Received: 13 March 2018 | Revised: 26 December 2018 | Accepted: 12 February 2019

DOI: 10.1111/jors.12436

ORIGINAL MANUSCRIPT

WILEY JOURNAL OF REGIONAL SCIENCE

Manufacturing (co)agglomeration in a transition country: Evidence from Russia

Ekaterina Aleksandrova[1] | Kristian Behrens[1,2,3] | Maria Kuznetsova[1]

[1]Department of Economics, National Research University Higher School of Economics, Moscow, Russia

[2]Department of Economics, Université du Québec à Montréal, Montreal, Canada

[3]CEPR, London, UK

Correspondence
Kristian Behrens, Department of Economics, Université du Québec à Montréal, Case postale 8888, Succursale Centre-ville, Montréal, QC, H3C 3P8, Canada; National Research University Higher School of Economics, Russia; and CEPR, UK.
Email: behrens.kristian@uqam.ca

Funding information
Social Sciences and Humanities Research Council of Canada, Grant/Award Number: Canada Research Chair; National Research University Higher School of Economics, Grant/Award Number: 5-100; Russian Government, Grant/Award Number: 11.G34.31.0059

Abstract

We document the geographic concentration patterns of Russian manufacturing using detailed microgeographic data. About 80% of three-digit industries are significantly agglomerated, and a similar share of three-digit industry pairs is significantly coagglomerated. Industry pairs with stronger buyer–supplier links—as measured using Russian input–output tables—tend to be slightly more coagglomerated. This result is robust to instrumental variable estimation using either Canadian or US instruments. Using Canadian ad valorem transport costs as a proxy for transport costs in Russia, we further find that industries with higher transport costs are more dispersed, and industry pairs with higher transport costs are less coagglomerated.

KEYWORDS
ad valorem transport costs, agglomeration, coagglomeration, input–output linkages, Russian manufacturing industries

JEL CLASSIFICATION
R12

1 | INTRODUCTION

The uneven spatial distribution of industries is a first-order feature of almost any country in the world. While this has been extensively documented for developed countries—especially for manufacturing—there is a dearth of evidence for developing or transition countries (see Duranton, 2015; World Bank, 2009). This is unfortunate because it is precisely for those countries that understanding geographic concentration—especially for manufacturing—and the associated productivity gains is important to assess economic development prospects and options.

There is now a broad consensus that agglomeration has a causal effect on productivity due to the existence of agglomeration economies: Doubling the size of an industry in a geographic area increases productivity by about 2–5% on average (Combes & Gobillon, 2015; Melo, Graham, & Noland, 2009; Rosenthal & Strange, 2004). Realizing these

productivity gains from geographic concentration may be especially important for transition countries, such as Russia. It has repeatedly been pointed out that Russia needs to reduce its dependence on oil and primary goods and that it must substantially improve its weak manufacturing productivity. According to Deloitte's *2016 Global Manufacturing Competitiveness Index*, Russia ranks 32 out of 40 countries—lower than Brazil, South Africa, and Poland. It has lost four ranks since 2013 and is projected to stay at its current rank by 2020.[1] There is clearly room for substantial improvements, and those may be partly achieved by policies that require a better understanding of geographic concentration patterns and their underlying determinants. How concentrated are manufacturing industries in Russia? Which industries are more concentrated than others? And what are the potential mechanisms explaining that differential concentration?

The aim of this paper is twofold. First, we provide a detailed picture of the geographic concentration patterns of Russian manufacturing industries. Using recent and highly disaggregated point-pattern data, we estimate the agglomeration of industries and the coagglomeration of industry pairs.[2] We pay special attention to Russia's "dual geographic structure," that is, the existence of a dense western and a scattered eastern part. Second, we investigate some of the potential determinants of the geographic concentration of individual industries and of the coagglomeration of industry pairs, paying special attention to "technological relationships" as embodied in input–output links and transport costs. We focus on these two variables since they are less likely to depend on the institutional setting, thereby allowing us to use proxies from other countries if needed. Also, input–output tables are one of the rare data that we have access to in Russia, which allows us to use them directly in our analysis.

Our key results may be summarized as follows. First, we document the existence of strong spatial patterns. About 80% of three-digit industries are significantly agglomerated, with a substantially higher share in the European part than in the Asian part of Russia. Roughly, the same share of industry pairs is significantly coagglomerated, mainly at short distances below 100 km and at distances between 650 and 800 km (the distance between the Moscow and the Saint Petersburg regions).

Second, the overall patterns of geographic concentration—their extent, strength, and composition—are surprisingly similar to those documented for manufacturing industries in other countries, such as the United Kingdom, Canada, or the United States. Hence, geographic concentration seems to obey similar rules, despite Russia's history of a centrally planned economy that explains in large part the geographic structure of industry before 1990 (see, e.g., Kofanov & Mikhailova, 2015).

Third, we find that stronger buyer–supplier links and lower transport costs are associated with more geographic concentration. While the former effect is relatively weak using Russian input–output data, it becomes stronger once we use US and Canadian instruments. This suggests that there is some downward bias, consistent with the view that the historically inherited patterns were established for reasons other than input–output linkages. We also find that industries that face higher transport costs—measured using industry-level ad valorem trucking costs from Canada—are consistently more geographically dispersed than industries that face lower transport costs. This finding is in line with recent evidence for Canada and suggests that geographic concentration is stronger when transport costs are low. We finally also document that industry pairs with stronger input–output links tend to be more coagglomerated only if transporting their outputs is more costly, thus further pointing to the role that buyer–supplier links and the costs of shipping goods play in driving the coagglomeration of industries.

[1]Available online at https://www2.deloitte.com/global/en/pages/manufacturing/articles/global-manufacturing-competitiveness-index.html, last accessed on February 15, 2018. According to the *Global Competitiveness Report 2014–2015*, Russia ranks 119 out of 144 countries in terms of its goods market competition and efficiency. See http://www3.weforum.org/docs/WEF_GlobalCompetitivenessReport_2014-15.pdf, last accessed on February 20, 2018.

[2]There are only few works on geographic concentration in Russia. They all use either the Herfindahl–Hirschman Index, or the Krugman Dissimilarity Index, or the Theil Index—and all rely on fairly aggregated regional and industrial data (see, e.g., Kolomak, 2015; Maslikhina, 2017; Rastvortseva & Chentsova, 2015). We are aware of two papers that use more disaggregated data. Vorobyev, Kislyak, and Davidson (2010) use a sample of about 10,000 firms coded to the city level to estimate localization and urbanization economies for different broad industries. Kofanov and Mikhailova (2015) use a sample of plants—taken from the industrial census of the USSR in 1989 and coded to the settlement level—to look for differences in the geographic structure of manufacturing industries between the Soviet state-planned economy and the free market economy starting in the early 1990s. Neither paper provides estimates for coagglomeration patterns of industry pairs or an analysis of the determinants of geographic concentration patterns.

The remainder of the paper is structured as follows. Section 2 briefly presents our data. Section 3 provides detailed results for the agglomeration of individual manufacturing industries and the coagglomeration of industry pairs. We provide results for all of Russia, as well as for the western and the eastern parts separately. Section 4 analyzes the importance of input–output linkages and of transport costs for agglomeration and coagglomeration in Russia. Finally, Section 5 concludes our study. Detailed explanations concerning our data and additional results are relegated to a set of appendices.

2 | DATA

We start with a brief overview of our data. Additional details on data collection and processing, as well as the different data sources, are relegated to Appendix A.1. Our main data set is the 2014 RUSLANA database, which contains information about Russian companies and establishments.[3] We focus on the manufacturing portion of that database and retain all establishments that were active in 2014 and whose contact information—especially address—were updated after 2012. Basic data cleaning and geocoding, using a three-stage procedure detailed in Appendix A.2, yield 345,384 geocoded establishments, of which 320,934 are geocoded accurately.

Each establishment reports a primary industry code from the National Industry Classification (OKVED 2007), which is similar to the NACE Rev.2 classification. We use industry codes up to the three-digit level. Although finer levels are reported by a number of establishments, doing so was not mandatory before 2012. Hence, industry codes beyond the three-digit level may be unreliable—some plants only report three-digit codes, whereas others in the same industry report four-digit codes or finer. Eliminating establishments that do not report three-digit information and converting the four- to three-digit codes, we end up with a final data set of 316,967 accurately geocoded establishments. Table S1 in the Supporting Information Appendix provides a breakdown of these establishments by three-digit industry codes.

Russia is a geographically large country with a quite dense European part (the west) and a less dense Asian part (the east). These two parts display very different settlement and population patterns. They are also separated by a natural barrier, the Ural mountains (see Appendix A for additional details). To account for this geographic heterogeneity, we will consider both the overall spatial distribution of industries in Russia and the distributions in the east and in the west separately. Figure 1 depicts the spatial distribution of all manufacturing establishments in our sample in Russia in 2014, as well as the east–west division along the Ural mountains. As shown, manufacturing establishments are densely packed in the western part, whereas the eastern part displays a much sparser and more scattered pattern that essentially follows the Trans-Siberian railway line.

3 | GEOGRAPHIC AGGLOMERATION AND COAGGLOMERATION PATTERNS

Our first aim is to document the geographic concentration patterns of Russian manufacturing industries. Figure 2 illustrates two types of patterns: the *agglomeration* of a single industry ("Manufacture of motor vehicles, trailers and semi-trailers," OKVED 34) in Figure 2a,[4] and the *coagglomeration* of two industries ("Spinning of textile fibers," OKVED 171; "Weaving manufacture," OKVED 172) in Figure 2b. We successively look at these two types of concentrations.

[3]We use interchangeably the terms "establishments" and "plants" in the paper. Both refer to a physical location where a firm operates (part of) its activities.

[4]This two-digit industry consists of OKVED 341 ("Manufacture of motor vehicles"), 342 ("Manufacture of bodies (coachwork) for motor vehicles;

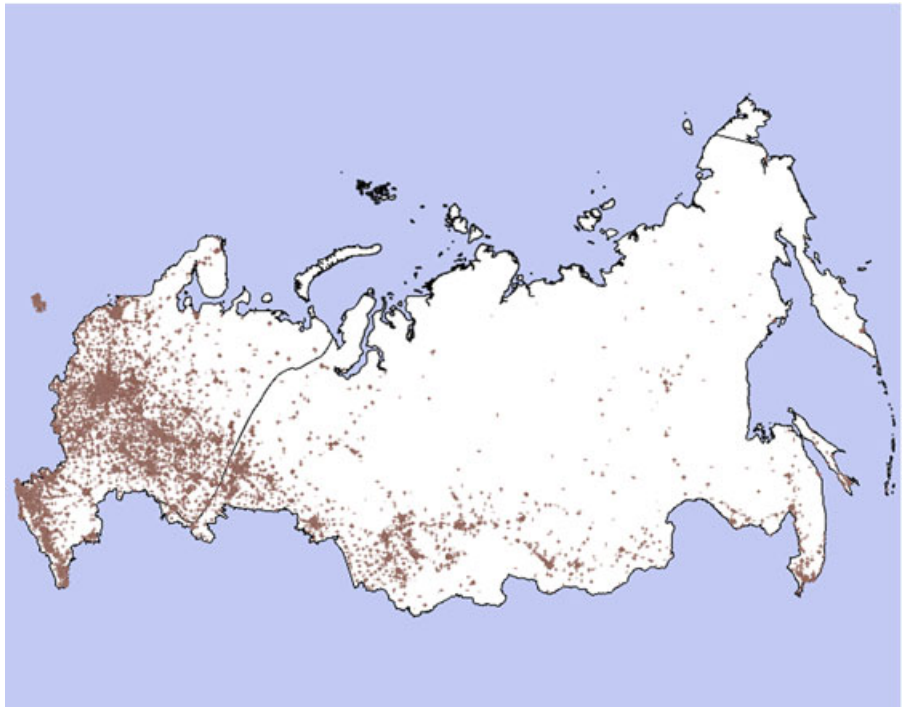

FIGURE 1 Distribution of manufacturing plants in Russia in 2014 and east–west divide [Color figure can be viewed at wileyonlinelibrary.com]

3.1 | Agglomeration: Methodology

We follow Duranton and Overman (2008, 2005) who develop a methodology that uses bilateral distances between plants to assess geographic concentration.[5] The idea is to estimate a kernel-smoothed distribution (K-density) of the bilateral distances between plants, which can then be used to: (a) identify localized industries, that is, industries that display significant geographic concentration *relative* to manufacturing in general, and (b) construct measures of *absolute* geographic concentration of those industries, that is, measures of the "spatial compactness" of their plants. The idea underlying (a) is to apply sampling and bootstrapping techniques to compare the observed distribution of bilateral distances between the plants in an industry to a set of bilateral distances obtained from samples of randomly drawn plants among all manufacturing plants. Doing so allows us to measure *relative geographic concentration*, that is, how much more—or less—industries are concentrated with respect to manufacturing in general. The idea in (b) is to construct the cumulative distribution of the K-density up to some distance d, which measures the *absolute geographic concentration* of an industry, namely, the share of bilateral distances between plants in that industry below the distance threshold d (see Behrens, Bougna, & Brown, 2018; Behrens & Brown, 2018). These two measures are complementary and capture two different, yet equally important, aspects of the geographic concentration process (see Marcon & Puech, 2017, for a recent survey of those measures).

manufacture of trailers and semi-trailers"), and 343 ("Manufacture of parts and accessories for motor vehicles and their engines").

[5]We use the DO index because we have access to geocoded data. Alternatively, we could compute area-based indices—for example, the Ellison–Glaeser index (Ellison & Glaeser, 1997) or the index by Mori, Nishikimi, and Smith (2005)—but this is problematic in Russia. The reason lies in the structure of the administrative divisions—there are either very large "Subject of Federation" regions (85 in all), or quite heterogeneous "municipalities" (about 2,300). The latter are based on population thresholds and, therefore, range from very small units in the west to sometimes gigantic units in the east. We thus strongly believe that the exercise using administrative divisions does not make much sense. Ideally, we would use "local labor markets" or similar divisions to compute the indices, but those simply do not exist for Russia.

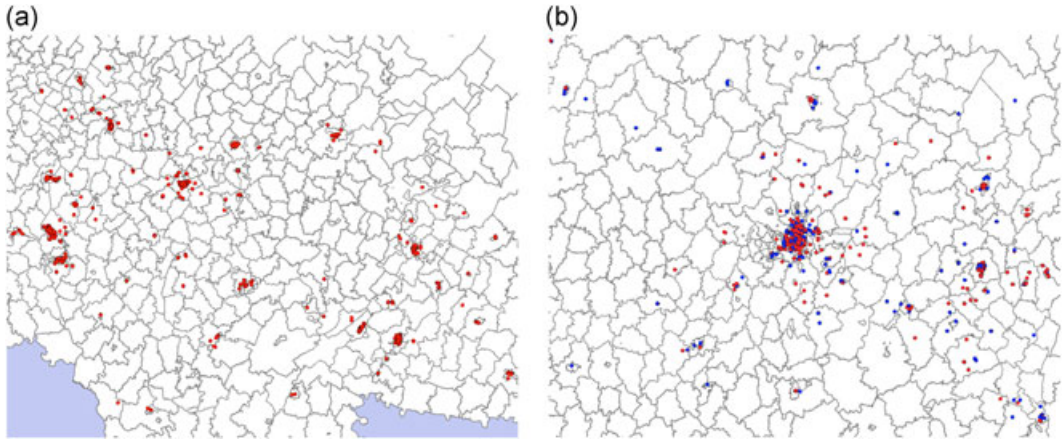

FIGURE 2 Examples of agglomeration and coagglomeration patterns: (a) Motor vechicles trailers, semi-trailers; (b) Spinning (blue) and weaving (red) [Color figure can be viewed at wileyonlinelibrary.com]

The methodology developed by Duranton and Overman (2008, 2005) has four steps. First, we compute the pairwise distances between all plants in an industry and estimate a kernel density of their distribution. Second, we construct a counterfactual distribution by assuming that the plants in a given industry are randomly reallocated among all possible locations where we observe manufacturing activity. We use that distribution to estimate a counterfactual kernel density. Third, to assess whether the observed location patterns depart statistically significantly from randomness, we repeat the second step 1,000 times to construct confidence intervals from the 1,000 counterfactual *K*-densities. Last, we test whether an industry is significantly localized, or significantly dispersed, or not significantly different from a random allocation, by comparing the actual distribution of bilateral distances with the confidence bands derived from the sampling procedure. We provide more information on these four steps, as well as other technical details concerning our implementation, in Appendix B.

3.2 | Agglomeration: Results

We estimate *K*-densities for 101 three-digit industries for all of Russia, and separately for the western and the eastern parts of Russia. For each industry, we also compute local and global confidence bands based on 1,000 random permutations, as explained above.

