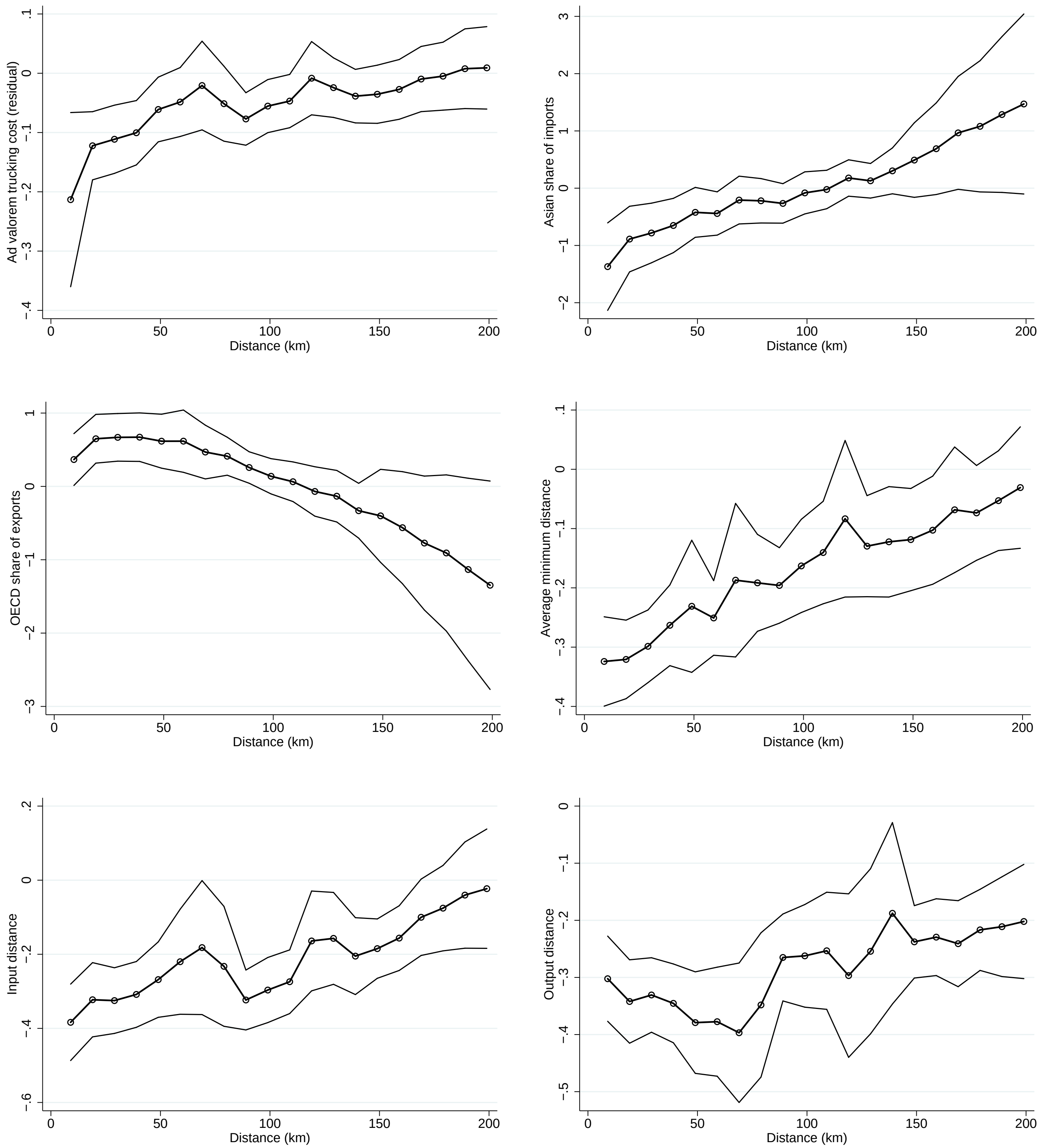
Notes: All estimations for 257 industries and 17 years (4,369 observations). ^{a}, ^{b}, ^{c} denote coefficients significant at the 1%, 5% and 10% levels, respectively. We use simple OLS. All specifications include industry and year fixed effects. Standard errors are clustered at the industry level and given in parentheses. Our measures of input and output distances are computed using $N = 5$. 'Ad valorem trucking costs (residual)' denotes the residual of the regression of 'Ad valorem trucking costs' on industry multi factor productivity. A constant term is included but not reported. All industry controls (Total industry employment; Firm Herfindahl index (employment based); Mean plant size; Share of plants affiliated with multiplant firms; Share of plants controlled by foreign firms; Natural resource share of inputs; Energy share of inputs; Share of hours worked by all workers with post-secondary education; In-house R&D share of sales) are included but not reported.

Third, we first re-estimate the model by averaging all variables over five year periods. Doing so reduces the year-on-year volatility of some variables (e.g., the trade variables), and allows for slowly moving variables like R&D expenditures or localization patterns to be potentially better identified in the regressions. It also deals potentially with business cycle aspects that may drive the changes in the geographical concentration of industries. The last three columns of Table 10 in Appendix E show that our basic findings are unchanged when replacing year-on-year variations with five-year averages.

Fourth, our results may be partly driven by sectoral 'outliers'. For example, as documented by Behrens (2014), the textile industries in Canada experienced a remarkable downward trend

and Strange (2001).

Figure 9: Transportation, trade, and input-output coefficients (marginal effect by distance).



in terms of number of plants and the geographical dispersion of activity in the wake of the end of the Multi-Fibre Arrangement in 2005. Given that these sectors were initially among the most strongly localized ones (see Table 1), and given that these industries have a tendency to display very strong co-agglomeration patters (see Ellison, Glaeser, and Kerr, 2010, p.1199), the large changes in these sectors may drive some of the results. That this is not the case, and that all of our main findings are robust to the exclusion of those sectors, is shown in Table 7. The left panel provides results when excluding the textile sectors, whereas the right panel provides results when excluding the high-tech sectors.[27] In both cases, our key coefficients are qualitatively unchanged. Note, however, two differences. First, the input-output linkages become more negative when excluding the textile industries. Second, the transport cost variable becomes more negative when excluding the high-tech industries. The former result suggests that textile industries are less dependent on input-output linkages than other industries (e.g., manufacturing durables). The latter result suggests that spatial patterns of high-tech industries are less impacted by changes in transportation costs, so that their inclusion tends to reduce the estimated coefficient on transport costs.

