Universidade Federal de Pernambuco - UFPE
Centro de Ciências Sociais Aplicadas - CCSA
Departamento de Economia
Programa de Pós-Graduação em Economia - PIMES

Edilberto Tiago de Almeida

Essays on Agglomeration Economies: location patterns, attenuation and human capital spillovers

Recife
2021

Universidade Federal de Pernambuco - UFPE
Centro de Ciências Sociais Aplicadas - CCSA
Departamento de Economia
Programa de Pós-Graduação em Economia - PIMES

Edilberto Tiago de Almeida

Essays on Agglomeration Economies: location patterns, attenuation and human capital spillovers

Thesis presented to the **Programa de Pós-Graduação em Economia - PIMES** of Departamento de Economia of **Universidade Federal de Pernambuco - UFPE** in partial fulfillment of the requirements for the degree of Doutor em Economia.

Advisor: Dr. Raul da Mota Silveira Neto
Co-advisor: Dra. Roberta de Moraes Rocha

Recife
2021

Catalogação na Fonte
Bibliotecária Ângela de Fátima Correia Simões, CRB4-773

A447e Almeida, Edilberto Tiago de
Essays on agglomeration economies: location patterns, attenuation and human capital spillovers / Edilberto Tiago de Almeida. - 2021.
188 folhas: il. 30 cm.

Orientador: Prof. Dr. Raul da Mota Silveira Neto e Coorientadora Prof.ª Dra. Roberta de Moraes Rocha
Tese (Doutorado em Economia) – Universidade Federal de Pernambuco, CCSA, 2021.
Inclui referências e apêndices.

1. Concentração industrial. 2. Externalidades (Economia). 3. Capital humano. I. Silveira Neto, Raul da Mota (Orientador). II. Rocha, Roberta de Moraes (Coorientadora). III. Título.

336 CDD (22. ed.) UFPE (CSA 2021 – 053)

Edilberto Tiago de Almeida

Essays on Agglomeration Economies: location patterns, attenuation and human capital spillovers

Thesis presented to the **Programa de Pós-Graduação em Economia - PIMES** of Departamento de Economia of **Universidade Federal de Pernambuco - UFPE** in partial fulfillment of the requirements for the degree of Doutor em Economia. **Approved** by the Examination Committee:

Recife, 31/05/2021

Prof. Dr. Raul da Mota Silveira Neto
Advisor
Universidade Federal de Pernambuco - UFPE

Prof. Dra. Tatiane Almeida de Menezes
Universidade Federal de Pernambuco - UFPE

Prof. Dr. Carlos Roberto Azzoni
Universidade de São Paulo - USP

Prof. Dr. Gervásio Ferreira dos Santos
Universidade Federal da Bahia - UFBA

Prof. Dr. Pedro Vasconcelos Maia do Amaral
Universidade Federal de Minas Gerais - UFMG

Aos meus pais

Acknowledgements

First of all, I'm immensely grateful to the two most important people in my life, my father Edmilson Almeida and my mother Eucrimaria Silva, for their unconditional support in all these years. Without their encouragement, dedication and love, this achievement would not be possible. I'm also grateful to my mother Maria José Almeida (in memory) who cannot accompany my trajectory, but is present through her love and protection. I dedicate this thesis to the three of them, the loves of my life.

I am grateful to my advisor, Prof. Raul Silveira Neto, for his motivation, patience, generosity and valuable guidance of this work and to my career. His dedication to research is an example to me. I thank my co-advisor, Prof. Roberta Rocha, for the teachings and guidance, which contributed to this work and to my career. I'm also grateful to the teaching from the professors of the Department of Economics of UFPE and PIMES, Francisco Ramos, Gustavo Sampaio, Rafael Azevedo, Paulo Vaz, Francisco Cribari Neto, Rafael Costa Lima, Rafael Vasconcelos, Tatiane Menezes and Álvaro Hidalgo. I also thank the secretaries of PIMES, Maria Luíza and Jackeline Ferreira, for their support whenever I needed it.

I thank to the members of the Examining Committee, professors Tatiane Menezes, Carlos Azzoni, Gervásio Santos and Pedro Amaral, for their availability, comments and suggestions.