3.2.1 | Results for all of Russia

Figure 3 depicts the *K*-densities and confidence bands corresponding to four different representative location patterns. First, "Manufacture of beverages" (OKVED 159) in Figure 3a is significantly less localized than manufacturing in general. This industry is hence geographically more dispersed than overall manufacturing activity. Second, "Cast iron and steel, other primary processing" (OKVED 273) in Figure 3b is neither localized nor dispersed. It closely follows the overall distribution of manufacturing in Russia and can, therefore, not be distinguished from an industry that would locate randomly. Third, "Manufacture of motor vehicles" (OKVED 341) in Figure 3c is significantly more localized than manufacturing in general, especially at short distances, at distances of about 400 km, and at longer distances. However, that industry is not jointly overrepresented in the Moscow and the Saint Petersburg regions (the largest metro areas in Russia), as can be seen from its *K*-density at about 600–800 km, which corresponds to the distance between these two major metropolitan areas. Last, "Weaving manufacture" (OKVED 172) in Figure 3d is the most strongly localized industry in our example, especially at short

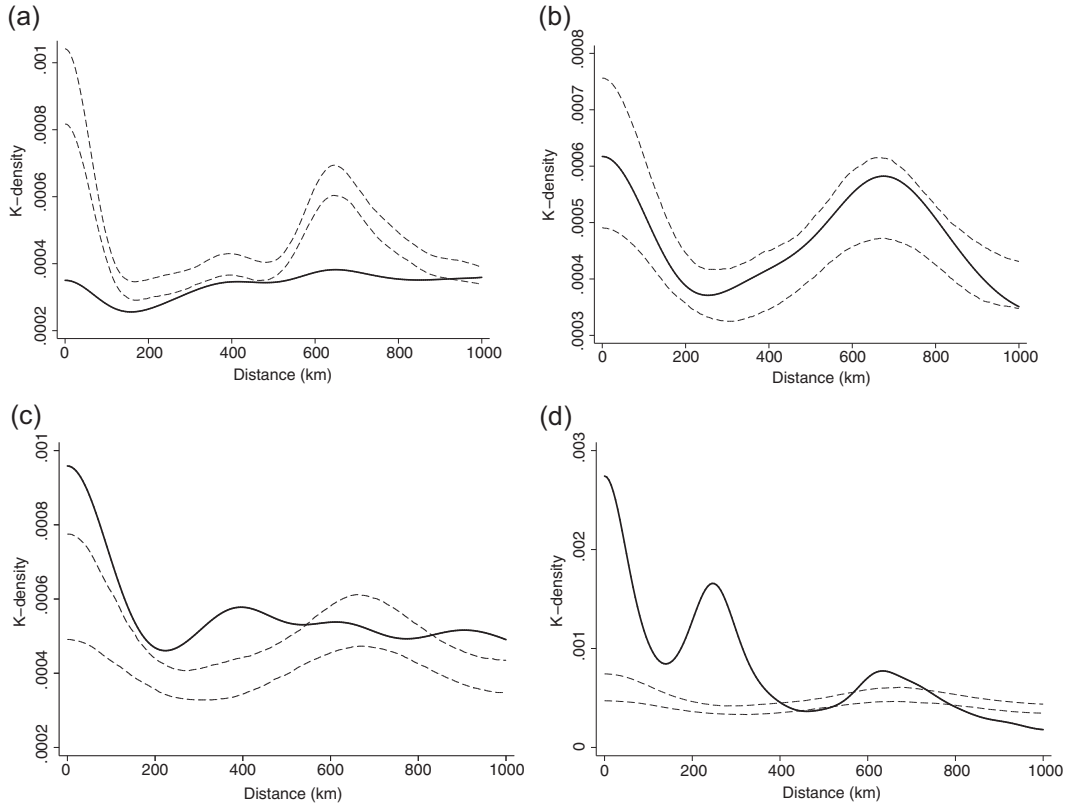

FIGURE 3 *K*-densities and confidence bands of selected OKVED three-digit industries, all of Russia: (a) Manufacture of beverages; (b) Cast iron and steel, other primary processing; (c) Manufacture of motor vehicles; (d) Weaving manufacture

TABLE 1 Summary of geographic concentration patterns for Russian manufacturing industries

	(a) All of Russia N = 316, 967 plants		(b) Western Russia N = 245, 616 plants		(c) Eastern Russia N = 71, 351 plants	
Status	**Number**	**Percentage (%)**	**Number**	**Percentage (%)**	**Number**	**s(%)**
Localized industry	81	80.20	89	88.12	67	66.34
Random	6	5.94	6	5.94	25	24.75
Dispersed industry	14	13.86	6	5.94	9	8.91
Excess localization $\overline{\Gamma}\|_{\Gamma_i>0}$	0.063		0.057		0.052	
Excess dispersion $\overline{\Psi}\|_{\Psi_i>0}$	0.043		0.028		0.013	
Total	101	100	101	100	101	100

Note. All *K*-densities are computed over a range of 0–1000 km, for 101 three-digit OKVED industries. The confidence bands are computed using 1,000 bootstrap replications. We compute the *K*-densities in 5 km steps. See Figure 1 and Appendix A.2 for details on how we split Russia into a western and an eastern part. The values of $\overline{\Gamma}|_{\Gamma_i>0}$ and $\overline{\Psi}|_{\Psi_i>0}$ are computed at the last point at which the *K*-densities are evaluated, that is, $\overline{d}$ = 1,000 km. We report average values for all significantly localized industries in the case of $\overline{\Gamma}|_{\Gamma_i>0}$, and for all significantly dispersed industries in the case of $\overline{\Psi}|_{\Psi_i>0}$.

TABLE 2 Localization patterns of OKVED three-digit industries by broad two-digit industry groups, all of Russia

OKVED2 industries	Industry name	No. of OKVED3 subindustries	No. of localized	No. of random	No. of dispersed	% localized
			Strong localization patterns			
34	Manufacture of motor vehicles, trailers, and semitrailers	3	3	0	0	100
22	Publishing, printing, and reproduction of recorded media	3	3	0	0	100
32	Manufacture of radio, television, and communication electronic components and apparatus	3	3	0	0	100
20	Woodworking and manufacture of wood and cork articles, except furniture	5	5	0	0	100
21	Manufacture of cellulose, pulp, paper, cardboard, and articles of these materials	2	2	0	0	100
18	Manufacture of wearing apparel; dressing and dyeing of fur	3	3	0	0	100
25	Manufacture of rubber and plastic products	2	2	0	0	100
24	Manufacture of chemicals and chemical products	7	7	0	0	100
17	Textile manufacture	7	7	0	0	100
36	Manufacture of furniture; manufacturing	6	6	0	0	100
19	Manufacturing of leather; leather articles and manufacture of footwear	3	3	0	0	100
			Intermediate localization patterns			
29	Manufacture of machinery and equipment	7	6	0	1	86
31	Manufacture of electrical machinery and apparatus not elsewhere classified	6	5	1	0	83
33	Manufacture of medical instruments, measure, control and test devices, optical devices,photo and cine equipment, watches	5	4	1	0	80
28	Manufacture of fabricated metal products	7	5	2	0	71
23	Manufacture of coke, refined petroleum products, and nuclear fuel	3	2	0	1	67

(Continues)

TABLE 2 (Continued)

OKVED2 industries	Industry name	No. of OKVED3 subindustries	No. of localized	No. of random	No. of dispersed	% localized
			Weak localization patterns			
27	Manufacture of basic metals	5	3	0	2	60
35	Manufacture of ships, aircraft and spacecraft, and other transport	5	3	1	1	60
15	Manufacture of food products and beverages	9	5	4	0	56
37	Recycling of secondary raw materials	2	1	1	0	50
26	Manufacture of other nonmetallic mineral products	8	3	4	1	38

Note. The localization status of all OKVED three-digit industries within the same two-digit industry is reported. We group two-digit industries by broad localization patterns ("Strong localization patterns," "Intermediate localization patterns," and "Weak localization patterns") based on the frequency of localization of the three-digit industries that make up the two-digit industry.

geographic distances, and at distances of about 200 km. That industry is also jointly overrepresented in the Moscow and the Saint Petersburg regions, as seen from the second peak at around 650–700 km.

Table 1 summarizes the agglomeration and dispersion patterns for the 101 three-digit industries for all of Russia, the western part, and the eastern part. As shown in panel (a), about 80% of the three-digit industries are localized in Russia as a whole. There is, hence, generally a substantial degree of localization. Panels (b) and (c) show that the patterns are stronger in the more dense western part of Russia (about 88% of localized industries) and weaker in the less dense eastern part of Russia (about 66% of localized industries). We return to this point below.

Table 2 shows how geographic concentration patterns differ systematically across broad industry groups. The top panel lists the two-digit industries that contain the largest shares of localized three-digit industries. As Table 2 shows, the most localized industry groups include textile and leather (OKVED 17–19), printing and publishing (OKVED 22), paper (OKVED 21), and different types of chemicals, rubber, and plastic (OKVED 24–25). Industries with intermediate localization patterns include different types of machinery (OKVED 29, 31, and 33), as well as coke and fabricated metal products (OKVED 23 and 28). Finally, the least localized industry groups are related to basic metals and nonmetallic mineral products (OKVED 26–27), food and beverages (OKVED 15), recycling (OKVED 37), and—somewhat surprisingly—manufacture of ships, aircraft, and spacecraft (OKVED 35). The latter finding is likely due to the level of aggregation since the two-digit industry masks substantial heterogeneity; it conflates, for example, shipbuilding and spacecraft, two very different industries in terms of potential location patterns.

Figure 4 depicts the number of significantly localized industries by distance. As shown, there are a large number of significantly localized industries at short distances, and localization falls off rapidly with distance. It rises and peaks again at about 650–700 km, which is the distance between the two major economic centers of the country, the federal cities of Moscow and Saint Petersburg. This suggests that many industries tend to cluster both at short distances and at intermediate distances between major economic centers. As we show later using the coagglomeration patterns of industry pairs, the major economic centers still display substantial specialization, that is, they host also a share of mutually exclusive industries.

As explained in Appendix B.1, we can construct a relative measure, $\bar{I}_i$, of the strength of localization of industry i, and an absolute measure given by the K-density cumulative distribution function (CDF). Figure 4c shows that there are only few highly localized three-digit industries, whereas most industries are not strongly localized. Figure 4d further shows that these patterns are less skewed when considering the CDFs instead of the strength of localization. When taken together, these two results suggest that there is substantial localization and geographic agglomeration of some industries, whereas the patterns for most other industries are less pronounced.

Which industries are the most strongly localized compared to manufacturing in general? And which industries are the most strongly agglomerated spatially? Table 3 lists the top 10 most localized and most geographically concentrated three-digit manufacturing industries in Russia. As shown, textile-related industries, recording, pharmaceuticals, aircraft and spacecraft, and some parts of the motor vehicles industries rank among the most localized industries. These patterns are fairly similar for both localization and geographic concentration, thus suggesting that the most agglomerated industries are also those that are the most strongly localized.

How do our results thus far compare to those of previous studies? Starting with Canada—another geographically large country—we find that there is a substantial overlap in the types of industries that are strongly localized. In Canada, for example, the industry groups that are among the most geographically localized include "Clothing Manufacturing," "Textile Mills," "Machinery Manufacturing," and "Printing and Related Support Activities," whereas those among the most dispersed include "Petroleum and Coal Products Manufacturing," "Food Manufacturing," "Beverage and Tobacco Product Manufacturing," and "Non-Metallic Mineral Product Manufacturing" (see Behrens & Bougna, 2015, Table 7). Hence, the overall pattern is fairly similar across both countries, as shown by Table 2. Given that the two countries are institutionally very different, this suggests that these pattern may be driven by technological considerations and therefore, be more general than what we would think. We return to that point later when investigating the importance of input–output linkages and transport costs—two arguably technological parameters—for the geographic concentration of industries.

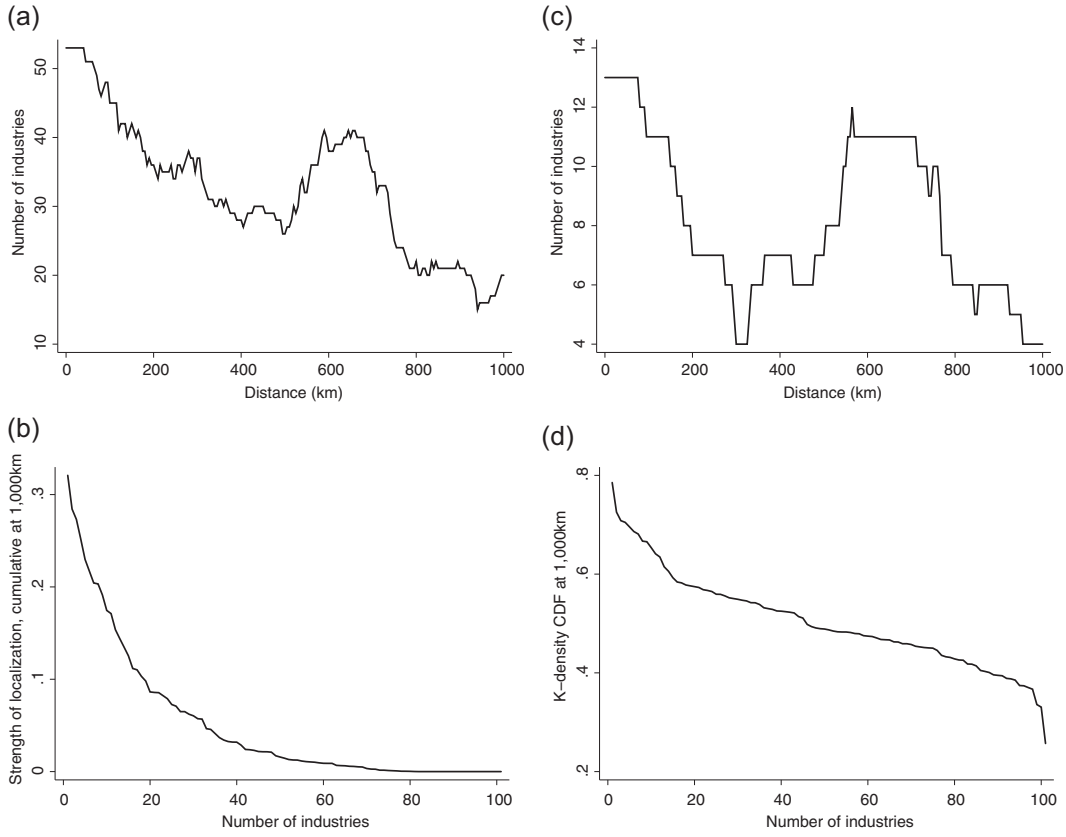

FIGURE 4 Localization patterns and strength of localization by distance, all of Russia: (a) Significant localization by distance; (b) Significant dispersion by distance; (c) Strength of localization Γi at 1,000 km; (d) CDF of K-density at 1,000 km. CDF: cumulative distribution function

Turning to studies of other countries, another point to note is the strong geographic concentration of textile- and clothing-related industries. This has been abundantly documented before for high-income countries like the United Kingdom (Duranton & Overman, 2005), the United States (Ellison, Glaeser, & Kerr, 2010), Japan (Nakajima, Saito, & Uesugi, 2012), Canada (Behrens, Boualam, & Martin, 2019; Behrens & Bougna, 2015), Germany (Riedel & Koh, 2014), and France (Barlet, Briant, & Crusson, 2013). Our results show that we also observe that concentration in transition countries like Russia. This suggests that agglomeration forces pushing towards geographic concentration are especially strong for those industries and do not depend substantially on the level of economic development or the institutional environment.

3.2.2 | Results for western and eastern Russia

We now report separate estimations for western and eastern parts of Russia. First, as shown in panels (b) and (c) of Table 1, geographic concentration patterns are stronger in the western part of Russia (88% of localized three-digit industries) than in the eastern part (66% of localized three-digit industries). The western part of Russia thus has more pronounced geographic concentration patterns, whereas the eastern part has a larger share of industries that are as good as randomly located—about one-quarter of all industries in the east.