As a final series of robustness checks, we ran a number of experiments that we do not report in detail. We used, for example, the ICT investment variables from the KLEMS database, interacted with the other variables of the model, to check whether changes in communication costs have the same effect than changes in transportation costs. We did not get any significant coefficients – neither for the direct effects, nor for the interaction terms. We also estimated models with heterogeneous coefficients since transportation costs differ across industries. To this end, we split our sample into high-vs-low transport cost industries, using a 'below median'–'above median' criteria. The two coefficients were statistically identical. We also treated decreasing/increasing transportation costs in an asymmetric way as they may have asymmetric impacts. Again, the two coefficients were fairly close. We also replaced our measures of input and output linkages with the industry 'material share to sales' ratio, a proxy for reliance on intermediate inputs. That variable turns out to be insignificant in our regressions, whereas the other coefficients are largely unaffected. We also ran the model in a pooled cross-section and by year using a between estimator and found roughly the same signs and significant coefficients for transportation costs and the input and output distance measures. The cross-sectional results are summarized by Table 12 in Appendix E. It is worth noting that, although the levels of trade costs do seem to matter for the geographical concentration of industries, the time-series changes in those costs are much more strongly associated with changes in that concentration.

[27]Our definition of high-tech sectors is based on the US Bureau of Labor Statistics classification by Hecker (2005). This definition of high-tech industries is 'input based'. An industry is 'high-tech' if it employs a high proportion of scientists, engineers or technicians. As shown by Hecker (2005), these industries are also usually associated with a high R&D-to-sales ratio, and they also largely – but not always – produce goods that are classified as 'high-tech' by the Bureau of Economic Analysis.

Last, we also tried to control for the 'labor intensity' of an industry (not just highly skilled workers vs low-skilled workers). We constructed different measures using the quantity index of labor and the quantity index of capital from the KLEMS data, but these variables turned out again to be insignificant in our regressions.

To summarize, our key findings are fairly robust and continue to hold true in a variety of alternative specifications. Imports are mostly dispersive, whereas exports play in the opposite direction. Sectors that see their transportation costs increase tend to disperse more.[28] Last, our micro-geographic measures of input and output linkages are across the board the most significant and stable variables. Since they are computed by taking into account the *relative positions of all industries with respect to each other*, our findings suggest that there are very strong regularities in how industries relate spatially to one another and on how changes in the spatial structure of some industries shape changes in the spatial structure of linked industries.

4.3 Controlling for endogeneity

We finally address the potential endogeneity concerns that we discussed at length in Section 3.3. The results of the different estimations are summarized in Table 8.

Model 4 replicates column 4 of Table 4. As explained previously, we use the residual of a regression of ad valorem trucking costs on sectoral multifactor productivity – including a set of industry and year fixed effects – in that specification. The residual from that regression is, by construction, orthogonal to multifactor productivity. Observe that Model 4 in Table 8 differs from Model 4 in Table 4 only by the standard errors, which are bootstrapped using 200 replications. Comparing the results in the two tables shows that no coefficient changes its significance level. The coefficient on the residual ad valorem trucking rate is larger in absolute value than the coefficient that is not purged from productivity effects (-0.260 instead of -0.208). The direction of the bias is consistent with an industry price-decreasing effect of agglomeration ($p_{m,t}$ decreases in (2)) or a transportation sector price-increasing effect ($p_{\text{trans},t}$ increases in (2)). Both of these effects could underlie the upward bias in the coefficient on transportation costs that we estimate.

Model 5 is a standard 2SLS instrumental variable regression. We instrument the ad valorem trucking rate using formula (2), where we replace Canadian price indices with their US counterparts to construct our instrument. The rationale underlying this instrumentation strategy was explained before in Section 3.3 and is similar in spirit to that in Ellison, Glaeser, and Kerr (2010). The first-stage results are summarized in Table 11 in Appendix E. As can be seen from

[28] Holmes and Stevens (2014) document for the case of US manufacturing that import competition is dispersive for big firms that produce 'primary segment goods' in clusters, whereas small firms outside are less affected since they produce 'specialty segment goods' that are more costly to transport. Higher transport costs shield those small firms, whereas more trade exposes the larger firms. Our results concerning the impacts of changes in transportation costs and trade exposure are broadly in line with those findings.

that table, the instrument is strong (with a first-stage F-test value of 19.07 and an R^2 of 0.62). Table 8 shows that the instrumented coefficient is substantially more negative than the coefficient for the residual ad valorem trucking rate, itself more negative than the coefficient using the unpurged trucking rate. The direction of the bias in the estimated coefficients is the same in Models 4 and 5, which suggests that OLS estimates significantly underestimate the impact of changes in transportation costs on the spatial concentration of industries.

Finally, models 6 and 7 in Table 8 use the Lewbel (2012) estimator with internal instruments for the input-output distances and a set of the trade shares (see Appendix D for more details on the implementation).[29] The excluded external instrument is the US price-based ad valorem trucking costs as before. As can be seen from the results, the instrumented coefficient on the Asian share of imports increases, as do most of the other trade share coefficients. At the same time, both the magnitude of transportation costs and of the input and output distances decreases slightly. However, these variables remain significant and their magnitude is in the same ballpark than in the case of OLS (-0.194 vs -0.208 from Model 3 in Table 4). Thus, our results appear to be robust. Changes in transportation costs, in international trade exposure, and in access to suppliers and clients all affect the geographical concentration of manufacturing industries even when potential endogeneity concerns are taken into account.

5 Concluding remarks

Using a long panel of micro-geographic concentration measures, we have substantiated evidence for the causal effects of changes in transport costs – broadly defined – on the geographical concentration of Canadian manufacturing industries. We find large effects. Holding all other variables fixed at their 1992 levels, changes in trucking rates explain about 20%, changes in input-output linkages about 30%, and changes in import exposure about 60% of the observed decline in spatial concentration over the 1992–2008 period. Our qualitative results are robust to endogeneity concerns and to the way we measure the spatial concentration of industries – in terms of plants, employment, or sales.