Figure 5 depicts the number of significantly localized and dispersed industries by distance (Figure 5a,b), as well as the strength of localization and the *K*-density CDF (Figure 5c,d) for western Russia. Figure 6 provides the same information for eastern Russia. These two figures confirm that the overall degree of geographic concentration is

TABLE 3 Top 10 most localized and most geographically concentrated three-digit industries, all of Russia

OKVED	Industry name	
Top 10 most localized industries		Γ_i
172	Weaving manufacture	0.323
353	Manufacture of aircraft and spacecraft	0.284
223	Reproduction of recorded media	0.273
176	Manufacture of textile fabrics	0.252
244	Manufacture of pharmaceuticals	0.230
173	Finishing of textiles	0.217
362	Manufacture of jewelry, medals and related articles of precious metals and stones; manufacture of coins	0.204
171	Spinning of textile fibers	0.203
321	Manufacture of electronic and radio components, electrovacuum devices	0.191
343	Manufacture of parts and accessories for motor vehicles and their engines	0.175
Top 10 most geographically concentrated industries		CDF
172	Weaving manufacture	0.785
353	Manufacture of aircraft and spacecraft	0.725
176	Manufacture of textile fabrics	0.708
363	Manufacture of musical instruments	0.705
223	Reproduction of recorded media	0.696
171	Spinning of textile fibers	0.686
343	Manufacture of parts and accessories for motor vehicles and their engines	0.681
173	Finishing of textiles	0.667
321	Manufacture of electronic and radio components, electrovacuum devices	0.666
192	Manufacture of luggage, handbags and the like, saddlery and harness	0.654

Note. Γ_i and the K-density CDF are computed at 1,000 km distance. We hence measure localization and geographic concentration over the whole distance range that we compute the K-densities for.
CDF: cumulative distribution function.

stronger in the west than in the east. Observe that the "gradients" in panels (a) and (b) of Figures 5 and 6 are different. Whereas they are basically monotonic in the east, there are local peaks in the west. This reflects the presence of several large cities in the western part, but their relative absence in the east. The distributions of the strength of localization in panels (c) and (d) look, on the contrary, quite similar in both regions: There are only a few strongly localized industries, whereas most industries display less extreme geographic patterns. The same conclusion can be drawn from panel (d).

Finally, Tables 4 and 5 list the most strongly localized and geographically most concentrated industries in the east and in the west. While different kinds of publishing and recording, textile, and pharmaceutical industries make the list in the west, the industries in the east are notably different, including metal-related industries and motor vehicles. These differences are linked to different broad specialization patterns that reflect a mix of natural advantage and the legacy of the Soviet planned economy, and to different concentration patterns of populations in the less dense eastern part and the more dense western part.

3.3 | Coagglomeration: Methodology

To date, we have only analyzed the geographic concentration patterns of individual industries. However, recent research on the determinants of agglomeration and clusters has emphasized that the *coagglomeration patterns of industry pairs* convey valuable information as to the underlying agglomeration mechanisms (see Behrens, 2016; Ellison et al., 2010; Faggio, Silva, & Strange, 2017). We hence now estimate the coagglomeration patterns of all three-digit industry pairs. We then use these estimates to investigate in more detail how input–output linkages and transport costs correlate with the geographic concentration of industries.

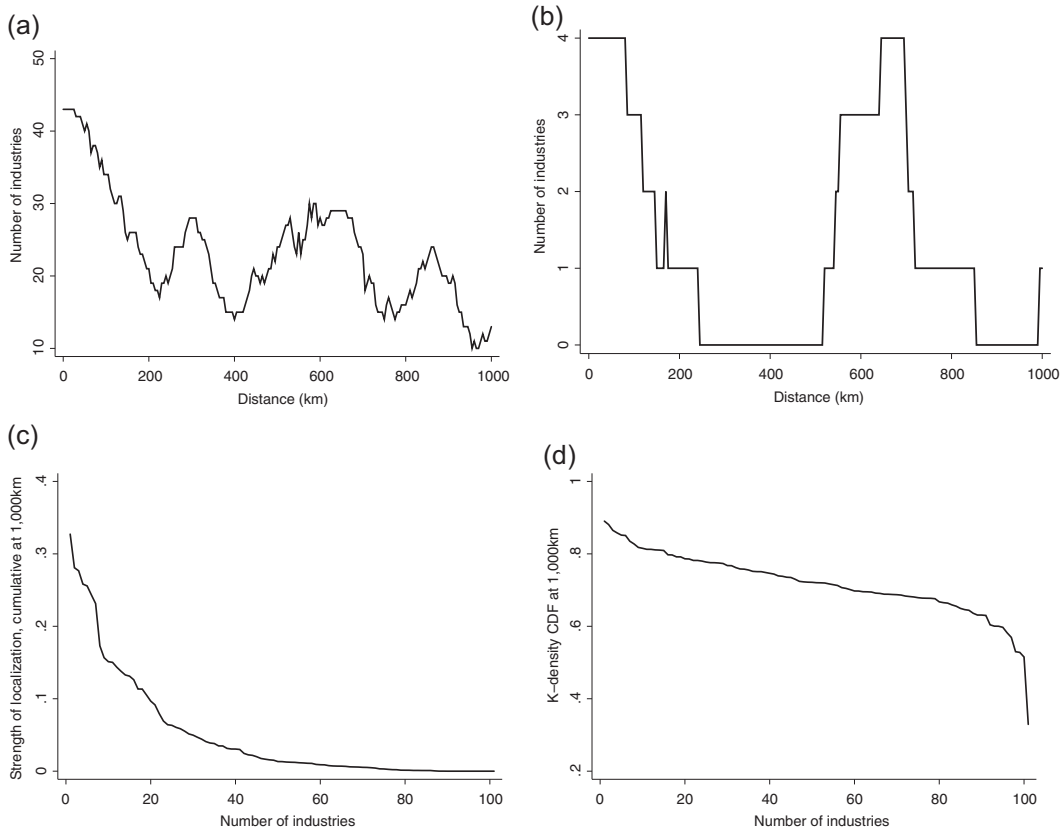

FIGURE 5 Localization patterns and strength of localization by distance, western Russia: (a) Significant localization by distance; (b) Significant dispersion by distance; (c) Strength of localization Γi at 1,000 km; (d) CDF of K-density at 1,000 km. CDF: cumulative distribution function

As shown by Duranton and Overman (2008), the methodology for estimating *K*-densities can be readily adapted to the coagglomeration of two different industries. As for the case of single industries in Section 3.1, we again estimate *K*-densities for the distribution of bilateral distances between manufacturing establishments. However, we now restrict these densities to pairs of establishments *in different industries*. In other words, we exclude the geographic concentration of the industry itself and only measure how that industry is located relative to some other industry. We relegate technical details of the procedure and the implementation to Appendix B.2.

Before proceeding, let us briefly pause to discuss the counterfactual distributions that we use to construct confidence bands for the coagglomeration of industries. As for the case of the agglomeration of single industries, we construct confidence bands by drawing 1,000 random samples of establishments. A key difference, however, is that we restrict the counterfactual to the locations that contain establishments of either of the two industries only. Put differently, we take the joint distribution of the establishments in the two industries as our benchmark. Hence, any departure from the counterfactual distribution measures how much closer establishments in the two industries are from establishments in their own industry rather than from establishments in the two industries in general. As explained by Duranton and Overman (2008), this is a stronger test for localization since the strength of coagglomeration—that is, the difference between the observed distribution and the counterfactual distribution—already controls for the agglomeration patterns of

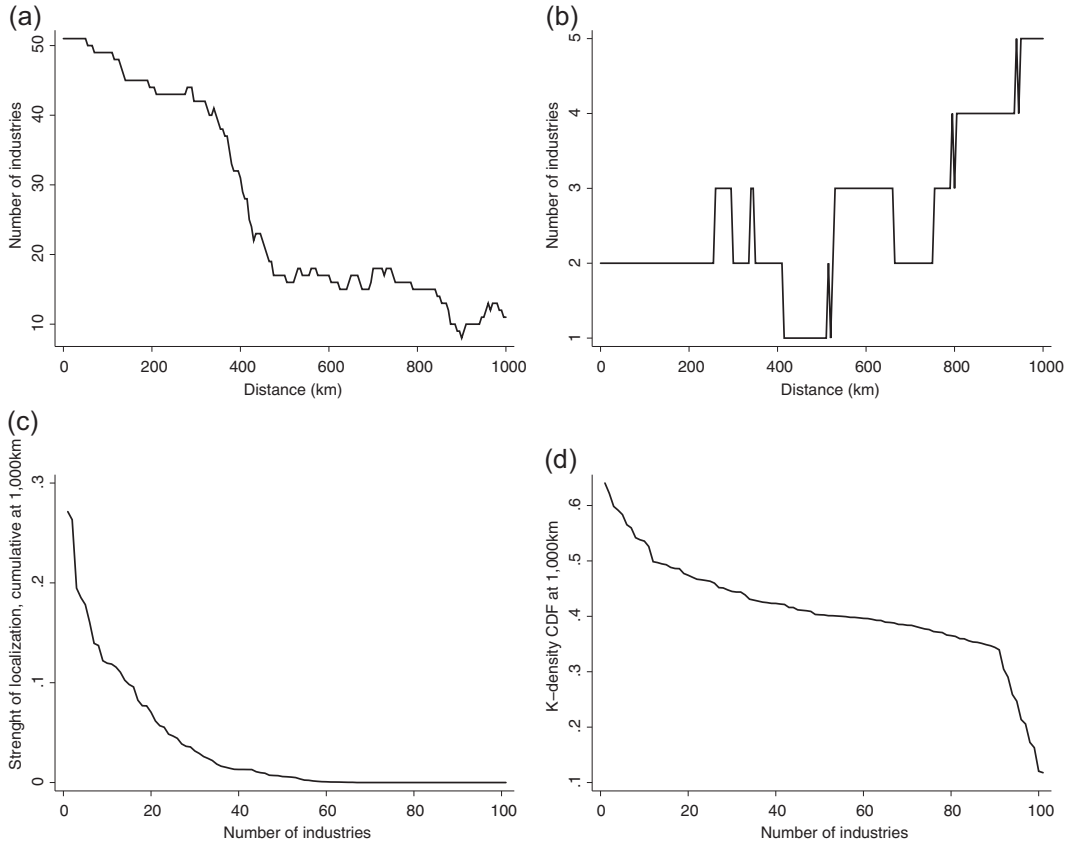

FIGURE 6 Localization patterns and strength of localization by distance, eastern Russia: (a) Significant localization by distance; (b) Significant dispersion by distance; (c) Strength of localization Γi at 1,000 km; (d) CDF of *K*-density at 1,000 km. CDF: cumulative distribution function

the two industries themselves.[6] A direct consequence of this is that some industry pairs can be strongly concentrated geographically, but not be significantly coagglomerated conditional on that geographic concentration. We provide an example below.

Our measure of absolute geographic concentration—the CDF of the estimated *K*-densities—retains the same interpretation as in the case of single industries. It measures the share of bilateral distances between pairs of plants in the two industries that is smaller than some given distance. Hence, larger values of that CDF correspond to more geographic concentration between the two industries.

3.4 | Coagglomeration: Results

We compute $(101 \times 100)/2 = 5,050$ distinct *K*-densities for our 101 three-digit industry pairs. Even when using the computationally more efficient Scholl–Brenner algorithm (Scholl & Brenner, 2015), this represents a large computational burden as some of our industries have more than 20,000 establishments. We hence only present results for all of Russia and no separate results for the eastern and western parts.

[6]Other choices are possible to construct the counterfactual benchmarks. One implication of our choice is that the estimated measures of excess concentration or dispersion of individual industries are not directly comparable to those of industry pairs. The reference distribution—the counterfactual—is different.

TABLE 4 Top 10 localized and concentrated industries (western Russia, three-digit OKVED)

OKVED	Industry name	
Top 10 most localized industries		Γ_i
223	Reproduction of recorded media	0.327
353	Manufacture of aircraft and spacecraft	0.281
244	Manufacture of pharmaceuticals	0.277
173	Finishing of textiles	0.258
172	Weaving manufacture	0.256
176	Manufacture of textile fabrics	0.243
362	Manufacture of jewelry, medals and related articles of precious metals and stones; manufacture of coins	0.231
335	Manufacture of watches and clocks and other time instruments	0.173
321	Manufacture of electronic and radio components, electrovacuum devices	0.157
171	Spinning of textile fibers	0.151
Top 10 most geographically concentrated industries		CDF
176	Manufacture of textile fabrics	0.890
172	Weaving manufacture	0.881
353	Manufacture of aircraft and spacecraft	0.865
335	Manufacture of watches and clocks and other time instruments	0.858
173	Finishing of textiles	0.852
223	Reproduction of recorded media	0.851
321	Manufacture of electronic and radio components, electrovacuum devices	0.835
363	Manufacture of musical instruments	0.827
343	Manufacture of parts and accessories for motor vehicles and their engines	0.818
244	Manufacture of pharmaceuticals	0.815

Note. Γ_i and the *K*-density CDF are computed at 1,000 km distance. We hence measure localization and geographic concentration over the whole distance range that we compute the *K*-densities for.
CDF: cumulative distribution function.

Figure 7 depicts four representative coagglomeration patterns of industry pairs. Figure 7a depicts "Publishing" (OKVED 221) and "Reproduction of recorded media" (OKVED 223). As shown, those industries are significantly coagglomerated, especially at short distances. They are thus found in the same places, for example, the same cities. Figure 7b depicts "Manufacture of other general purpose machinery" (OKVED 292) and "Manufacture of parts and accessories for motor vehicles and their engines" (OKVED 343). Those two industries are significantly codispersed at short distances, but coagglomerated at longer distances. They thus do not tend to significantly share the same locations but are found in different cities—either nearby cities at about 400 km, or far away ones at about 800–1,000 km. Figure 8a shows the corresponding *K*-density CDF for these two industries. As shown, the two industries are relatively dispersed geographically, which suggests that the coagglomeration at longer distances is essentially due to different regions specializing in these two industries, but with little geographic concentration at short distances. Figure 7d depicts "Processing and preserving of fish and fish products" (OKVED 152) and "Manufacture of other wearing apparel and accessories" (OKVED 182). As expected, those industries are codispersed across all distances, meaning that these industries tend to agglomerate into separate clusters.

Figures 7c and 8b are especially interesting. They depict the coagglomeration *K*-density and CDF of "Spinning of textile fibers" (OKVED 171) and "Weaving manufacture" (OKVED 172), respectively. As shown, these two industries are not significantly coagglomerated *conditional on the overall concentration of those two industries*. Indeed, the observed *K*-density falls into the 90% confidence band. Yet, the coagglomeration CDF of the industry pair 171–172 is significantly larger than the average or the median CDF across industry pairs, consistent with Figure 2b which shows that these two industries are both strongly concentrated geographically and also close to each other. In other

TABLE 5 Top 10 localized and concentrated industries (eastern Russia, three-digit OKVED)

OKVED	Industry name	
Top 10 most localized industries		Γ_i
341	Manufacture of motor vehicles	0.271
343	Manufacture of parts and accessories for motor vehicles and their engines	0.263
272	Manufacture of crude iron and steel pipes	0.195
342	Manufacture of bodies (coachwork) for motor vehicles; manufacture of trailers and semitrailers	0.185
271	Manufacture of crude iron, ferroalloy, steel	0.178
192	Manufacture of luggage, handbags and the like, saddlery and harness	0.160
275	Casting of metals	0.139
156	Manufacture of grain mill products, starches and starch products	0.137
154	Manufacture of vegetable and animal oils and fats	0.122
285	Treatment and coating of metals; general mechanical engineering	0.120
171	Spinning of textile fibers	0.151
Top 10 most geographically concentrated industries		CDF
192	Manufacture of luggage, handbags and the like, saddlery and harness	0.641
342	Manufacture of bodies (coachwork) for motor vehicles; manufacture of trailers and semitrailers	0.622
272	Manufacture of crude iron and steel pipes	0.598
343	Manufacture of parts and accessories for motor vehicles and their engines	0.592
275	Casting of metals	0.584
341	Manufacture of motor vehicles	0.565
156	Manufacture of grain mill products, starches and starch products	0.560
271	Manufacture of crude iron, ferroalloy, steel	0.542
154	Manufacture of vegetable and animal oils and fats	0.538
172	Weaving manufacture	0.536

Note. Γ_i and the K-density CDF are computed at 1000 km distance. We hence measure localization and geographic concentration over the whole distance range that we compute the K-densities for.
CDF: cumulative distribution function.

words, the pair 171–172 is strongly concentrated geographically because both industries are strongly concentrated and tend to locate in the same areas. However, conditional on this, the two industries are *not closer to each other than predicted by a random allocation*. This finding suggests that these industries may be attracted by unobserved local factors such as an adequate labor force or infrastructure. It also highlights that, as explained before, the significance test for coagglomeration is a fairly stringent one since it controls for the geographic concentration of the individual industries in the industry pair under consideration.

Panel (a) of Table 6 summarizes the numbers of significantly coagglomerated, significantly codispersed, and randomly located industry pairs for all of Russia. As can be seen from that table, a large share of industry pairs (more than 81%) are significantly coagglomerated. In other words, there is substantial cross-industry structure in the Russian agglomeration patterns. This information is useful and likely to allow us to better understand some of the underlying drivers of the (co)agglomeration of industries. We return to this point in Section 4.

Figure 9 shows the number of significantly coagglomerated industry pairs (Figure 9a) and the number of significantly codispersed industry pairs (Figure 9b) by distance. As shown, there is substantial coagglomeration at short distances and even more at around 600–650 km—recall that this corresponds roughly to the distance between Moscow and Saint Petersburg. This suggests that some industry pairs tend to cluster at short distances within major metro areas, whereas others tend to cluster separately in different metro areas. Panel (b) of Table 6 shows that many of the industry pairs that are significantly coagglomerated at short distances are different from those that are significantly coagglomerated at intermediate distances. Although there are a lot of industry pairs that are coagglomerated both at short and at long distances, the industrial tissues of Moscow and Saint Petersburg appear to be substantially different (the second spike in Figure 9a)—as panel (b) of Table 6 shows, more than 30% of industry pairs are coagglomerated on 0–170 km but not at long distance, meaning that they are not jointly overrepresented in both large cities but at short distances within those cities.

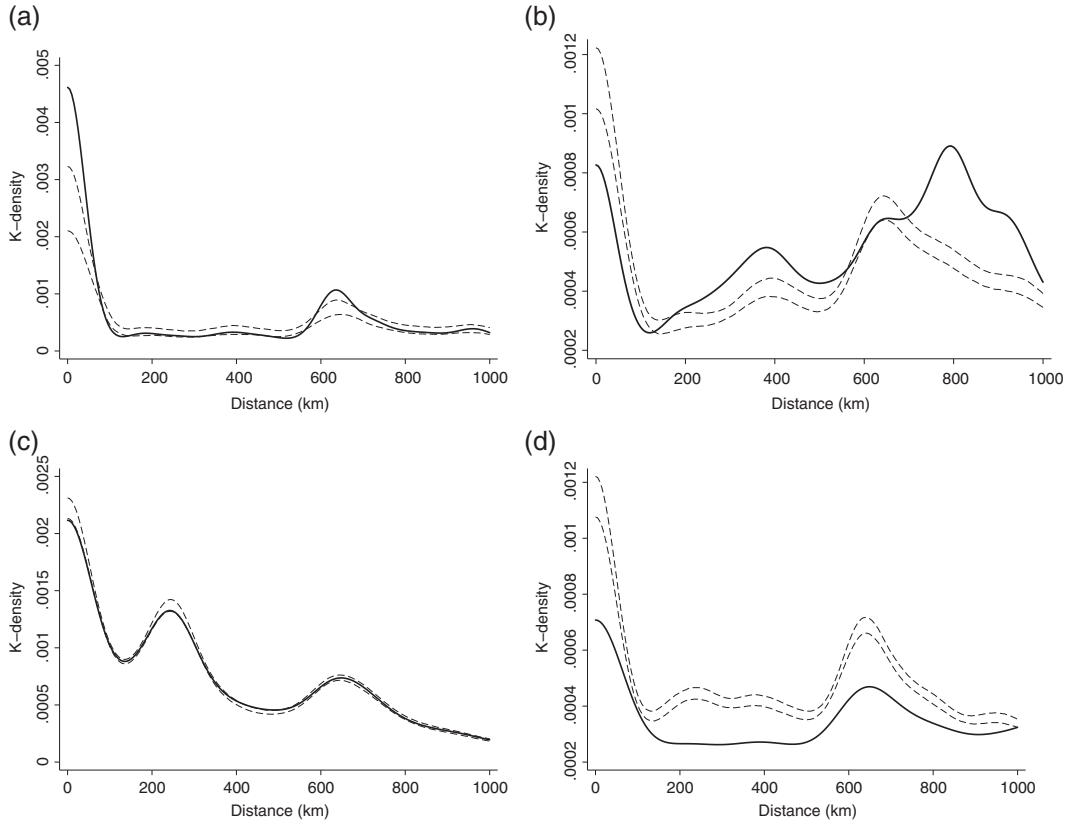

FIGURE 7 *K*-densities and confidence bands of selected OKVED three-digit industry pairs, all of Russia: (a) OKVED 221 and 223; (b) OKVED 292 and 343; (c) OKVED 171 and 172; (d) OKVED 152 and 182

Turning next to the strength of the coagglomeration and the skewness in its distribution, Figure 9c,d show that there are only a small number of highly localized industry pairs. For most industry pairs, the strength of coagglomeration is not very large. Note, however, that this result needs to be interpreted with caution. Indeed, as explained before, the strength of coagglomeration is measured *conditional on the geographic concentration of the two*

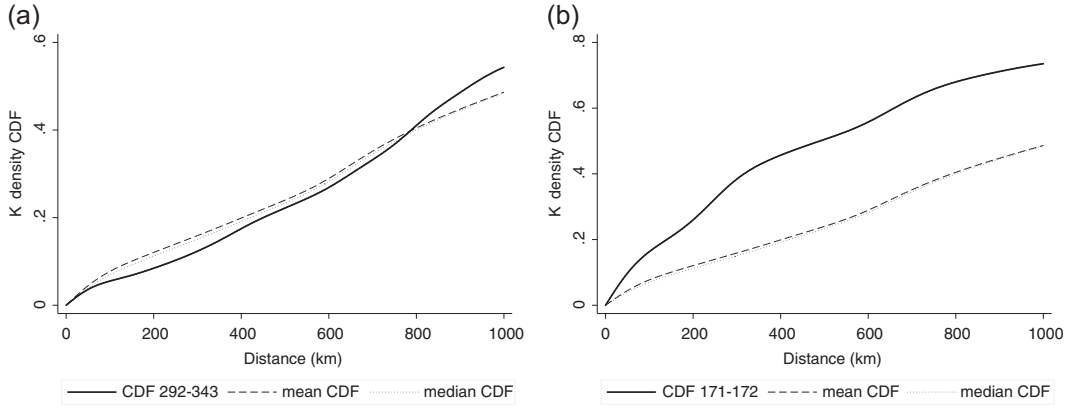

FIGURE 8 *K*-density CDFs for selected OKVED three-digit industries, and mean and median, all of Russia: (a) OKVED 292 and 343; (b) OKVED 171 and 172. CDF: cumulative distribution function

TABLE 6 Summary statistics of K-density estimates for coagglomeration patterns

	No. of industry pairs	Percentages
(a) Coagglomeration status		
Significantly coagglomerated	4,109	81.37
Random	269	5.33
Significantly codispersed	672	13.31
Total	5,050	100
$\overline{\Gamma}\vert_{\Gamma_i>0}$	0.012	
$\overline{\Psi}\vert_{\Psi_i>0}$	0.011	
(b) Type of coagglomeration		
Coagglomerated on 0–170 km, but not on 550–750 km	1,000	30.70
Coagglomerated on 550–750 km, but not on 0–170 km	725	22.26
Coagglomerated on 0–170 km and on 550–750 km	1,543	47.04

Note. All K-densities are computed for a range of 0–1,000 km for 5,050 three-digit industry pairs. The values of $\overline{\Gamma}\vert_{\Gamma_i>0}$ and $\overline{\Psi}\vert_{\Psi_i>0}$ are computed at the last point at which the K-densities are evaluated, that is, 1,000 km. We report average values for all significantly localized industry pairs in the case of $\overline{\Gamma}\vert_{\Gamma_i>0}$, and for all significantly dispersed industry pairs in the case of $\overline{\Psi}\vert_{\Psi_i>0}$. The bottom panel provides the breakdown of all coagglomerated industry pairs on 0–170 km and on 550–750 km.

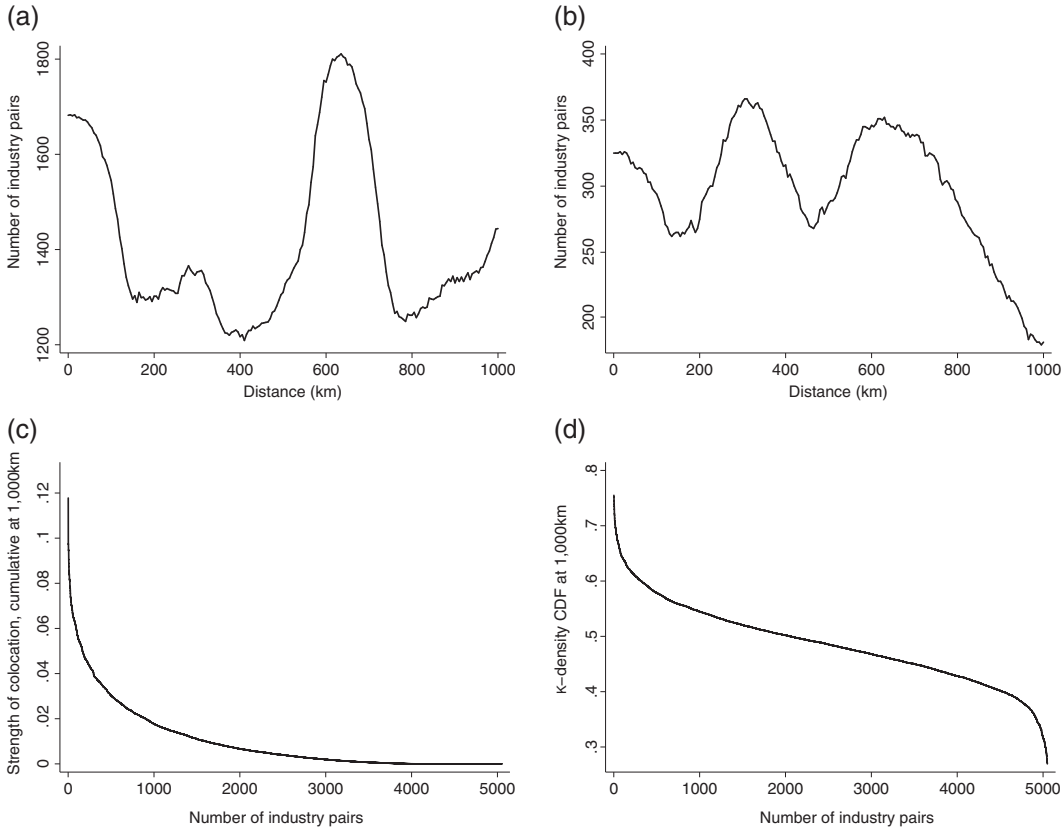

FIGURE 9 Coagglomeration patterns and strength of coagglomeration by distance, all of Russia: (a) Significant colocalization by distance; (b) Significant codispersion by distance; (c) Strength of colocalization Γi at 1,000 km; (d) CDF of K-density at 1,000 km. CDF: cumulative distribution function

industries. Controlling for that own-industry concentration can make two very strongly concentrated industries appear to be only weakly coagglomerated (or not at all; see Figure 7c for an illustration).

Finally, Table 7 summarizes the coagglomeration patterns by two-digit industries. Panel (a) reports the coagglomeration of three-digit industries (broken down by their two-digit industry) with other three-digit industries that do not belong to the same two-digit industry. Panel (b) reports the coagglomeration of the three-digit industries within the two-digit industry with other industries that do belong to the same two-digit industry. As can be seen, some industries display strong coagglomeration patterns within the same two-digit industry (e.g., "Publishing, printing and reproduction of recorded media"), whereas other industries are relatively codispersed (e.g., "Manufacture of coke, refined petroleum products and nuclear fuel"). As can be further seen, some industries also display strong coagglomeration patterns with most other industries that are not in the same two-digit industry. This might indicate industries that are very "urban," and which appear to be coagglomerated with most other "urban" industries. Note, finally, that the overall share of significantly coagglomerated industry pairs is roughly similar within and between two-digit industries, around 80%. Hence, coagglomeration patterns seem to be pervasive and to cut across most industrial boundaries.

4 | INPUT–OUTPUT LINKAGES, TRANSPORT COSTS, AND (CO)AGGLOMERATION

To date, we have documented that there are many localized and geographically concentrated industries in Russia. For some of those industries, especially in the west, the extent of geographic concentration is substantial. While the foregoing analysis is informative by itself, it tells us little about the underlying mechanisms. What are the potential drivers of agglomeration or coagglomeration in Russia? And are those drivers substantially different from those put forward in more developed countries where firms and industries operate in a different—and historically more market-oriented—institutional environment?

A long line of research on agglomeration has put forward a number of channels that may contribute to the agglomeration or coagglomeration of industries (see Combes & Gobillon, 2015; Rosenthal & Strange, 2004, for surveys). Industries may (co)agglomerate, among other things, because of buyer–supplier links, because they require specialized workers that can be "pooled" geographically, because they require access to localize natural advantages, because their goods are costly to transport and they want cheap access to markets, or because they exchange knowledge and information. In all of these cases, industries are close together because they benefit from that geographic proximity. Various empirical strategies have been put forward to substantiate causal evidence for these channels. One of them—pioneered by Ellison et al. (2010)—uses the revealed colocation patterns of industries to see if industries that are more connected along some dimensions (e.g., input–output linkages) tend to colocate more.[7] Since we have computed the colocation patterns of Russian industries, we can use them to learn more about the potential mechanisms that may generate them.

Doing so is, however, difficult for the case of Russia. The main reason is a lack of data to construct measures of industrial connectedness. This lack of data—and more generally the absence of work on the geographic concentration of industries in Russia—is one of the key reasons that explains why the drivers of geographic concentration have, to our knowledge, never been directly investigated for Russia to date. Furthermore, as already mentioned, the institutional environment in Russia is very different from that of other countries that have been the focus of research on agglomeration. For example, many location patterns in Russia are a legacy of the Soviet past

[7]Helsley and Strange (2014) show that coagglomeration patterns do not necessarily reflect beneficial agglomeration forces. However, numerical experiments performed by O'Sullivan and Strange (2018) suggest that this is, on average, the case. There is a growing empirical literature exploiting colocation patters (see, among others, Behrens & Brown, 2018; Behrens et al., 2018; Ellison et al., 2010; Faggio et al., 2017; Rosenthal & Strange, 2001). Combes and Gobillon (2015) provide a critical discussion of this approach and of its potential pitfalls and limitations.

TABLE 7 Coagglomeration patterns by broad industry groups

OKVED2 industries	Industry name	Number of three-digit industries in the two-digit sector that are coagglomerated with ... (a) Outside same two-digit				(b) Within same two-digit			
		Localized	Dispersed	Random	% localized	Localized	Dispersed	Random	% localized
15	Manufacture of food products and beverages	691	120	17	83.45	26	9	1	72.22
17	Textile manufacture	616	30	12	93.62	17	2	2	80.95
18	Manufacture of wearing apparel; dressing and dyeing of fur	219	56	19	74.49	3	0	0	100.00
19	Manufacturing of leather; leather articles and manufacture of footwear	248	29	17	84.35	3	0	0	100.0
20	Woodworking and manufacture of wood and cork articles, except furniture	416	51	13	86.67	9	1	0	90.00
21	Manufacture of cellulose, pulp, paper, cardboard and articles of these materials	145	36	17	73.23	0	0	1	0.00
22	Publishing, printing, and reproduction of recorded media	278	9	7	94.26	3	0	0	100.00
23	Manufacture of coke, refined petroleum products and nuclear fuel	113	153	28	38.44	0	3	0	0.00
24	Manufacture of chemicals and chemical products	544	68	46	82.67	19	0	2	90.48
25	Manufacture of rubber and plastic products	163	25	10	82.32	1	0	0	100.00
26	Manufacture of other nonmetallic mineral products	572	136	36	76.88	24	4	0	85.71
27	Manufacture of basic metals	381	61	38	79.37	9	0	1	90.00
28	Manufacture of fabricated metal products	516	119	23	78.42	12	9	0	57.14
29	Manufacture of machinery and equipment	520	87	51	79.03	16	1	4	76.19
31	Manufacture of electrical machinery and apparatus	482	68	20	84.56	13	2	0	86.67
32	Manufacture of radio, television and communication electronic components and apparatus	264	21	9	89.80	3	0	0	100.00
33	Manufacture of medical instruments, measure, control	388	75	17	72.65	8	1	1	80.00

(Continues)

TABLE 7 (Continued)

OKVED2 industries	Industry name	Number of three-digit industries in the two-digit sector that are coagglomerated with ... (a) Outside same two-digit Localized	Dispersed	Random	% localized	(b) Within same two-digit Localized	Dispersed	Random	% localized
	and test devices, optical devices, photo and cine equipment, watches								
34	Manufacture of motor vehicles, trailers and semitrailers	248	37	9	84.35	2	1	0	66.67
35	Manufacture of ships, aircraft and spacecraft and other transport	361	27	92	75.21	8	1	1	80.00
36	Manufacture of furniture	499	45	26	87.54	15	0	0	100.00
37	Recycling of secondary raw materials	170	23	5	85.86	1	0	0	100.00
Total		7,834	1,276	512	–	192	34	13	–

Note. There is a total of 10,100 (nonunique) industry pairs. The total in panel (a) includes the reciprocal pairs *ij* and *ji*; since we allocate the pair *ij* to the two-digit sector of *i* and the pair *ji* to the two-digit sector of *j*, they do not enter the computations symmetrically. The 239 industry pairs *ij* within the same two-digit sector are included in the figures reported in panel (b). In that case, we exclude these reciprocal pairs from the totals as they enter the computations symmetrically. Hence, the total number of reported pairs is 9,861 (panel (a)) plus 239 (panel (b)) plus the 239 reciprocal pairs excluded in panel (b).

where nonmarket considerations were key drivers of where industries were established (see, e.g., Kofanov & Mikhailova, 2015, for a more detailed discussion).