Our research makes three distinct contributions. First, we construct new and finer measures of the costs of trading goods across space than in the previous literature. We use detailed microdata on freight transportation to estimate industry-level time-varying measures of transport costs, and we propose a new way of constructing micro-geographic input-output linkages based on location patterns and national input-output tables. Second, we are – to the best of our knowledge – among the first to exploit the time-series variation in the data to shed light on what drives *changes* in the spatial concentration of industries. The panel nature of the

[29]Since there is an insignificant correlation between the OECD export share and the squared residuals, we did not include it. We substituted instead the NAFTA import share because it is consistently significant in the baseline set of models and it meets the criteria for being internally instrumented.

Table 8: Controlling for potential endogeneity of $\mathbf{T}_{m,t}$ in specification (1).

	Dependent variable is the CDF at 50 kilometers			
	(Model 4)	**(Model 5)**	**(Model 6)**	**(Model 7)**
Variables	Base	IV-2SLS	Lewbel 1	Lewbel 2
Asian share of imports	-1.119^a	-1.110^a	-1.589^a	-1.621^a
	(0.420)	(0.377)	(0.533)	(0.495)
OECD share of imports	-0.476	-0.486		-0.673
	(0.393)	(0.341)		(0.416)
NAFTA share of imports	-0.549	-0.558^c	-0.756^c	-0.850^b
	(0.374)	(0.323)	(0.435)	(0.419)
Asian share of exports	0.482	0.452		0.641
	(0.409)	(0.398)		(0.580)
OECD share of exports	0.443^b	0.422^b		0.638^c
	(0.202)	(0.189)		(0.360)
NAFTA share of exports	0.318	0.297		0.532
	(0.206)	(0.194)		(0.365)
Ad valorem trucking costs		-0.346^a	-0.180^b	-0.194^b
		(0.095)	(0.091)	(0.089)
Ad valorem trucking costs (residual)	-0.260^a			
	(0.083)			
Input distance	-0.358^a	-0.359^a	-0.132^c	-0.223^a
	(0.053)	(0.054)	(0.077)	(0.076)
Output distance	-0.318^a	-0.314^a	-0.385^a	-0.349^a
	(0.040)	(0.042)	(0.086)	(0.086)
Average minimum distance	-0.294^a	-0.293^a		
	(0.041)	(0.039)		
R^2	0.518	0.514	0.316	0.328

Notes: The dependent variable is the unweighted (count based) Duranton-Overman K-density CDF. ^{a}, ^{b} and ^{c} denote coefficients significant at the 1%, 5% and 10% levels, respectively. Our measures of input and output distances are computed using $N = 5$. 'Ad valorem trucking costs (residual)' denotes the residual of the regression of 'Ad valorem trucking costs' on industry multi factor productivity. Model 4 replicates our preferred model but the standard errors are bootstrapped because of the generated regressor. Model 5 instruments the 'Ad valorem trucking costs' using costs constructed from US price indices. Models 6 and 7 use the Lewbel (2012) methodology to instrument input-output distances and trade shares. In model 6 only a subset of the import shares is instrumented, while all trade shares are instrumented in model 7. See Appendix D for details. A constant term is included but not reported. All industry controls (Total industry employment; Firm Herfindahl index (employment based); Mean plant size; Share of plants affiliated with multiplant firms; Share of plants controlled by foreign firms; Natural resource share of inputs; Energy share of inputs; Share of hours worked by all workers with post-secondary education; In-house R&D share of sales) are included but not reported.

data allows us to control for unobserved heterogeneity and a battery of other time-varying factors. We have highlighted a hitherto unnoticed tradeoff when using time-varying geographical concentration measures constructed from micro-geographic data: the need to smooth out the time-series volatility at short distances versus the potential underestimation bias of the concentration measures due to the smoothing. More work is called for here to propose better measures of concentration in the presence of substantial plant-level churning in the data. Last, by exploiting the spatially continuous nature of our data, we have also shed light on the spatial scale at which the aforementioned effects operate. In line with previous research that has looked at the geographical scale of knowledge spillovers, labor market pooling, and input-output linkages, we find that the costs of trading goods influence the spatial structure of industries at small geographical scales: whereas the effects are sizable at short distances up to 50 kilometers, they basically vanish beyond about 100–200 kilometers.

We believe that our results are important because they show that, although the costs of trading goods across space have hit historical lows, changes in those costs still do shape location patterns of industries. In a world where profit margins have become tiny, even small changes in trade costs can have large effects on firm location, specialization patterns, and trade. In a nutshell, the often heralded 'death of distance' is premature. The world is not yet flat: transport costs matter!

Acknowledgements. We thank Gabriel Ahlfeldt, John Baldwin, Antoine Bonnet, Pierre-Philippe Combes, Wulong Gu, Julien Martin, Se-il Mun, Yasusada Murata, Frédéric Robert-Nicoud, Mathieu Parenti, Eugenia Shevtsova, Peter Warda, and conference participants at the 2013 NARSC Meetings in Atlanta, the 2014 Meetings of the Société canadienne de science économique in Ottawa, the 2014 Western Regional Science Meetings in San Diego, the 2014 African Econometric Society Meetings, the Summer School on 'New Structural Economics and Development in Africa' at Peking University, the Center for Market Studies and Spatial Economics in Saint Petersburg, the University of Geneva, the University of Quebec at Montreal, and the 2014 ERSA–UEA Meetings in Saint Petersburg, for very helpful comments and suggestions. Financial support from Statistics Canada's 'Trade Cost Analytical Projects Initiative' (API) is gratefully acknowledged. Tim Pendergast provided excellent technical assistance with program codes. Behrens and Bougna gratefully acknowledge 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*. Bougna further gratefully acknowledges financial support from CIRPÉE. This work was carried out while Behrens and Bougna were Deemed Employees of Statistics Canada. The views expressed in this paper and all remaining errors are ours. This paper has been screened to ensure that no confidential data are revealed.

References

[1] 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.

[2] Behrens, Kristian. 2014. "Unweaving clusters: Manufacturing localization and international trade in Canada, 2001–2009." *In progress*, Univ. of Quebec at Montreal, Canada.