To make some progress on those questions, we proceed as follows. First, we use detailed information from Russian input–output tables to construct measures of the strength of interactions between industries (see Appendix C for details). Since input–output relationships are mainly technological—especially at a higher level of aggregation such as the three-digit industry level—we think that these relationships should be relatively independent from the precise institutional setting and operate in most market-oriented economies. Of course, we have to deal with reverse causality because the centrally planned location patterns inherited from the Soviet period may have been based on other considerations than input–output links. Since there is inertia in industry location, past location patterns may thus lead to less exchange in intermediate goods than what we would observe in a market outcome where industries pick locations based, in part, on minimizing the costs of accessing suppliers and customers. We deal with that reverse causality by providing instrumental variables estimates the instruments being input–output linkages as measured using Canadian and US data.[8] We show that these instruments are strong, in line with the view that input–output relationships are mainly technological. We also argue that our instruments are valid since they are exogenous to location patterns in Russia. Indeed, it is unlikely that the geographic concentration of industries in Russia is a driver of the Canadian or US input–output tables, thereby removing all problems of reverse causality. This instrumentation strategy is similar to that in Ellison et al. (2010) who instrument US industry-level characteristics using their UK counterparts. Overall, we find that there is a some downward bias in the estimates when using the Russian data. This suggests that input–output linkages are partly driven by past location patterns rather than location patterns being driven by input–output linkages.

Our second goal is to investigate how transport costs—another arguably technological parameter—relate to the agglomeration and coagglomeration of Russian industries. We do not have good measures for industry-level ad valorem transport costs in Russia (or, that being, for most other countries), so we will proxy them using detailed Canadian data developed previously by Behrens et al. (2018). While the general level of transport costs in Russia may be very different from that in Canada—wages, fuel prices, and the quality of infrastructure are quite different—what matters for our identification is the variation in transport costs between industries. We think that this is largely similar for broad industries across countries. Cement and gypsum products have a very high ad valorem transport cost (transport cost-to-price ratio), whereas gold has a very low ad valorem transport cost. This should be largely independent of the country, that is, it should be the case in both Canada and Russia. Hence, we think that our measures of Canadian ad valorem transport costs should also capture the relative costs of transporting different types of goods in Russia. Ideally, we would of course like to measure these costs directly using Russian data and then use the Canadian data as instrument. This is, unfortunately, not possible so that we will directly use the Canadian data as a proxy.

Last, we also run regressions where we use proxies for the other Marshallian agglomeration forces—labor market pooling and knowledge exchange—as additional controls: (a) the similarity of the workforce hired by the industries to capture thick local labor markets, and (b) the intensity with which patents originating in one industry—or being used by an industry—come from other industries. Ideally, we would again like to use Russian industry-level data. Unfortunately, we do not have them and there is little hope of getting them.[9] We thus use high-quality data from the United States as proxies. For that country, we have good measures of labor force similarity and patent citations across industries. Since these data are arguable less technological than either transport costs or input–output coefficients, we view these regressions as robustness checks, the main objective being to verify that our results using input–output data and transport costs are robust to the inclusion of these additional controls.

[8]We construct a crosswalk between the OKVED three-digit classification and the NAICS four-digit classification. Details concerning that crosswalk, as well as our data to construct input–output linkages, transport costs, and other controls (labor market pooling, knowledge exchange), are provided in Appendix C.

[9]To our knowledge, detailed data on the labor force composition across industries or patent citations do not exist for Russia. Labor force composition of industries is available only for "letters/characters," that is, the most aggregated level of OKVED (see http://www.gks.ru/bgd/regl/b15_11/IssWWW.exe/Stg/d01/06-04.htm). Concerning patent data, there exists only a coarse measure of the total number of patents by region. There is no information about patent citations across industries.

4.1 | Agglomeration of industries

We regress our measures of geographic concentration—both for individual industries and for industry pairs—on our measures of input–output linkages and ad valorem transport costs. We start with the geographic concentration of individual industries. The input–output linkages of an industry with itself are measured by the share of manufacturing input the industry buys from itself (hence sells to itself). The ad valorem transport costs are a measure of how expensive it is to ship a ton of its output relative to the price of the output. Appendix C provides additional details on the data. As explained before, we believe that input–output linkages and transport costs capture essentially technological relationships and we expect that they impact location patterns in Russia in similar ways than in other countries: High transport cost industries should be relatively dispersed to serve their demands, in line with Krugman (1991), whereas industries that buy a lot from themselves should be relatively agglomerated, in line with Krugman and Venables (1995).[10]

Table 8 summarizes the results of simple OLS regressions of geographic concentration on our industry-level input–output linkages and ad valorem transport costs. Three results stand out: (a) All coefficients for transport costs are negative and strongly significant, both for absolute and for relative geographic concentration; (b) the coefficients on input–output linkages, despite having positive point estimates, are not significant; and (c) the effects are stronger in the western part of Russia and weaker in the eastern part of Russia, but overall qualitatively quite similar. These findings are in line with findings for Canada (see Behrens & Brown, 2018; Behrens et al., 2018): Lower transport costs are associated with more geographic concentration (the "CDF" columns) and with more specialization (the "I_i" columns).

Our results for input–output linkages, however, are through the board insignificant.[11] This suggest that there is either no effect of input–output linkages on geographic concentration in Russia, or that our measures are imprecise and the samples too small to obtain statistically significant results. Combes and Gobillon (2015) argue that these two problems explain why this type of approach is unlikely to yield strong results. More fundamentally still, it is even unclear what input–output coefficients of an industry with itself precisely measure. For example, an industry that buys a lot from itself should, ceteris paribus, have more incentives to be spatially concentrated. However, this also depends on how much it buys from all other industries and how those other industries are located. The latter aspects are very hard to control for, and thus the ceteris paribus assumption is unlikely to hold. The relationship between input–output coefficients and the agglomeration of an industry is hence ambiguous and, combined with small sample sizes, unlikely to yield strong results (see also Rosenthal & Strange, 2001).

4.2 | Coagglomeration of industry pairs

As seen above, using only data for individual industries does not allow us to assess the role of input–output linkages for geographic concentration in Russia. To make progress, we hence now follow Ellison et al. (2010) and run regressions at the industry-pair level, using the coagglomeration K-density CDF as our dependent variable.[12]

To fix ideas, we begin with some descriptive correlations. Figure 10a depicts the unconditional correlation of the coagglomeration K-density CDF with the average input coefficient between each pair of industries. As

[10]The relationship between transport costs and geographic concentration is ambiguous in economic geography models. It depends, among other things, on how the dispersion forces are modeled. For example, in Helpman (1998), who uses a fixed supply of housing as a dispersion force, lower transport costs lead to the dispersion of economic activity. More generally, the relationship can be either non monotone, with lower transport costs leading first to more agglomeration and then to redispersion (see Baldwin, Forslid, Martin, Ottaviano, & Robert-Nicoud, 2003), or "discontinuous," with small changes in transport costs leading either to no change or to sudden "catastrophic" change. We focus on the geographic concentration of individual manufacturing industries, so we think that—conditional on controls for other agglomeration forces—dispersed demand plays a more important role than the inelastic supply of housing in explaining location patterns. Furthermore, we abstract from nonlinearities and run simple linear regressions. Our aim is not to estimate a structural economic geography model. See Behrens et al. (2018) for a more detailed discussion of these points.

[11]In unreported regressions, we also instrumented the Russian input coefficients with their Canadian and US counterparts. None of the IV estimates were significant. The point estimates increase slightly, but the standard errors increase too. We do not think that these regressions are informative.

[12]We do not run regressions using the strength of localization, I_{ij}, as dependent variable. Contrary to the geographic concentration of one industry—where the benchmark is the overall distribution of manufacturing—the benchmark for the coagglomeration measures is the joint distribution of both industries. Hence, I_{ij} is more difficult to use in regressions since it only measures the concentration of the industries above their own concentration.

TABLE 8 Transport costs, input–output linkages, and geographic concentration of Russian manufacturing industries

	All				West				East			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Dependent variables	CDF	CDF	Γ_i	Γ_i	CDF	CDF	Γ_i	Γ_i	CDF	CDF	Γ_i	Γ_i
Own-industry input coefficient (Russian data)		0.105		0.106		0.138		0.148		0.020		0.082
		(0.132)		(0.140)		(0.136)		(0.150)		(0.120)		(0.113)
Ad valorem transport cost	−0.270*	−0.286*	−0.204**	−0.219**	−0.274*	−0.294*	−0.208**	−0.228**	−0.167**	−0.176**	−0.195**	−0.207**
	(0.092)	(0.093)	(0.090)	(0.095)	(0.097)	(0.095)	(0.096)	(0.099)	(0.064)	(0.069)	(0.078)	(0.081)
Observations	93	92	93	92	93	92	93	92	93	92	93	92
R^2	0.078	0.090	0.049	0.060	0.085	0.104	0.059	0.080	0.040	0.040	0.042	0.049

Note. Robust standard errors in parentheses are clustered by NAICS four-digit industries. The dependent variables are either the DO K-density CDF or Γ_i at 100 km distance. All variables are standardized. See Appendix C for information on transport costs and the input–output variables. All regressions include a "quality" dummy, which takes value 1 for industry matches that are not clearly one-to-one, and 0 otherwise. This dummy is never significant, and regressions excluding those industries yield similar results.

CDF: cumulative distribution function.

*, **Significant at 1% and 5%, respectively.

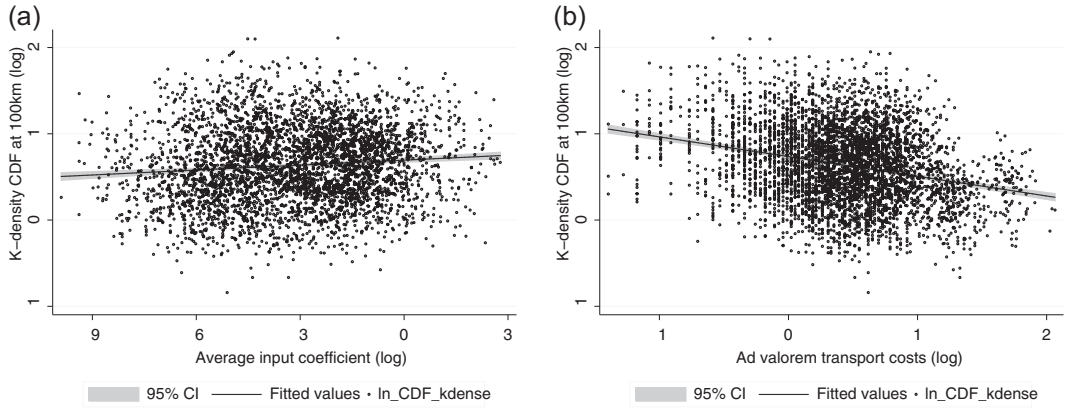

FIGURE 10 Coagglomeration patterns and ad valorem transport costs: (a) Input coefficients; (b) Transport costs. CDF: cumulative distribution function; CI: confidence interval

shown, there is a positive correlation: Industry pairs with stronger buyer–supplier links tend to be more coagglomerated. Figure 10b depicts the unconditional correlation of the coagglomeration *K*-density CDF with our measure of ad valorem transport costs at the industry-pair level. We measure the latter as the average of the ad valorem transport costs of the two industries. As shown, there is a negative correlation, that is, industry pairs with higher average transport costs tend to be less coagglomerated than industry pairs with lower average transport costs.

Table 9 summarizes our baseline results for input coefficients and transport costs. Column (1) shows that, in line with Figure 10, there is a positive relationship between geographic concentration and the strength of buyer–supplier links. The coefficient is, however, fairly small and not precisely measured. Column (2) shows that there is a strong negative association between transport costs of the two industries and their coagglomeration patterns. Industries whose products are relatively expensive to trade tend to be less coagglomerated. The reason may be that these industries are relatively more dispersed—in line with our results in the foregoing section—and may thus interact less. These result hold also if we measure our pairwise variables using the maximum of the coefficients instead of the average (see Columns (6) and (7)). The remaining columns in Table 9 provide additional robustness checks. In particular, our results do not change when including both input–output links and transport costs. They are also robust to the exclusion of industry pairs within the same three-digit industry. In the latter case, the coefficients for input–output links increase slightly, but the overall results are similar.

The regressions in Table 9 suffer from two potential shortcomings: endogeneity and measurement error.[13] Both can be addressed using two-stage least squares instrumental variables (2SLS IV) estimations. We proceed as follows. First, we construct instruments for the Russian input coefficients. As argued above, we believe that input coefficients largely reflect technological relationships. Hence, the input coefficients from other countries should provide relevant and valid instruments. They are relevant because the technological relationships at the three-digit level are fairly stable across countries. This is shown by Table 10, which summarizes the raw correlations between the Russian, Canadian, and US average input coefficients (and ad valorem transport costs). As shown, the correlations between the different input coefficients are high, thus suggesting that the first stage of the 2SLS IV regressions will be strong.

[13]There are two additional problems: (a) There can be unobserved fundamentals that drive the agglomeration or coagglomeration of industries, without any agglomeration forces, and (b) there can be spurious agglomeration or coagglomeration, which is not based on efficiency considerations. These two points are very hard to control for in this type of analysis (see Ellison et al., 2010). They require us to have an idea of the counterfactual distribution that would be observed in the absence of agglomeration forces (see, e.g., Carillo & Rothbaum, 2016; Klier & McMillen, 2008, for different approaches).

TABLE 9 Baseline regressions for the geographic concentration of Russian industry pairs

	Average					Maximum				
Sample	(1)	(2)	Exclude3 (3)	(4)	Exclude3 (5)	(6)	(7)	Exclude3 (8)	(9)	Exclude3 (10)
Input coefficient (Russia)	0.019			0.020	0.034***	0.015			0.011	0.027
	(0.016)			(0.015)	(0.018)	(0.016)			(0.015)	(0.018)
Ad valorem transport cost		−0.247*	−0.246*	−0.247*	−0.245*		−0.215*	−0.212*	−0.215*	−0.211*
		(0.017)	(0.018)	(0.017)	(0.018)		(0.017)	(0.017)	(0.017)	(0.017)
Observations	4,242	4,242	4,053	4,242	4,053	4,242	4,242	4,053	4,242	4,053
R^2	0.001	0.061	0.060	0.061	0.060	0.000	0.046	0.045	0.046	0.045

Note. Our mapping from OKVED to NAICS is not one-to-one, so some different OKVED industry pairs may be associated with the same NAICS industry pairs. To control for that, we report robust standard errors (in parentheses) that we cluster by NAICS four-digit pairs. The dependent variables is the DO coagglomeration K-density at 100 km distance. All variables are standardized. "Exclude3" regressions exclude all four-digit industry pairs within the same three-digit industries. All regressions include our "quality flag." See Appendix C for a detailed description of our data.

*, ***Significant at 1% and 10%, respectively.

TABLE 10 Correlations of the input coefficients and ad valorem transport costs

	Average input coefficient			AV transport cost
	Russia	Canada	United States	
Russia	1.000			
Canada	0.445*	1.000		
United States	0.475*	0.647*	1.000	
AV transport cost	−0.009	0.015	0.011	1.000

Note. See Appendix C for additional details on our data.
*Significant at the 1% level.

Our instruments are valid if the underlying unobserved causes that drive the coagglomeration of industries in Russia are not the same than those that drive the coagglomeration of industries in the countries we use as instruments. Given their geographic distance and fairly insignificant trade relationships, as well as their geographic, historical, and institutional differences, it is very unlikely that the Canadian or US instruments suffer from the same endogeneity problems than the Russian input coefficients. Using detailed Canadian and US input–output data to instrument the Russian variables has also a second advantage. The Canadian and US tables are likely to be better measured than the Russian table, thereby also correcting potential measurement error in our data.