[3] Behrens, Kristian, and Théophile Bougna. 2013. "An anatomy of the geographical concentration of Canadian manufacturing industries." CIRPÉE Discussion Paper #13–27.

[4] Behrens, Kristian, Carl Gaigné, Gianmarco I.P. Ottaviano, and Jacques-François Thisse. 2007. "Countries, regions, and trade: On the welfare impacts of economic intergration." *European Economic Review* 51**(5)**, 1277–1301.

[5] Behrens, Kristian, Giordano Mion, Yasusada Murata, and Jens Südekum. 2012. "Spatial frictions." CEPR Discussion Paper #8572, Center for Economic Policy Research, London, UK.

[6] Behrens, Kristian, and Pierre M. Picard. 2011. "Transportation, freight rates, and economic geography." *Journal of International Economics* 85**(2)**: 280–291.

[7] Brown, W. Mark, and William P. Anderson. 2015. "How thick is the border: the relative cost of Canadian domestic and cross-border truck-borne trade, 2004-2009." *Journal of Transportation Geography* 42: 10–21.

[8] Brülhart, Marius (2011). "The spatial effects of trade openness: a survey." *Review of World Economics* 147**(1)**: 59–83.

[9] Brülhart, Marius, Céline Carrère, and Frédéric Robert-Nicoud. 2014. "Trade and towns: On the uneven effects of trade liberalization." *In progress*, Université de Lausane *and* Université de Genève, Switzerland.

[10] Brülhart, Marius, Céline Carrère, and Federico Trionfetti. 2012. "How wages and employment adjust to trade liberalization: Quasi-experimental evidence from Austria." *Journal of International Economics* 86**(1)**: 68–81.

[11] Combes, Pierre-Philippe, Gilles Duranton, and Laurent Gobillon. 2011. "The identification of agglomeration economies." *Journal of Economic Geography* 11**(2)**: 253–266.

[12] Combes, Pierre-Philippe, and Laurent Gobillon. 2014. "The empirics of agglomeration economies." In: G. Duranton, J.V. Henderson, and W.C. Strange (eds.), *Handbook of Regional and Urban Economics, vol. 5*. North-Holland: Elsevier, *forthcoming*.

[13] D'Costa, Sabine (2010). "Trade liberalization and the geographical concentration of industries: Evidence from Hungarian manufacturing industries." *PhD Dissertation*, London School of Economics, UK.

[14] Dauth, Wolfgang, Sebastian Findeisen, and Jens Suedekum. 2014. "The rise of the east and the far east: German labor markets and trade integration." *Journal of the European Economic Association* 12**(6)**: 1643–1675.

[15] Dumais, Guy, Glenn D. Ellison, and Edward L. Glaeser. 2002. "Geographic concentration as a dynamic process." *Review of Economics and Statistics* 84**(2)**: 193–204.

[16] 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.

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

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

[19] Duranton, Gilles, Peter M. Morrow, and Matthew Turner. 2014. "Roads and trade: Evidence from the US." *Review of Economic Studies* 81**(2)**, 681–724.

[20] Duranton, Gilles, and Matthew Turner. 2012. "Urban growth and transportation." *Review of Economic Studies* 79**(4)**, 1407–1440.

[21] 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.

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

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

[24] Fujita, Masahisa, Krugman, Paul R., and Anthony J. Venables (1999). *The Spatial Economy: Cities, Regions and International Trade*. MIT Press, Cambridge, MA.

[25] Guimaraes, Paulo, Octávio Figueiredo, and Douglas Woodward. 2000. "Agglomeration and the location of foreign direct investment in Portugal." *Journal of Urban Economics* 47**(1)**: 115–135.

[26] Hecker, Daniel E. 2005. "High-technology employment: a NAICS-based update." *Monthly Labor Review* 128**(7)**: 57–72.

[27] Head, Keith, John Ries, and Debrorah Swenson. 1995. "Agglomeration benefits and location choice: Evidence from Japanese manufacturing investments in the United States." *Journal of International Economics* 38**(3-4)**: 223–247.

[28] Helpman, Elhanan. 1998. "The size of regions." In: D. Pines, E. Sadka, I. Zilcha (eds.), *Topics in Public Economics. Theoretical and Empirical Analysis*, Cambridge University Press, pp. 33–54.

[29] Henderson, J. Vernon. 1997. "Medium-sized cities." *Regional Science and Urban Economics* 27**(6)**: 583–612.

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

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

[32] Holmes, Thomas J. and John J. Stevens (2004). "Spatial distribution of economic activities in North America." In: J.V. Henderson, J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics, vol. 4*. North-Holland: Elsevier B.V., pp. 2797–2843.

[33] Holmes, Thomas J., and John J. Stevens. 2002. "Geographic concentration and plant scale." *Review of Economics and Statistics* 84**(4)**: 682–690.

[34] Jonkeren, O., Erhan Demirel, Jos van Ommeren, and Piet Rietveld. 2009. "Endogenous transport prices and trade imbalances." *Journal of Economic Geography* 11**(3)**: 509–527.

[35] Kim, Sukkoo. 1995. "Expansion of markets and the geographic distribution of economic activities: The trends in U.S. regional manufacturing structure, 1860-1987." *Quarterly Journal of Economics* 110**(4)**: 881–908.

[36] Koenig, Pamina, Florian Mayneris, and Sandra Poncet. 2010. "Local export spillovers in France." *European Economic Review* 54**(4)**: 622–641.

[37] Krugman, Paul R. 1991. "Increasing returns and economic geography." *Journal of Political Economy* 99**(3)**: 483–499.

[38] Krugman, Paul R., and R. Livas Elizondo. 1996. "Trade policy and the third world metropolis." *Journal of Development Economics* 49**(1)**: 137–150.

[39] Krugman, Paul R., and Anthony J. Venables (1995). "Globalization and the inequality of nations." *Quarterly Journal of Economics* 110**(4)**: 857–880.

[40] Lewbel, Arthur. 2012. "Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models." *Journal of Business and Economic Statistics* 30**(1)**: 67–80.

[41] Lu, Jiangyong, and Zhigang Tao. 2009. "Trends and determinants of China's industrial agglomeration." *Journal of Urban Economics* 65**(2)**: 167–180.