Table 11 summarizes our 2SLS IV results. The bottom panel of the table shows that both the Canadian and the US instruments are strong and highly relevant, with good first stages. Both instruments separately add to the first stage, which justifies their joint inclusion. Having more than one instrument also allows us to run an overidentification test. As shown by the Hansen *J*-statistic, our instruments pass that test. The top panel reveals that the instrumented input coefficients are significant in half of the specifications, and that their estimated values increase, especially when excluding industry pairs in the same three-digit industries. Overall, our results suggests that there may be a downward bias when using Russian data to estimate the impact of input–output linkages on the coagglomeration patterns of industries. One explanation could be that industry location in Russia was largely determined in the Soviet era by considerations other than those of a market economy. Hence, industry pairs that are only weakly linked may have ended up in the same places and, therefore, bias the input coefficients towards zero. Last, the coefficient on transport costs is largely unaffected in the IV specification, with values that are close to those in our baseline regressions.

4.3 | Additional results and robustness

Table 12 presents additional results where we include proxies to control for industries' similarity in terms of the types of workers they hire ("Occupational employment similarity" correlation) and potential knowledge spillovers between them as measured by patent citations (Patent citations, use-based). Both have been highlighted as potentially important sources of agglomeration economies in the literature. Since we do not have these variables for Russia, we proxy them with US data (see Appendix C for details).

Table 12 shows that our main results are unchanged: Industries with stronger input–output linkages are more coagglomerated and industries with higher transport costs are less coagglomerated. We further find that industries that exchange more knowledge are substantially more coagglomerated. These results are robust and in line with what has been found for other countries like Canada and the United States. However, as can also be seen from Table 12, industries that hire more similar workers appear to be less coagglomerated in Russia.

The latter result differs starkly from what has been found previously in the literature, where labor market similarity is one of the strongest predictors of coagglomeration (see Ellison et al., 2010; Faggio et al., 2017). One

TABLE 11 2SLS instrumental variable regressions, using Canadian and US instruments

	Average			Maximum		
Samples	(1)	(2)	Exclude3 (3)	(4)	(5)	Exclude3 (6)
	Second-stage results					
Average input coefficient (Russia)	0.045	0.064***	0.122*			
	(0.036)	(0.034)	(0.042)			
Average AV transport cost		−0.248*	−0.243*			
		(0.017)	(0.018)			
Maximum input coefficient (Russia)				0.048	0.045	0.103**
				(0.035)	(0.033)	(0.043)
Maximum AV transport cost					−0.214*	−0.208*
					(0.017)	(0.017)
Observations	4,242	4,242	4,053	4,242	4,242	4,053
	First-stage results					
Average input coefficient (Canada)	0.337*	0.337*	0.449*			
	(0.064)	(0.064)	(0.071)			
Average input coefficient (United States)	0.415*	0.416*	0.318*			
	(0.077)	(0.077)	(0.072)			
Maximum input coefficient (Canada)				0.278*	0.277*	0.373*
				(0.062)	(0.063)	(0.064)
Maximum input coefficient (United States)				0.361*	0.361*	0.265*
				(0.070)	(0.070)	(0.061)
Observations	4,242	4,242	4,053	4,242	4,242	4,053
Hansen-*J* (*p*-value)	0.557	0.660	0.132	0.290	0.412	0.205
F-statistic	49.16	48.93	43.18	36.23	36.12	36.33
R^2	0.355	0.356	0.268	0.324	0.325	0.253

Note. We only report the key coefficients in the first stage and do not report the others. Our mapping from OKVED to NAICS is not one-to-one, so some different OKVED industry pairs may be associated with the same NAICS industry pairs. To control for that we report robust standard errors (in parentheses) that we cluster by NAICS four-digit pairs. The dependent variable is the DO coagglomeration *K*-density at 100 km distance. All variables are standardized. "Exclude3" regressions exclude all four-digit industry pairs within the same three-digit industries. All regressions include our "quality flag." See Appendix C for a detailed description of our data.

2SLS: two-stage least squares; AV: ad valorem.

*, **, ***Significant at 1%, 5%, and 10%, respectively.

possible explanation may lie in the fact that we use a proxy constructed from US data that works poorly for Russia. Another possible explanation lies in the low mobility of workers between firms and regions (see, e.g., Guriev & Vakulenko, 2015, for evidence on "geographic poverty traps" in Russia). Especially low-skilled workers—the bulk of the workforce in Russian manufacturing—are not mobile: Low salaries, combined with a substantial demand for

TABLE 12 Baseline and instrumental variable regressions, additional controls

	Average				Maximum			
Sample/method	(1)	Exclude3 (2)	IV (3)	IV Exclude3 (4)	(5)	Exclude3 (6)	IV (7)	IV Exclude3 (8)
	Second-stage results							
Average input coefficient (Russia)	0.029**	0.038**	0.091*	0.141*				
	(0.015)	(0.018)	(0.035)	(0.042)				
Average AV transport cost	−0.228*	−0.230*	−0.229*	−0.228*				
	(0.017)	(0.018)	(0.017)	(0.018)				
Maximum input coefficient (Russia)					0.020	0.031***	0.073**	0.123*
					(0.014)	(0.017)	(0.033)	(0.043)
Maximum AV transport cost					−0.199*	−0.199*	−0.199*	−0.196*
					(0.017)	(0.017)	(0.017)	(0.017)
OES correlation	−0.073*	−0.059**	−0.084*	−0.066*	−0.076*	−0.062**	−0.084*	−0.067*
	(0.023)	(0.025)	(0.024)	(0.025)	(0.023)	(0.025)	(0.024)	(0.025)
Patent citations (use-based)	0.322*	0.274*	0.327*	0.277*	0.343*	0.294*	0.348*	0.297*
	(0.109)	(0.103)	(0.109)	(0.102)	(0.113)	(0.106)	(0.113)	(0.106)
Observations	4,242	4,053	4,242	4,053	4,242	4,053	4,242	4,053
	First-stage results							
Average input coefficient (Canada)			0.342*	0.452*				
			(0.066)	(0.072)				
Average input coefficient (United States)			0.416*	0.319*				
			(0.077)	(0.072)				
Maximum input coefficient (Canada)							0.279*	0.374*
							(0.064)	(0.064)
Maximum input coefficient (United States)							0.361*	0.265*
							(0.070)	(0.061)
Hansen-*J* (*p*-value)			0.508	0.101			0.314	0.159
F-statistic			46.42	41.96			34.61	35.49
R^2			0.356	0.268			0.325	0.253

Note. We only report the key coefficients in the first stage and do not report the others. Our mapping from OKVED to NAICS is not one-to-one, so some different OKVED industry pairs may be associated with the same NAICS industry pairs. To control for that, we report robust standard errors (in parentheses) that we cluster by NAICS four-digit pairs. The dependent variable is the DO coagglomeration *K*-density at 100 km distance. All variables are standardized. "Exclude3" regressions exclude all four-digit industry pairs within the same three-digit industries. All regressions include our "quality flag." See Appendix C for a detailed description of our data.

AV: ad valorem; IV: instrumental variable; OES: Occupational Employment Survey.

*, **, ***Significant at 1%, 5%, and 10%, respectively.

TABLE 13 OLS and 2SLS IV regressions, with interaction terms and in log form

	Base (log)				Interactions (log)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sample/method				IV	Interact	Interact IV	Interact	Interact IV
Second stage								
Average input coefficient (Russia, log)	0.019*		0.014*	−0.006	0.012**	−0.030***	0.005	−0.023
	(0.004)		(0.004)	(0.012)	(0.005)	(0.017)	(0.006)	(0.021)
Average AV transport cost (log)		−0.230*	−0.225*	−0.213*	−0.211*	−0.108*	−0.096*	0.022
		(0.016)	(0.017)	(0.022)	(0.024)	(0.048)	(0.027)	(0.052)
Average AVTC (log) × average input (Russia, log)					0.005	0.049**	0.009	0.067*
					(0.007)	(0.021)	(0.008)	(0.023)
OES correlation (log)							−0.136*	−0.182*
							(0.016)	(0.021)
Patent citations (use-based, log)							0.080*	0.086*
							(0.007)	(0.009)
Observations	3,906	4,242	3,906	2,565	3,906	2,565	3,080	2,082
First stage 1 (average input coefficient)								
Average input coefficient (Canada, log)				0.545*		0.491*		0.471*
				(0.047)		(0.057)		(0.067)
Average input coefficient (United States, log)				0.196*		0.161*		0.180*
				(0.028)		(0.034)		(0.038)
				(0.087)		(0.124)		(0.144)
Average AVTC (log) × average input (Canada, log)						0.139**		0.085
						(0.066)		(0.072)
Average AVTC (log) × average input (United States, log)						0.071		0.079
						(0.045)		(0.051)
F-statistic				217.45		130.34		79.80
Hansen-*J* (*p*-value)				0.122		0.133		0.612
R^2				0.274		0.280		0.260
×*First stage 2 (average input AVTC)*								
Average input coefficient (Canada, log)						0.007		−0.006
						(0.028)		(0.035)
Average input coefficient (United States, log)						0.013		0.017
						(0.020)		(0.023)
Average AVTC (log) × average						0.581*		0.573*

(Continues)

TABLE 13 (Continued)

	Base (log)				Interactions (log)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sample/method				IV	Interact	Interact IV	Interact	Interact IV
input (Canada, log)								
						(0.053)		(0.059)
Average AVTC (log) × average input (United States, log)						0.219*		0.239*
						(0.038)		(0.045)
Hansen-*J* (*p*-value)						0.133		0.612
F-statistic						120.43		86.47
R^2						0.611		0.600

Note. We only report the key coefficients in the first stage and do not report the others. Our mapping from OKVED to NAICS is not one-to-one, so some different OKVED industry pairs may be associated with the same NAICS industry pairs. To control for that, we report robust standard errors (in parentheses) that we cluster by NAICS four-digit pairs. The dependent variable is the DO coagglomeration *K*-density at 100 km distance. All variables are taken in logs. In the IV regressions, the interaction term is instrumented using the Canadian and US input–output shares. All regressions include our "quality flag." See Appendix C for a detailed description of our data.

2SLS: two-stage least squares; AV: ad valorem; AVTC: ad valorem transport costs; IV: instrumental variable; OES: Occupational Employment Survey; OLS: ordinary least squares.

*, **, ***Significant at 1%, 5%, and 10%, respectively.

cheap labor in manufacturing, does not stimulate labor mobility or investment in human capital.[14] Another peculiarity of Russian institutions lies in the mismatch between the education system and the demand for manual professions. Employers solve this problem by providing firm-specific training right in the workplace. Since workers are contractually bound to firms in return for training—and since the training is firm-specific and therefore not easily portable—this impedes the mobility between firms and reduces the need to be close to other employers requiring similar labor types.[15] Hence, many of the labor pooling aspects of agglomeration economies may be inoperational in Russia because of the institutional setting.

As an additional robustness check, we also run log-log regressions. Columns (1)–(3) of Table 13 replicate some of our baseline regressions from Table 9. As shown, the results are qualitatively identical and quantitatively very similar; the estimated coefficients on our key variables are all significant and fairly precisely estimated. However, the IV regression in Column (4) yields a different results from before: The coefficient on the input–output links becomes insignificant. This may be due to the fact that we have a substantial number of zeros in the US input–output table, which leads to the loss of a large number of observations when using logs in the instrumentation.

Finally, we extend the model and add the interaction term between transport costs and input coefficients in Columns (5)–(8) of Table 13 as in Behrens and Brown (2018). The idea is that high transport costs should be more important drivers of coagglomeration if industries buy and sell a lot from each other. As argued by Combes and Gobillon

[14]For example, the average wage across all sectors in Russia was 34,029.5 rubles in 2015. In manufacturing, the average was 31,910.2 rubles, while in textile manufacturing it was only 15,757.6 rubles. See http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/wages/labour_costs//# for more information. One explanation for the low manufacturing wages may be the weak bargaining power of trade unions (see, e.g., Lukiyanova & Vishnevskaya, 2016). Another may be simply the weak manufacturing productivity itself, which should translate into low wages.

[15]This may be problematic for economic performance for several reasons. First, it is known that whether human capital investments lead to more or less geographic concentration depends on whether they are industry- or firm-specific (see Matouschek & Robert-Nicoud, 2005). The presence of monopsony power on the employer side—due to less coagglomeration of industries that require similar workers—reduces wages and stifles workers' investment decisions to acquire general human capital. Second, the positive benefits of insurance against idiosyncratic shocks are lost. Last, knowledge exchange and innovation due to the rapid mobility of workers across jobs and firms ("job hopping"; see, e.g., Fallick, Fleischman, & Rebitzer, 2006) are also foregone.

(2015, p. 334), this should be done since "according to theory, two industries sharing inputs have more incentive to colocate when trade costs for these inputs are large. In that perspective, variables capturing input–output linkages should be caused to interact with a measure of trade costs, but this is not done in the literature." Our IV results in Table 13 show that industries that buy a lot from each other—as measured by their average input coefficient—tend indeed to be more coagglomerated if transporting their output is relatively expensive. We do not find this effect in the OLS regressions, which suggests again that there may be a downward bias when using Russian data. We do not want to read too much out of these results since they are not robust to using a nonlogged specification.[16]

5 | CONCLUSIONS

Using disaggregated microgeographic data, we paint a detailed picture of geographic concentration patterns of manufacturing industries in Russia. We also provide some evidence for the importance of input–output linkages and ad valorem transport costs for the agglomeration of individual industries and the coagglomeration of industry pairs. Our results show that the geographic patterns of Russian manufacturing are broadly comparable to those that have been documented for other countries: There are many localized industries and industry pairs, and a few very strongly concentrated ones. The latter—which include, for example, textile-related industries—are not very different from those substantiated for other countries, thus pointing to the robustness of these findings. In a nutshell, the prevalence and magnitude of geographic concentration in Russia are largely comparable to those in more developed economies.

Furthermore, the role of input–output linkages and transport costs seems to also be in line with that established in other countries: Stronger buyer–supplier links and lower transport costs are associated with more geographic concentration. This result is robust to a battery of checks and instrumental variables estimation. We find a small downward bias in the coefficient for input–output linkages, in line with the fact that historically inherited location patterns may be due to other considerations than buyer–supplier relationships between industries: The geographic concentration patterns in the 1990s were largely a legacy of the Soviet past, and this may lead to a downward bias when using Russian input–output data.

While most of our results are in line with previous findings, one substantive difference is that, compared to other countries, industries with a more similar workforce appear to be less coagglomerated. This may be due to our indirect measure of this variable or to differences in the institutional environment of the labor market. We cannot disentangle these two potential causes so that more work is called for here.

ACKNOWLEDGMENTS

Behrens gratefully acknowledges financial support from the CRC Program of the Social Sciences and Humanities Research Council (SSHRC) of Canada for the funding of the *Canada Research Chair in Regional Impacts of Globalization*. The article was prepared within the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project "5-100." The views expressed in this paper and any remaining errors are ours.

ORCID

Kristian Behrens http://orcid.org/0000-0003-3595-6443

[16]When using the standardized (non logged) variables, the point estimates on the interaction terms are positive but insignificant. In the nonlogged version, the correlation of the transport cost and the interaction term is very high, and it drops in the logged specification. Collinearity hence seems less of an issue in the log-log specification.

REFERENCES

Aleksandrova, E., Behrens, K., & Kuznetsova, M. (2018). *Manufacturing (co)agglomeration in a transition country: Evidence from Russia* (HSE Working Paper No. WP BRP 186/EC/2018). National Research University Higher School of Economics. https://publications.hse.ru/mirror/pubs/share/direct/218658802

Baldwin, R., Forslid, R., Martin, P., Ottaviano, G. I. P., & Robert-Nicoud, F. L. (2003). *Economic geography and public policy*. Princeton, NJ: Princeton University Press.

Barlet, M., Briant, A., & Crusson, L. (2013). Location patterns of service industries in France: A distance-based approach. *Regional Science and Urban Economics*, *43*(2), 338–351.

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

Behrens, K., Boualam, B., & Martin, J. (2019). Are clusters resilient? Evidence from Canadian textile industries. *Journal of Economic Geography*. Forthcoming

Behrens, K., & Bougna, T. (2015). An anatomy of the geographical concentration of Canadian manufacturing industries. *Regional Science and Urban Economics*, *51*(C), 47–69.

Behrens, K., Bougna, T., & Brown, M. W. (2018). The world in not yet flat: Transport costs matter! *Review of Economics and Statistics*, *100*(4), 712–724.

Behrens, K., & Brown, M. W. (2018). Transport costs, trade, and geographic concentration: Evidence from Canada. In B. A. Blonigen, & W. W. Wilson (Eds.), *Handbook of international trade and transportation* (pp. 188–235). Cheltenham, UK, Northampton, MA: Edward Elgar Publishing Inc.

Brown, W. M. (2015). How much thicker is the Canada-US border? The cost of crossing the border by truck in the pre- and post-9/11 eras. *Research in Transportation Business and Management*, *16*, 50–66.

Carillo, P. E., & Rothbaum, J. L. (2016). Counterfactual spatial distributions. *Journal of Regional Science*, *56*(5), 868–894.

Combes, P.-P., & Gobillon, L. (2015). The empirics of agglomeration economies. In G. Duranton, J. V. Henderson, & W. C. Strange (Eds.), *Handbook of regional and urban economics* 5, pp. 247–348). Amsterdam: Elsevier.

Duranton, G. (2015). Growing through cities in developing countries. *The World Bank Research Observer*, *30*(1), 39–73.

Duranton, G., & Overman, H. G. (2008). Exploring the detailed location patterns of U.K. manufacturing industries using microgeographic data. *Journal of Regional Science*, *48*(1), 213–243.

Duranton, G., & Overman, H. G. (2005). Testing for localization using micro-geographic data. *Review of Economic Studies*, *72*(4), 1077–1106.

Ellison, G. D., Glaeser, E. L., & Kerr, W. R. (2010). What causes industry agglomeration? Evidence from coagglomeration patterns. *American Economic Review*, *100*(3), 1195–1213.

Ellison, G. D., & Glaeser, E. L. (1997). Geographic concentration in US manufacturing industries: A dartboard approach. *Journal of Political Economy*, *105*(5), 889–927.

Faggio, G., Silva, O., & Strange, W. C. (2017). Heterogeneous agglomeration. *Review of Economics and Statistics*, *99*(1), 80–94.

Fallick, B., Fleischman, C. A., & Rebitzer, J. B. (2006). Job-hopping in Silicon Valley: Some evidence concerning the microfoundations of a high-technology cluster. *Review of Economics and Statistics*, *88*(3), 472–481.

Guriev, S., & Vakulenko, E. (2015). Breaking out of poverty traps: Internal migration and interregional convergence in Russia. *Journal of Comparative Economics*, *43*(3), 633–649.

Helpman, E. (1998). The size of regions. In D. Pines, E. Sadka, & I. Zilcha (Eds.), *Topics in public economics. Theoretical and empirical analysis* (pp. 33–54). Cambridge: Cambridge University Press.

Helsley, R. W., & Strange, W. C. (2014). Coagglomeration, clusters, and the scale and composition of cities. *Journal of Political Economy*, *122*(5), 1064–1093.

Kerr, W. R. (2008). Ethnic scientific communities and international technology diffusion. *Review of Economics and Statistics*, *90*(3), 518–537.

Klier, T., & McMillen, D. P. (2008). Evolving agglomeration in the U.S. auto supplier industry. *Journal of Regional Science*, *48*(1), 245–267.

Kofanov, D., & Mikhailova, T. (2015). Geographical concentration of Soviet industries: A comparative analysis. *Journal of the New Economic Association*, *28*(4), 112–141.

Kolomak, E. A. (2015). Evolution of spatial distribution of economic activity in Russia. *Regional Research of Russia*, *5*(3), 236–242.

Krugman, P. R., & Venables, A. J. (1995). Globalization and the inequality of nations. *Quarterly Journal of Economics*, *110*(4), 857–880.

Krugman, P. R. (1991). Increasing returns and economic geography. *Journal of Political Economy*, *99*(3), 483–499.

Lukiyanova, A., & Vishnevskaya, N. (2016). Decentralisation of the minimum wage setting in Russia: Causes and consequences. *The Economic and Labour Relations Review*, *27*(1), 98–117.

Marcon, E., & Puech, F. (2017). A typology of distance-based measures of spatial concentration. *Regional Science and Urban Economics*, *62*(C), 56–67.

Maslikhina, V. (2017). Spatial concentration of the manufacturing industry: Evidence from Russia. *Journal of Applied Engineering Science*, *15*(4), 509–517.
Matouschek, N., & Robert-Nicoud, F. L. (2005). The role of human capital investments in the location decision of firms. *Regional Science and Urban Economics*, *35*(5), 570–583.
Melo, P. C., Graham, D. J., & Noland, R. B. (2009). A meta-analysis of estimates of urban agglomeration economies. *Regional Science and Urban Economics*, *39*(3), 332–342.
Mori, T., Nishikimi, K., & Smith, T. E. (2005). A divergence statistic for industrial localization. *Review of Economics and Statistics*, *87*(4), 635–651.
Nakajima, K., Saito, Y. U., & Uesugi, I. (2012). Measuring economic localization: Evidence from Japanese firm-level data. *Journal of the Japanese and International Economies*, *26*(2), 201–220.
O'Sullivan, A., & Strange, W. C. (2018). The emergence of coagglomeration. *Journal of Economic Geography*, *18*(2), 293–317.
Rastvortseva, S. N., & Chentsova, A. S. (2015). Regional specialization and geographical concentration of industry in Russia. *Regional Science Inquiry*, *II*(2), 97–106.
Riedel, N., & Koh, H.-J. (2014). Assessing the localization pattern of German manufacturing and service industries. *Regional Studies*, *48*(5), 823–843.
Rosenthal, S. S., & Strange, W. C. (2004). Evidence on the nature and sources of agglomeration economies. In J. V. Henderson, & J.-F. Thisse (Eds.), *Handbook of regional and urban economics* (pp. 2119–2172). Amsterdam: Elsevier B.V, North-Holland.
Rosenthal, S. S., & Strange, W. C. (2001). The determinants of agglomeration. *Journal of Urban Economics*, *50*(2), 191–229.
Scholl, T., & Brenner, T. (2015). Optimizing distance-based methods for large data sets. *Journal of Geographical Systems*, *17*(4), 333–351.
Vorobyev, P., Kislyak, N., & Davidson, N. (2010). *Spatial concentration and firm performance in Russia* (Working Paper No. 10/05). Economics Education and Research Consortium (EERC). http://eercnetwork.com/default/download/creater/working_papers/file/adeb78a4a5c2ac588433
World Bank (2009). *World development report 2009: Reshaping economic geography*. Washington, DC: The World Bank Group.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Aleksandrova E, Behrens K, Kuznetsova M. Manufacturing (co)agglomeration in a transition country: Evidence from Russia. *J Regional Sci*. 2019;1–41. https://doi.org/10.1111/jors.12436

APPENDIX

Appendix A contains detailed information on the way that we collected and processed our plant-level data, Appendix B explains the details of the Duranton–Overman *K*-density estimations. Last, Appendix C contains details on the other data we use in the paper.

APPENDIX: DATA

A.1 | Overview, sources, and data cleaning

Our main data set is the 2014 version of the RUSLANA database from Bureau Van Dijk Electronic Publishing (BvDEP; see http://www.ruslana.bvdep.com). This database contains information about Russian companies, most notably contact information (addresses) and activity codes. The database provides legal, operational, postal, and de facto

address information. We use the de facto addresses, which come from open sources (call-center Credinform, business catalogs, companies' websites, etc.). According to BvDEP, there are a large number of yearly updates of these contact details. Identification of companies and establishments in our data is based on the Taxpayer's Identification Number (INN) and the All-Russian Classifier of Enterprises and Organizations (OKPO) pair.

We only look at the manufacturing portion of the RUSLANA database. The manufacturing sector is delimited by OKVED 15.00.00 to 37.20.70. From official statistics ("Monthly report of the socioeconomic situation in Russia," Russian Bureau of Statistic, January 2014 and 2015), there were 405,000 registered manufacturing companies (8.2% of all companies in Russia) in 2013, and 403,100 registered manufacturing companies (8.3% of all companies in Russia) in 2014. The raw version of the 2014 RUSLANA database contains 774,469 manufacturing establishments. Primary data cleaning—removing establishments with no address information and those in the Republic of Crimea—reduce our sample to 726,897 establishments. We then discard plants based on their activity status. The raw version of the database contains many establishments that have been liquidated, are in the process of being liquidated, or have otherwise been removed from the state register. We drop them all. We further keep only those establishments whose contact information were updated between 2012 and 2014. These operations yield a database of 345,384 establishments with address information in 2012–2014. Of these, we can precisely geocode 320,934 companies (see Appendix A.2 below). Since we use the 2014 version of the data, this implies that we may include a number of plants that changed address between 2012–2013 and the last year of the data, 2014. In other words, 2012 or 2013 address information may be obsolete in 2014. We do not believe that this is a big problem since plants do not move frequently. Hence, using the sample including the three address years should provide a more accurate picture of the geographic distribution of economic activity in Russia.

To look at the geographic distribution of industries, we further require industry codes for each establishment. Each establishment reports a primary industry code from the National Industry Classification OKVED (OK 029-2007, used from January 01, 2008–January 01, 2011), which is similar to the NACE Rev.2 classification at the four-digit level. We henceforth refer to it as OKVED 2007 or just OKVED for short. We use industry codes for establishments up to the three-digit level. Although finer levels of industrial classification are reported by a number of companies, this was not legally required before 2012. Hence, samples with industry codes beyond the three-digit level may be unreliable—some plants only report three-digit codes, whereas others in the same industry report four-digit codes. We end up with a final data set of 316,967 establishments out of the 320,934 that are precisely geocoded and which report industry information at the three-digit level. Table S1 in the Supporting Information Appendix provides a detailed breakdown of establishments by three-digit industry codes and by east–west location status (see Appendix A.2 for details on the latter).

A.2 | Geocoding and east–west split

We use a geocoding procedure that involves three steps:

(1) First, each establishment's location is geocoded using the Google Maps API engine (see http://www.google.com/MapsAPI). The geocoding procedure returned approximate geographic coordinates for about 70% of establishments in the sample, based mainly on the postal code area. Note that this inaccuracy in the geocoding can be of the order of about 2–3 km. The fact that only about 20% of the establishments were rooftop-geocoded using Google can be due to human mistakes (mistakes in postal codes, house numbers, noise in the address information like office numbers), changes in street names and numbers, an inadequate treatment of the building numbers by Google, or—most likely—ambiguities and errors in the translation from the Cyrillic to the roman alphabet.

(2) Second, we repeat the geocoding based on the romanized versions of the addresses but using the Russian map API provided by Yandex (see http://tech.yandex.ru/maps). The Yandex geocoding service provides a finer geographical coverage of Russian localities compared to Google. Yet, the Yandex map API calls do not allow for

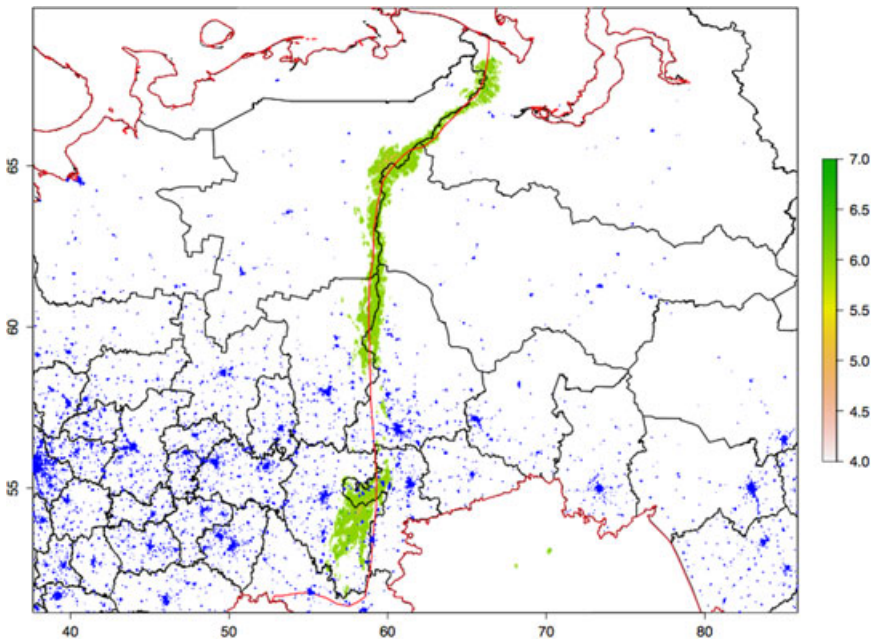

FIGURE A1 East–west split of the European and Asian parts along the Ural mountains

postal code parameters, which can lead to multiple results (in different regions) for the same street address. To take advantage of both geocoding engines, we utilize the geographic coordinates received in Step (1) as centroids and construct 55 km buffers around them. We then explicitly restricting the Yandex search among the localities contained in those buffers. In that case, 66% of the establishments are exactly geocoded with rooftop precision. Only few establishments are not assigned coordinates with a precision of at least the street number.

(3) Last, we run a geocoding procedure based on the Cyrillic versions of the addresses, which have been retranslated from their romanized spelling in the original data set. We again use the Russian map API provided by Yandex. As expected, this procedure yields a worse success rate, but still allows us to retrieve a number of establishments that could not be geocoded based on romanized names.[17]

We finally keep the following establishments that we consider to be geocoded precisely enough: (a) those that are rooftop coded or approximately (postcode) coded by Google API, and that are at the same time precisely coded by Yandex with a difference of less than 2 km between the two results—we then retain the Yandex coordinates, which we consider to be more accurate in general; and (b) the establishments that are precisely coded by Google API and not precisely coded by Yandex, in which case we retain the Google coordinates.

Turning to the east–west split of our sample, the elevation map in Figure A1 shows that the Ural mountain range forms a natural north–south boundary between the Asian and the European parts of Russia.[18] We thus use it to split our sample along that line. Note that the northern part of the Ural mountains runs along the boundary between the regions of the Republic of Komi in the west, and Yamalo-Nenets and the Khanty–Mansiisk

[17]Although the original Cyrillic names do seem to be available in the RUSLANA database, switching to "English" in the options leads to a download where the names are automatically translated using the Roman alphabet. We noticed this only later once the download and geocoding had been done.

[18]This map is provided by the UN Environment World Conservation Monitoring Centre (http://datadownload.unep-wcmc.org/?dataset=Mountains_and_Forests_in_Mountains_2000). The green areas correspond to zones with elevation between 1,000 and 1,500 meters and with a slope of more that 5° or with a local (7 km radius) elevation range in excess of 300 m.

autonomous districts in the east. Whereas the northern part of the mountains follows the administrative boundaries, this is no longer the case in the middle and southern parts. There, the mountains run through the Bashkortostan, Orenburg, Chelyabinsk, Perm, and Sverdlovsk regions. There is no clear cut along administrative lines. This shows that the east–west split can only be meaningfully implemented with detailed microgeographic data.

We use data from the Natural Earth Data page (see http//www.naturalearthdata.com/download/10 m/cultural/ne_10m_admin_0_scale_rank.zip) to split Russia along the Ural. To this end, we dissolve the polygons for Russia by identifier "RUE," which are related to the European Part, and by identifier "RUA," which are related to the Asian Part. Figure 1 shows the north–south division of Russia along the Ural into its eastern and western parts. Using that division, we create an indicator east–west for each plant in our data set.

APPENDIX B: DURANTON–OVERMAN METHODOLOGY

This appendix provides details on the Duranton–Overman methodology (Duranton & Overman, 2005, 2008) to compute K-densities to assess the agglomeration of industries and the coagglomeration of industry pairs.

B.1 | Agglomeration

The Duranton–Overman procedure involves four main steps.

First step (kernel densities). Consider an industry A with n plants. We compute the great circle distance d_{ij}, using latitude and longitude coordinates, between each pair (i, j) of establishments in that industry as follows:

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

Since $d_{ij} = d_{ji}$, this yields $n(n-1)/2$ distinct bilateral distances. The kernel-smoothed estimator of the density of these pairwise distances, henceforth called K-density, at distance d is

$$\hat{K}(d) = \frac{2}{n(n-1)h} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f\left(\frac{d - d_{ij}}{h}\right), \tag{B1}$$

where h is the optimal bandwidth—set according to Silverman's rule of thumb—and $f(\cdot)$ is a Gaussian kernel function. We estimate expression (B1) for all $d \leq x$, where x is a cutoff distance that we need to specify in the application. The K-density (B1) thus describes the distribution of bilateral distances between establishments in a given industry. Since the K-density is a distribution function, we can also compute its cumulative (CDF) up to some distance $\overline{d} \leq x$:

$$\text{CDF}(\overline{d}) = \sum_{d=0}^{\overline{d}} \hat{K}(d). \tag{B2}$$

The CDF at distance $\overline{d}$ thus measures the share of establishment pairs that are located less than distance $\overline{d}$ from each other. Alternatively, we can view this as the probability that two randomly drawn establishments in an industry will be at most $\overline{d}$ kilometers away from one another. Larger values of the CDF for a given distance indicate industries that have more compact geographic location patterns.