[42] Murata, Yasusada, Ryo Nakajima, and Ryuichi Tamura. 2014. "Testing for localization using micro-geographic data: A new approach." *In progress*, Nihon University, Japan.

[43] Redding, Stephen J., and Daniel M. Sturm. 2008. "The Costs of Remoteness: Evidence from German Division and Reunification." *American Economic Review* 98**(5)**, 1766–1797.

[44] Rosenthal, Stuart S., and William C. Strange. 2004. "Evidence on the nature and sources of agglomeration economies." In: J.V. Henderson, J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics, vol. 4*. North-Holland: Elsevier B.V., pp. 2119–2171.

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

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

Appendix

This set of appendices is structured as follows. Appendix A describes our datasets, data sources, and key variables. Appendix B provides details on the Duranton-Overman K-density computations. Appendix C describes the construction of the weights used in our input-output measures. Appendix D provides details for the implementation of the Lewbel (2012) estimates. Last, Appendix E contains supplemental tables and results.

A. Data and data sources

This appendix provides details on the data used and the data sources. A description of the key variables and the associated descriptive statistics are given in Table 3 in the main text.

Plant-level data and industries. Our analysis is based on the Annual Survey of Manufacturers (ASM) Longitudinal Microdata file. This data cover the years from 1990 to 2010. Our focus is on manufacturing plants only. For every plant we have information on: its primary 6-digit NAICS code (the codes are consistent over the 20 year period); its year of establishment; its total employment; whether or not it is an exporter in selected years; its sales; the number of non-production and production workers; and its 6-digit postal code. The latter, in combination with the Postal Code Conversion files (PCCF), allows us to effectively geo-locate the plants by associating them with the geographical coordinate of their postal code centroids.

The survey frame of the ASM has evolved over time. Early in the period, it was relatively stable with, on average, about 32,000 plants per sample year. The sample of plants was restricted to those with total employment (production plus non-production workers) above zero, and plants must have sales in excess of $30,000. Also, aggregate records were excluded. These records represent multiple (typically small) plants without latitudes and longitudes. In 2000, however, the number of plants in the survey increased substantially as the ASM moved from its own frame to Statistics Canada's centralized Business Register, increasing the sample to an average of 53,000 plants. In 2004, however, the number of plants in the frame was once again restricted, with many of the small plants once again excluded, or included in aggregate records. With this in place, the sample returned to near previous levels, averaging about 33,000 plants between 2004 and 2009. The expanded survey scope in the early 2000s had little effect on trends in the CDFS, but there was an effect on the number of industries found to be localized or dispersed (see Table 9 in the Appendix). Our econometric analysis deals with the change in the sample frame through the inclusion of year fixed effects.

We also use the ASM to construct controls for the labor market variables, for some natural advantage proxies, and for industry ownership structure variables that we include in the regressions. All variables are constructed by aggregating plant-level data to the industry level.

***L*-level input-output tables.** We use these tables to construct our plant-level proxies for the importance of input and output linkages (see Appendix C and Section 3.2.4 for more details). The L-level tables are at a more aggregate level than the 6-digit NAICS level. We break them down to the 6-digit level based on industries' weights in terms of sales.

KLEMS database. This database, which covers the period from 1961 to 2008, contains various industry-level informations useful for constructing proxies for natural advantage (e.g., energy intensity, water usage etc.).

Trucking micro-data. The trucking micro-data comes from Statistics Canada's Trucking Commodity Origin-Destination Survey and from the 'experiment export trade file' produced in 2008 (see Brown and Anderson, 2015, for details). Section 3.2.2 provides details on the methodology used to estimate ad valorem rates by industry and year.

Geographical data. To geolocate firms, we use latitude and longitude data of postal code centroids obtained from Statistics Canada's Postal Code Conversion files (PCCF). These files associate each postal code with different Standard Geographical Classifications (SGC) that are used for reporting census data in Canada. We match firm-level postal code information with geographical coordinates from the PCCF.

Trade data. The industry-level trade data come from Industry Canada and cover the years 1992 to 2009. The dataset reports imports and exports at the NAICS 6-digit level by province and by country of origin and destination. We aggregate the data across provinces and compute the shares of exports and imports that go to or originate from a set of country groups: Asian countries, OECD countries, and NAFTA countries. Since the trade data is available from 1992 on, whereas the KLEMS data is available until 2008, we restrict our sample to the 1992–2008 period in all estimations to maintain comparability of results.

US **price indices.** We use detailed year-by-year NAICS 6-digit price indices from the NBER-CES Manufacturing Productivity Databas (`http://nber.org/data/nberces5809.html`) to construct instruments for Canadian industry-level transportation costs. Methodological details are provided in Sections 3.3 and 4.3.

B. The distance-based approach to measuring localization

Following Duranton and Overman (2005, 2008), hereafter DO, the estimator of the kernel density (probability density function or PDF) of bilateral distances between plants at a given distance d, is given by:

$$\widehat{K}(d) = \frac{1}{n(n-1)h} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f\left(\frac{d-d_{ij}}{h}\right), \tag{B.1}$$

where h is Silverman's optimal bandwidth and f is a Gaussian kernel function. The distance d_{ij} (in kilometers) between plants i and j is computed as:

$$d_{ij} = 6378.39 \cdot \text{acos}\left[\cos(|\text{lon}_i - \text{lon}_j|)\cos(\text{lat}_i)\cos(\text{lat}_j) + \sin(\text{lat}_i)\sin(\text{lat}_j)\right]. \tag{B.2}$$

Alternatively, rather than using plant counts as the unit of observation in (B.1), we can characterize the localization of employment or sales at the industry level. This can be accommodated by adding weights to (B.1):

$$\widehat{K}_W(d) = \frac{1}{h\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(e_i+e_j)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (e_i+e_j) f\left(\frac{d-d_{ij}}{h}\right), \tag{B.3}$$

where e_i and e_j are the employment or sales levels of plants i and j, respectively.[30] The weighted K-density thus describes the distribution of bilateral distances between plants weighted by either employees or sales in a given industry, whereas the unweighted K-density describes the distribution of bilateral distances between plants in that industry. When required, as in

[30]Contrary to Duranton and Overman (2005), who use a multiplicative weighting scheme, we use an additive one. The additive scheme gives less weight to pairs of large plants and more weight to pairs of smaller plants than the multiplicative scheme does. Using a multiplicative scheme would imply that our results may be too strongly driven by a few very large firms in a given industry.

Table 9, we follow Duranton and Overman (2005) and implement a Monte Carlo approach for measuring the statistical significance of localization of industries.

To construct the K-densities, we need to fix a cutoff distance. Following Behrens and Bougna (2014), we choose a cutoff distance of 800 kilometer for computing the K-densities. The interactions across 'neighboring cities' mostly fall into that range in Canada. In particular, a cutoff distance of 800 kilometer includes interactions within the 'western cluster' (Calgary, AB; Edmonton, AB; Saskatoon, SK; and Regina, SK); the 'plains cluster' (Winnipeg, MB; Regina, SK; Thunder Bay, ON); the 'central cluster' (Toronto, ON; Montréal, QC; Ottawa, ON; and Québec, QC); and the 'Atlantic cluster' (Halifax, NS; Fredericton, NB; and Charlottetown, PE). Setting the cutoff distance to 800 kilometer allows us to account for industrial localization at both very small spatial scales, but also at larger interregional scales for which market-mediated input-output and demand linkages, as well as market size, might matter much more.

While the K-density PDF provides a clear picture of localization at every distance d, and while it allows for statistical testing, it is not well suited in capturing globally the location patterns of industries up to some distance d. This can, however, be achieved by using the K-density cumulative distribution up to distance d. In all our econometric estimations, we use as dependent variable the CDF of the K-densities. Those are given by:

$$\text{CDF}(d) = \sum_{\delta=1}^{d} \widehat{K}(\delta) \quad \text{and} \quad \text{CDF}_W(d) = \sum_{\delta=1}^{d} \widehat{K}_W(\delta). \tag{B.4}$$

Finally, for the purpose of comparision of our results, we also compute the 'raw' unweighted CDFS of the distribution of bilateral distances, which are given by

$$\text{RAW}(d) = \frac{1}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \chi(d_{ij} \leq d), \tag{B.5}$$

where n is the number of plants in the industry and where $\chi(\cdot)$ is an indicator function that takes value 1 if the bilateral distance d_{ij} is less than d and zero otherwise. While (B.4) provides a kernel-smoothed distribution, (B.5) provides a raw distribution.

Table 1 provides the (unweighted) K-density CDFS in 1990, 1999, and 2009 for the most strongly localized industries in Canada; while Table 2 summarizes the industry-average K-densities across years and using different weighting schemes. Last, Table 9 summarizes the year-on-year location patterns of industries based on the formal significance test of Duranton and Overman (2005) that we have described in the foregoing.

C. Input-output shares

We use the L-level national input-output tables from Statistics Canada at buyers' prices. These tables – which constitute the finest sectoral public release – feature 42 sectors that are somewhere in between the NAICS 3- and NAICS 4-digit levels. For each industry, i, we allocate total inputs purchased or outputs sold in the L-level matrix to the corresponding NAICS 6-digit sectors. We allocate total sales to each subsector in proportion to that sector's sales in the total sales to obtain a 257 × 257 matrix of NAICS 6-digit inputs and outputs, which we use in constructing the linkages.[31] From that table, we compute the share α_{ij} that sector i sells to sector j. We also compute the share β_{ij} that sector i buys from sector j. We systematically exclude within-sector transactions where $i = j$, as those may be capturing all sorts of intra-sectoral agglomeration economies that are conducive to clustering but not correlated with input-output linkages. Thus, the weights we use in equations (3) and (4) are given by

$$\omega^{\text{in}}_{\Omega(\ell),s} \equiv \alpha_{\Omega(\ell),s} \quad \text{and} \quad \omega^{\text{out}}_{\Omega(\ell),s} \equiv \beta_{\Omega(\ell),s}. \tag{C.1}$$

Using the L-level matrix provides smoother series of input-output linkages than those obtained using the confidential W-level national input-output tables (which are directly in the 257 × 257 industries format).

D. Applying the Lewbel (2012) method

To apply the Lewbel (2012) procedure, we need to verify two conditions: heteroscedasticity and correlation. First, we regress the potentially endogeneous variables (input and output distances, trade shares, and trucking costs) on all other exogeneous variables of the model. We then predict the residuals of that regression and run a standard heteroscedasticity test. We need to reject the homoscedasticity assumption for the Lewbel method to be applicable. In our case, we strongly reject the null hypothesis of homoscedasticity for all series of residuals (the p-value is zero in all tests). Second, we take the square of the predict residuals from the foregoing regression, and check the correlation between the dependent variable of the regression (input distances, or output distances, or the different trade shares, or trucking costs) and those squared residuals. The correlation needs to be 'strong' and statistically strongly significant for the instruments to work properly. In our case, this condition holds true for transportation costs, the input and output distances, and for all import shares: the correlation of the squared residuals with the variable itself is significant at 1% in all cases. It is 0.067 for transportation costs, -0.081 for input distances, -0.089 for output distances, 0.130 for the Asian share of imports, and -0.079 for the NAFTA share of imports. We find no statistically significant correlation for the export shares.

Since the two conditions (heteroscedasticity of the residuals and correlation of the squared residuals with the variable) are met in our case, we can apply the Lewbel estimator. Since fixed effects cannot be included in the estimation (see `ivreg2h` in Stata), we de-mean all variables

[31]Because of confidentiality reasons, we do not use the finer W-level matrices since this would make disclosure of results more problematic. However, the tests we ran using those matrices yield very similar results to the ones we report in this paper.

by industry first. The exogeneous variables are partialled-out for the Lewbel estimator and so their coefficients are not reported. Since we have an exogeneous instrument for transportation costs, we apply the Lewbel estimator only to deal with potential endogeneity concerns of trade shares and input-output distances.

E. Additional tables and results

Table 9 summarizes the location patterns by year and by statistical significance following the methodology developed by Duranton and Overman (2005). It contains information on the percentage of industries with random, localized, and dispersed point patterns for all years between 1990 and 2009. Table 10 contains robustness checks for the estimation of model (1) using the employment- and sales-weighted K-density CDFs, respectively. It also replicates our main results by averaging all variables over five-year intervals to reduce the volatility of some variables, and to allow slow-changing variables to be better identified. Table 11 contains the first-stage estimates for the IV regression, whereas Table 12 contains the cross-sectional estimates (both pooled and year-by-year) for transportation costs.