Second step (counterfactual densities). Using the locations of all manufacturing establishments in our sample, we randomly draw as many locations as there are plants in industry A. To each of these locations, we randomly assign a plant from industry A. We then compute the bilateral distances of this hypothetical distribution and estimate the associated counterfactual K-density (B1) of these bilateral distances. This procedure ensures that we control for

the overall pattern of geographic concentration in the manufacturing sector, as well as for the differences in industry sizes.

Third step (confidence bands). For each industry A, we repeat the second step 1,000 times. This yields a set of 1,000 estimated values of the K-density at each distance d. We then use our bootstrap distribution of K-densities, generated by the counterfactuals, to construct a two-sided confidence band that contains 90% of these estimated values. The upper bound, $\overline{K}(d)$, of this interval is given by the 95th percentile of the counterfactual distribution, and the lower bounds, $\underline{K}(d)$, by the fifth percentile of that distribution. We construct only *global* confidence bands such that deviations by randomly generated K-densities are equally likely across all levels of distances (see Duranton & Overman, 2005, for details).

Fourth step (identification of location patterns). Industries whose observed K-densities fall into their confidence band could be "as good as random" and are, therefore, not considered to be either localized or dispersed. Any deviation from the confidence band constructed in the third step indicates localization or dispersion of the industry. If $\widehat{K}(d) > \overline{K}(d)$ for at least one $d \in [0, x]$, whereas it never lies below $\underline{K}(d)$ for all $d \in [0, x]$, industry A is said to be globally localized at the 5% confidence level. Contrarily, if $\widehat{K}(d) < \underline{K}(d)$ for at least one $d \in [0, x]$, industry A is said to be globally dispersed.[19] We can also define an index of global localization, $\gamma_i(d) \equiv \max\{\widehat{K}(d) - \overline{K}(d), 0\}$, as well as an index of global dispersion:

$$\psi_i(d) \equiv \begin{cases} \max\{\underline{K}(d) - \widehat{K}(d), 0\} & \text{if } \sum_{d=0}^{x} \gamma_i(d) = 0, \\ 0 & \text{otherwise.} \end{cases}$$

The strength of localization and dispersion up to some distance $\overline{d} \leq x$ can be measured by

$$\Gamma_i(\overline{d}) \equiv \sum_{d=0}^{\overline{d}} \gamma_i(d) \quad \text{and} \quad \Psi_i(\overline{d}) \equiv \sum_{d=0}^{\overline{d}} \psi_i(d), \tag{B3}$$

which corresponds to the integral between the observed distribution and the upper and lower bounds of the confidence band. These two measures capture how "strongly" an industry deviates from randomness. Of course, $\Gamma_i = \Psi_i = 0$ for all distances for industries that do not deviate significantly from randomness.

Implementation details. We need to determine over what distance range x we compute the K-densities. In our application, we consider a range of distances between 0 and 1,000 km. Since Russia is a large country, it is important to evaluate the K-densities over a sufficiently long range. However, the computational burden increases substantially with the number of points on which we evaluate the K-densities. We believe that 1,000 km strikes the right balance between the need for longer distances, the geographic structure of Russia, and the computation time.[20] Furthermore, we need to determine the step size between two successive distances for evaluating the K-density. There is again a tradeoff between using a fine grid (many points) and the computational burden. We compute the K-densities using step sizes of 5 km. We use the interpolated value $[\widehat{K}(d) + \widehat{K}(d+\text{step size})] \times (\text{step size}/2)$ in the computations of the CDF. We do the same to compute $\Gamma_i(\overline{d})$ and $\Psi_i(\overline{d})$. Last, to speed up computations—given the large sample sizes that we have—we use the algorithm with discrete binning proposed by Scholl and Brenner (2015). We use 1 km distance bins for the approximation. Using smaller bins makes numerically no difference in our computations.[21]

[19]Barlet et al. (2013, p. 345) show that "the DO test for localization suffers from a systematic upward bias in small samples, and, more importantly, that this bias increases with the number of plants in the industry." We acknowledge this problem but do not think that this changes systematically the results of our subsequent analysis.

[20]Behrens and Bougna (2015) use 800 km for Canada, which is also a geographically large country. In Canada, most distances between neighboring large cities fall into that distance range. The same is true for Russia using a 1,000-km cutoff.

[21]In a previous version of the paper (Aleksandrova, Behrens, & Kuznetsova, 2018), we did all computations using the "exact" K-densities that do not rely on the Scholl–Brenner approximation, including for the coagglomeration measures. This proved to be computationally infeasible for the large sample that

B.2 | Coagglomeration

The Duranton–Overman procedure can be easily adapted to compute measures of the coagglomeration of industry pairs. The measures that we present in Appendix B.1 may be viewed as the "coagglomeration" of an industry with itself. The extension to two different industries is then straightforward. It involves the same four main steps. Below, we highlight only where the procedures differ to save space.

First step (kernel densities). Consider two industries A and B with n_A and n_B plants, respectively. There are $n_A \times n_B$ unique bilateral distances d_{ij} between all pairs of plants in the two industries. Hence, analogously to (B1), the kernel-smoothed estimator of the density of these pairwise distances at distance d is

$$\hat{K}_c(d) = \frac{1}{n_A n_B h} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} f\left(\frac{d - d_{ij}}{h}\right), \tag{B4}$$

where h is the optimal bandwidth—set using Silverman's rule of thumb—and $f(\cdot)$ is a Gaussian kernel function. We again estimate expression (B4) for all $d \leq x$, where x is the cutoff distance of 1,000 km. The K-density (B4) gives the distribution of bilateral distances between establishments in the two industries. As before, its CDF up to some distance $\overline{d} \leq x$ is

$$\text{CDF}_c(\overline{d}) = \sum_{d=0}^{\overline{d}} \hat{K}_c(d), \tag{B5}$$

which measures the share of pairs in the two industries—one from each industry—that are located less than distance $\overline{d}$ from each other. Larger values of the CDF for a given distance indicate industry pairs that have more compact geographic location patterns with respect to each other.

Second step (counterfactual densities). As for the case of the agglomeration of single industries, we construct confidence bands by drawing 1,000 random samples of establishments. A key difference is that we restrict the counterfactual to the locations that contain establishments of either industry A or B. Put differently, we take the joint distribution of the establishments in the two industries as our benchmark. This means that any departure from the counterfactual distribution measures how much closer establishments in the two industries are from each other than from establishments in the two industries in general. This is a strong test since the strength of coagglomeration—that is, the difference between the observed distribution and the counterfactual distribution—already controls for the agglomeration patterns of the two industries.[22] A direct consequence of this is that some industry pairs can be strongly concentrated geographically, but not be significantly coagglomerated conditional on that geographic concentration. For each industry pair, we compute global confidence bands based on 1,000 random permutations of the two industries.

Third step (confidence bands). This step is identical to that in Appendix B.1, except that the counterfactuals are drawn from the joint distribution of the two industries.

Fourth step (identification of location patterns). This step is also identical to that in Appendix B.1, using the confidence bands derived for the coagglomeration measures. We can compute the measures $\Gamma_i(\overline{d})$ and $\Psi_i(\overline{d})$ of excess agglomeration and dispersion in the same way.

Implementation details. We compute the coagglomeration patterns for Russian manufacturing industry pairs. For computational reasons, we only do so for the OKVED three-digit industries for all of Russia—101 industries, for a total of $(101 \times 100)/2 = 5,050$ unique industry pairs—using 5 km steps. We do not compute separate coagglomeration patterns for western and for eastern Russia. Even when using the Scholl–Brenner algorithm,

we have so that we restricted our sample to smaller sizes. We no longer require this in this version of the paper.

[22]Other choices are possible for the counterfactuals. One implication of our specific choice as in Duranton and Overman (2008) is that the excess agglomeration or dispersion measures of individual industries are not directly comparable to those of industry pairs. The reference distribution—the counterfactual—is different. This explains why we do not use the excess agglomeration measures in our coagglomeration regressions.

this proves computationally too demanding. All remaining implementation details are as in Appendix B.1 for the case of single industries.

APPENDIX C: INPUT–OUTPUT TABLES, AD VALOREM TRANSPORT COSTS, AND PROXIES FOR MARSHALLIAN COVARIATES

We first explain how we link the NAICS four-digit classification used in North America to the Russian OKVED three-digit classification. We then detail our data and data sources.

C.1 | Crosswalk between OKVED three-digit and NAICS four-digit industries

The 2007 OKVED three-digit classification—with 101 manufacturing industries—and the 2002 NAICS four-digit classification—with 86 manufacturing industries—are broadly comparable. We classify sectoral matches into four categories (which correspond to the indicator "flag" in Table S2 in the Supporting Information Appendix). First, there are sectors in either the NAICS or the OKVED classification that have no obvious matching counterpart. These include NAICS 3113 ("Sugar and Confectionery Product Manufacturing"), 3118 ("Bakeries and Tortilla Manufacturing"), 3141 ("Textile Furnishings Mills"), 3313 ("Alumina and Aluminum Production and Processing"), 3325 ("Hardware Manufacturing"), 3327 ("Machine Shops, Turned Product, and Screw, Nut and Bolt Manufacturing"), 3334 ("Ventilation, Heating, Air-Conditioning and Commercial Refrigeration Equipment Manufacturing"), (3335, "Metalworking Machinery Manufacturing"), 3344 ("Semiconductor and Other Electronic Component Manufacturing"); and OKVED 371 ("Recycling of metal waste and scrap"), 372 ("Recycling of nonmetallic trash and scrap"), 233 ("Processing of nuclear fuel"), 267 ("Cutting, shaping and finishing of decorative and building stone"), 174 ("Manufacture of made-up textile articles, except apparel"), 221 ("Publishing"). We flag these sectors with 0. We also flag all sectors that contain "other" or "not elsewhere classified" with 0. Those sectors, even when they have very similar names, are likely to have a somewhat different composition across classifications (since they are residual categories). Second, there are sectors where either many OKVED match to one NAICS or the other way round. We flag those with 2 or 3 (depending on the direction of the one-to-many correspondence) and try to match the best we can. There are also a small number of many-to-many correspondences. Since there is no satisfying way to deal with them, we flag them with 0. Finally, there is a fairly exact correspondence between about half of the sectors based on the sectors' names. One such example is NAICS 3115 ("Dairy Product Manufacturing") and OKVED 155 ("Manufacture of dairy products"). These "exact matches" are flagged with 1.

Table S2 in the Supporting Information Appendix below provides the full crosswalk that we use. There are 46 exact matches, 32 one-to-many matches (in either direction), and 33 matches that may be of worse "quality" (based on many-to-many, or the absence of matches, or "not elsewhere classified" categories). We use these matches to construct a "quality dummy" that takes value 0 if the flag equals 1 ("precise matches") and value 1 if the flag equals 0, 2, or 3 ("imprecise matches").

Since our crosswalk is not one-to-one, several different OKVED three-digit industry pairs may be associated with the same NAICS four-digit industry pairs. This creates some correlations in the error terms, which we control for in our regressions by clustering all standard errors at the NAICS four-digit industry-pair level.

C.2 | Trucking commodity origin destination survey

We use a series of recent ad valorem transport cost measures developed by Brown (2015) and used by Behrens et al. (2018) and Behrens and Brown (2018). These *ad valorem* rate series are estimated using Statistics Canada's Trucking Commodity Origin-Destination Survey (TCOD). The TCOD is a for-hire carrier-based survey that collects data on a per shipment basis, including the origin and destination, (network) distance shipped, revenue to the carrier,

tonnage, and the commodity of the shipment. To calculate ad valorem rates, the value of the shipment is also required. Unfortunately, the TCOD does not report the value of goods shipped. Hence, value per tonne estimates by six-digit Harmonized System (HS) commodity from an "experiment export trade file" produced in 2008 is used to estimate the value of the shipments. Commodity export price indices are used to project the value per tonne estimates through time (see Brown, 2015, for details). The commodity value per tonne estimates are used to estimate the value of shipments. This "augmented" TCOD file is the basis that is used to estimate ad valorem trucking rates by industry. This particular analysis requires a long time period to improve the accuracy of the predicted rates, and the estimates are based on survey weights that ensure trucking rates are representative of the population of carriers (see Brown, 2015; Behrens & Brown, 2018, for additional details). We use the ad valorem transport costs (AVTC) for the year 2008, the last year available in the data.

C.3 | Input–output coefficients

We construct input–output measures based on Russian, Canadian, and US input–output tables. We work with the manufacturing portion of the input–output tables, excluding hence services, primary industries, private consumption, government items, and imports/exports. We aggregate the data to the four-digit level for Canada and the United States, and the three-digit level for Russia. All input–output coefficients are computed as shares relative to total manufacturing inputs and outputs of each industry.

To make our input–output measures symmetric for the industry pairs *ij* and *ji*, we take either the (simple) average of the respective coefficients or their maximum value. Hence, in our coagglomeration regressions the input coefficient for industries *ij* is either the simple average of the two input coefficients *ij* and *ji* or the maximum of the two. We report results for both and they are fairly similar (the raw correlation between the series is larger than 0.95 in all cases). See Ellison et al. (2010) for additional details and discussion.

In what follows, we provide details on the data and the data sources.

Russia. The Russian Bureau of Statistics provides the basic input–output tables for 2011 (see http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/accounts/). We use the symmetric tables, which are compatible with the international standards in terms of both methodology and the official documents used to construct them. The System of National Accounts provides all the relevant groundwork for the entire range of our basic tables. The tables are produced using the two- and three-digit Russian Classification of Products by Economic Activities (OKPD), which is mostly identical to the three-digit Russian National Industry Classification (OKVED 2007) that we make use of in our work. Both classifications (OKPD and OKVED) are product-based, thus allowing for simple crosswalks as in Canada and the United States. When required, we disaggregate two-digit sectors to the three-digit level using weights that are constructed using the figures for the average employment across regions and sectors for the period 2012–2014. These data are from the Russian Bureau of Statistics (see https://fedstat.ru/indicator/43211).

Canada. We use the 2007 input–output matrix for Canada. The finest public release of the input–output matrices is at the *L*-level (link level), which is between NAICS three- and four-digit. We disaggregated the matrix to the *W*-level (NAICS six-digit) using either sales or employment data as sectoral weights. We use the input–output tables at buyers' prices. For each manufacturing industry, *i*, we allocate inputs purchased or outputs sold in the *L*-level matrix (at the three- or four-digit level) to the corresponding NAICS six-digit subsectors. To do so, we allocate the total sales of each sector to all subsectors in proportion to those sectors' sales in the total sales to obtain a 257×257 matrix of NAICS six-digit inputs and outputs for manufacturing. We reaggregate these matrices to the four-digit level to compute the shares that sectors buys from and sell to each other. The input–output shares for the manufacturing submatrix are rescaled to sum to unity.

United States. We use the 2002 input–output benchmark tables from the Bureau of Economic Analysis. We use the detailed six-digit table and construct the same input–output shares as explained above for Canada and Russia. Note that the US matrix features a larger number of zeros than our treated version of either the Russian or

Canadian tables. This may be due to either our treatment of the Russian and Canadian tables (where we have to break down some industries into subindustries) or to the existence of thresholds for reporting numbers in the US data. In any case, the larger number of zeros implies that we are losing more observations when taking the log of the US data. This explains why we do not use these data as instruments in the log specifications.

C.4 | Occupational employment similarity

We compute measures of worker similarity in the different industries. To this end, we use US Occupational Employment Survey data from the *Bureau of Labor Statistics* for 2011 to compute the share of each of 554 occupations in each four-digit NAICS industry. We only retain occupations for which there is at least some employment in manufacturing (e.g., there are no "Surgeons" in manufacturing industries, hence we exclude them completely from our data). Our measure of occupational employment similarity is computed as the correlation coefficient between the vectors of occupational shares of industries i and j. By construction, this measure is symmetric in ij and ji.

C.5 | Patent citations across industries

Last, we construct proxies for "knowledge spillovers" or "knowledge sharing" by using the NBER Patent Citation database (which builds on US Patent and Trade Office data) and by following previous work by Kerr (2008). Our proxy for knowledge flows is the maximum of the shares of patents that industry i (or j) manufactures ("make-based") or uses ("use-based") and which originate from the other industry j (or i). We take the maximum of the shares ij and ji to obtain a symmetric measure for each pair.