Table 9: Percentage of industries with random, localized, and dispersed point patterns, 1990 to 2009.

	Unweighted (plant counts)			Employment weighted			Sales weighted		
Year	Random	Localized	Dispersed	Random	Localized	Dispersed	Random	Localized	Dispersed
1990	52.53	34.63	12.84	52.53	36.96	10.51	54.86	37.35	7.78
1991	51.36	36.19	12.45	52.92	38.52	8.56	55.25	36.19	8.56
1992	53.70	36.19	10.12	56.42	35.02	8.56	58.37	33.46	8.17
1993	53.70	34.24	12.06	58.37	33.46	8.17	59.53	31.52	8.95
1994	49.81	36.96	13.23	57.20	33.07	9.73	60.70	30.74	8.56
1995	55.25	33.46	11.28	58.37	33.07	8.56	59.53	32.30	8.17
1996	54.09	35.41	10.51	56.03	35.41	8.56	59.53	33.46	7.00
1997	55.25	35.41	9.34	60.70	32.30	7.00	61.09	32.68	6.23
1998	55.64	34.24	10.12	58.37	35.02	6.61	61.87	32.68	5.45
1999	55.25	34.63	10.12	58.75	35.41	5.84	61.48	32.30	6.23
2000	47.86	37.74	14.40	51.75	40.47	7.78	53.31	40.47	6.23
2001	43.58	41.25	15.18	52.92	40.86	6.23	50.58	42.41	7.00
2002	45.91	39.69	14.40	50.97	41.63	7.39	54.86	37.35	7.78
2003	47.47	36.58	15.95	50.58	40.86	8.56	55.64	35.41	8.95
2004	60.31	30.35	9.34	60.31	33.07	6.61	60.70	32.30	7.00
2005	58.75	33.46	7.78	62.65	31.13	6.23	64.20	31.52	4.28
2006	60.31	30.35	9.34	60.31	33.46	6.23	62.26	33.85	3.89
2007	57.59	33.46	8.95	60.70	33.85	5.45	62.65	32.30	5.06
2008	56.03	34.24	9.73	61.48	31.91	6.61	64.59	29.96	5.45
2009	59.53	33.07	7.39	63.04	31.52	5.45	63.04	31.13	5.84

Source: Authors' computations using the *Annual Survey of Manufacturers* Longitudinal Microdata file. The statistical significance of the location patterns is computed using Monte Carlo simulations with 1,000 replications following the procedure developped by Duranton and Overman (2005).

Table 10: Estimation of specification (1) using employment-weighted CDFs, sales-weighted CDFs, and five year averages.

Dependent variable	Employment weighted CDF			Sales weighted CDF			Unweighted CDF, five year averages		
Variables	CDF 10km	CDF 100km	CDF 500km	CDF 10km	CDF 100km	CDF 500km	CDF 10km	CDF 100km	CDF 500km
Total industry employment	0.289^a	0.235^a	0.074^b	0.309^a	0.257^a	0.091^a	0.313^a	0.242^a	0.077^b
	(0.049)	(0.041)	(0.029)	(0.050)	(0.043)	(0.029)	(0.052)	(0.040)	(0.033)
Firm Herfindahl index (employment based)	-0.001	-0.009	0.003	0.023	0.015	0.021	0.011	-0.004	0.003
	(0.028)	(0.025)	(0.020)	(0.029)	(0.025)	(0.020)	(0.037)	(0.027)	(0.022)
Mean plant size	-0.230^a	-0.177^a	-0.044	-0.258^a	-0.199^a	-0.062	-0.286^a	-0.233^a	-0.069
	(0.055)	(0.049)	(0.039)	(0.054)	(0.048)	(0.037)	(0.063)	(0.052)	(0.043)
Share of plants affiliated with multiplant firms	-0.010	-0.133	-0.204^b	0.024	-0.113	-0.199^b	-0.003	-0.115	-0.213^b
	(0.123)	(0.110)	(0.081)	(0.126)	(0.112)	(0.080)	(0.143)	(0.118)	(0.086)
Share of plants controlled by foreign firm	0.233	0.274^b	0.261^a	0.238	0.271^c	0.223^a	0.163	0.256^c	0.264^b
	(0.144)	(0.128)	(0.081)	(0.161)	(0.141)	(0.085)	(0.169)	(0.137)	(0.105)
Natural resource share of inputs	-0.005	0.007	0.003	-0.010	0.003	-0.001	0.027	0.035^b	0.010
	(0.017)	(0.011)	(0.008)	(0.017)	(0.012)	(0.008)	(0.025)	(0.016)	(0.011)
Energy share of inputs	0.058	0.033	0.020	0.048	0.024	0.013	0.045	0.032	0.037
	(0.051)	(0.048)	(0.035)	(0.053)	(0.049)	(0.035)	(0.058)	(0.047)	(0.033)
Share of hours worked by all workers with post-secondary education	0.028	0.041	0.035	0.014	0.023	0.021	-0.214	-0.126	-0.058
	(0.064)	(0.054)	(0.033)	(0.071)	(0.056)	(0.033)	(0.137)	(0.114)	(0.087)
In-house R&D share of sales	0.004	0.019	0.018^b	-0.005	0.011	0.016^c	0.019	0.041^b	0.032^a
	(0.017)	(0.013)	(0.009)	(0.018)	(0.014)	(0.009)	(0.023)	(0.018)	(0.012)
Asian share of imports	-0.684^b	-0.531^b	-0.241^c	-0.713^b	-0.604^b	-0.285^c	-1.463^b	-1.012^a	-0.383^c
	(0.312)	(0.252)	(0.145)	(0.349)	(0.276)	(0.162)	(0.579)	(0.357)	(0.202)
OECD share of imports	-0.377	-0.232	0.008	-0.305	-0.186	0.043	-0.770	-0.351	-0.006
	(0.264)	(0.217)	(0.164)	(0.286)	(0.236)	(0.176)	(0.566)	(0.336)	(0.236)
NAFTA share of imports	-0.312	-0.208	-0.018	-0.262	-0.195	0.003	-0.821	-0.477	-0.104
	(0.244)	(0.198)	(0.141)	(0.276)	(0.226)	(0.159)	(0.518)	(0.317)	(0.201)
Asian share of exports	0.264	0.368	0.065	0.217	0.299	0.082	0.322	0.366	0.051
	(0.483)	(0.389)	(0.130)	(0.507)	(0.398)	(0.106)	(0.539)	(0.439)	(0.211)
OECD share of exports	0.212	0.330	0.181^c	0.349	0.424^c	0.280^a	0.360	0.450	0.266
	(0.295)	(0.210)	(0.094)	(0.288)	(0.216)	(0.096)	(0.386)	(0.314)	(0.191)
NAFTA share of exports	0.111	0.276	0.098	0.190	0.318	0.169^b	0.265	0.442	0.180
	(0.310)	(0.206)	(0.075)	(0.303)	(0.213)	(0.076)	(0.383)	(0.296)	(0.149)
Ad valorem trucking costs (residual)	-0.158^b	-0.150^b	-0.148^a	-0.134^c	-0.127^c	-0.137^a	-0.377^a	-0.361^a	-0.315^a
	(0.077)	(0.072)	(0.053)	(0.076)	(0.070)	(0.045)	(0.085)	(0.076)	(0.060)
Input distance	-0.256^a	-0.238^a	-0.186^a	-0.256^a	-0.239^a	-0.180^a	-0.258^a	-0.246^a	-0.221^a
	(0.063)	(0.054)	(0.032)	(0.064)	(0.056)	(0.033)	(0.073)	(0.059)	(0.043)
Output distance	-0.234^a	-0.222^a	-0.127^a	-0.200^a	-0.193^a	-0.113^a	-0.374^a	-0.383^a	-0.239^a
	(0.053)	(0.048)	(0.030)	(0.056)	(0.048)	(0.029)	(0.069)	(0.062)	(0.044)
Minimum distance	-0.312^a	-0.246^a	-0.119^a	-0.327^a	-0.249^a	-0.131^a	-0.400^a	-0.297^a	-0.141^a
	(0.050)	(0.039)	(0.026)	(0.054)	(0.039)	(0.026)	(0.067)	(0.043)	(0.032)
Number of NAICS industries	257	257	257	257	257	257	257	257	257
Number of years	17	17	17	17	17	17	4	4	4
Industry dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations (NAICS $\times$ years)	4,369	4,369	4,369	4,369	4,369	4,369	1,028	1,028	1,028
R^2	0.318	0.371	0.381	0.294	0.359	0.376	0.517	0.599	0.598

Notes: ^{a}, ^{b}, ^{c} denote coefficients significant at the 1%, 5% and 10% levels, respectively. We use simple OLS. Standard errors, given in parentheses, are clustered at the industry level. Our measures of input and output distances are computed using $N = 5$. A constant term is included but not reported.

Table 11: First-stage results for the IV regression.

Variables	Dependent variable: Ad valorem trucking costs
Total industry employment	0.017
	(0.014)
Firm Herfindahl index (employment based)	0.002
	(0.010)
Mean plant size	0.006
	(0.014)
Share of plants affiliated with multiplant firms	0.026
	(0.039)
Share of plants controlled by foreign firm	0.055
	(0.044)
Natural resource share of inputs	-0.008
	(0.006)
Energy share of inputs	0.084^a
	(0.018)
Share of hours worked by all workers with post-secondary education	-0.057^a
	(0.014)
In-house R&D share of sales	0.024^a
	(0.009)
Asian share of imports	-0.056
	(0.107)
OECD share of imports	0.067
	(0.095)
NAFTA share of imports	0.021
	(0.109)
Asian share of exports	-0.156^c
	(0.089)
OECD share of exports	-0.104
	(0.072)
NAFTA share of exports	-0.065
	(0.069)
Ad valorem trucking costs US (instrument)	0.485^a
	(0.111)
Input distance	0.035^c
	(0.020)
Output distance	-0.011
	(0.015)
Average minimum distance	0.005
	(0.014)
First-stage R^2	0.628
First-stage F test of excluded instruments	19.07

Notes: ^{a}, ^{b}, ^{c} denote coefficients significant at the 1%, 5% and 10% levels, respectively. OLS regression of 'ad valorem trucking cost' on the ad valorem trucking cost US (our instrument) and all control variables. We report the first-stage R^2 and note from the first-stage test that the instrument is strong.

Table 12: Cross-sectional estimates, pooled and year-by-year.

Dependent variable: CDF at 50 kilometers			
Pooled cross section		Yearly cross sections (ad valorem trucking costs (residual))	
Asian share of imports	-0.044	1992	-0.128[a]
	(0.272)		(0.045)
OECD share of imports	-0.094	1993	-0.116[b]
	(0.268)		(0.046)
NAFTA share of imports	-0.062	1994	-0.097[b]
	(0.207)		(0.041)
Asian share of exports	0.531	1995	-0.109[b]
	(0.552)		(0.043)
OECD share of exports	0.288	1996	-0.090[b]
	(0.336)		(0.041)
NAFTA share of exports	0.201	1997	-0.074[c]
	(0.248)		(0.040)
Ad valorem trucking costs (residual)	-0.065[b]	1998	-0.064
	(0.031)		(0.041)
Input distance	-0.306[a]	1999	-0.060
	(0.098)		(0.046)
Output distance	-0.428[a]	2000	0.008
	(0.099)		(0.042)
Average minimum distance	-0.380[a]	2001	-0.039
	(0.062)		(0.040)
Observations	4,369	2002	-0.038
R^2	0.773		(0.041)
		2003	-0.041
			(0.039)
		2004	-0.043
			(0.047)
		2005	-0.028
			(0.045)
		2006	-0.044
			(0.044)
		2007	-0.062
			(0.040)
		2008	-0.068[c]
			(0.036)

Notes: [a], [b], [c] denote coefficients significant at the 1%, 5% and 10% levels, respectively. OLS regressions, dependent variables is the CDF at 50 kilometers distance. All specifications include the same controls than in the main text. There are no time fixed effects in the pooled cross section. Huber-White robust standard errors in parentheses.