I thank to my friends from PIMES, Gilberto Nogueira, Yuri Barreto, Andrews Barros, Joebson Maurílio, and Francisco Bustos for sharing moments of joy in this trajectory.

I was encouraged by many professors from my undergraduate and master's studies, especially Carla Calixto, and my professors of UFRPE-UAST and UFPE-CAA.

I express special thanks to Robson Santana and Ana Maria Santana for the support and encouragement during these years. I'm also thankful to my friend, Rosy Nascimento, for the encouragement since my undergraduate studies.

Finally, I acknowledge and am grateful for financial support from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Abstract

This dissertation addresses the issue of agglomeration economies from a perspective still little studied in developing countries. Using a unique geocoded database for the manufacturing activities in Brazil, we provide a detailed set of new evidence on the geographical distribution of these activities and on the spatial extent of agglomeration economies on the birth of new establishments and on labor productivity in a micro-geographic context. The second chapter documents the patterns of location of manufacturing activity in Brazil using a distance-based measure, thus not susceptible to the modifiable areal unit problem (MAUP) that is common in traditional concentration measures. We show that in the period 2006-2015 there were no significant changes in the pattern of location of manufacturing industries in the country and that these activities are more concentrated in Brazil than in other developing countries such as China and Russia, and much more concentrated than the pattern observed in developed countries. The third chapter evaluates the spatial extent of the agglomeration economies on the location choice of new establishments and on the employment levels chosen. More specifically, we estimate the local determinants of the number of births per square kilometer and the associated employment levels as functions of the own-industry employment and other economic environment characteristics when the location decisions were made. The main results show that agglomeration economies are attenuated with distance. Moreover, in nearly all cases for both births and new-establishment employment, localization effects disappear after 5 km from the site chosen by the new establishment. The fourth chapter explores the spatial extent of human capital externalities on the wage in cities. We show that human capital spillovers are highly localized and stronger at short distances, specifically up to 1 km from the individual's workplace. These effects are also attenuated with distance, which is consistent with the idea that knowledge spillovers occur mainly from face-to-face interaction among workers.

Keywords: Location Patterns. Industrial Concentration. Distance-Based Measures. Attenuation. Human Capital Spillovers.

Resumo

Esta tese aborda a questão das economias de aglomeração numa perspectiva ainda pouco estudada nos países em desenvolvimento. Usando um conjunto de dados georreferenciados único para a indústria manufatureira no Brasil, fornecemos um conjunto detalhado de novas evidências sobre a distribuição geográfica destas atividades e sobre a extensão espacial das economias de aglomeração sobre o surgimento de novos estabelecimentos e sobre a produtividade do trabalho em um contexto microgeográfico. O segundo capítulo documenta os padrões de localização da atividade manufatureira no Brasil usando uma medida baseada em distâncias e, portanto, não suscetível ao problema da unidade geográfica modificável (MAUP), comum em medidas de concentração tradicionais. Mostramos que no período 2006-2015 não houve mudanças significativas no padrão de localização da indústria de manufatura no país e que estas atividades são mais concentradas no Brasil do que em outros países em desenvolvimento como China e Rússia e muito mais concentradas do que o padrão observado nos países desenvolvidos. O terceiro capítulo avalia a extensão espacial das economias de aglomeração sobre a escolha locacional dos novos estabelecimentos e sobre os níveis de emprego que eles escolhem. Mais especificamente, estimamos os determinantes locais do número de nascimentos de firmas por quilômetro quadrado e seus níveis de emprego como funções do emprego na própria indústria controlando outras características do ambiente econômico quando as decisões de localização foram tomadas. Os principais resultados mostram que as economias de aglomeração são rapidamente atenuadas com a distância. Além disso, para a maioria das indústrias estudadas, os efeitos não são estatisticamente significantes após 5 km do local escolhido pelo novo estabelecimento. O quarto capítulo explora o escopo espacial das externalidades de capital humano sobre o salário dos trabalhadores. Mostramos que os spillovers de capital humano são altamente localizados e mais fortes a curtas distâncias, precisamente, mais fortes até 1 km do local de trabalho do indivíduo. Estes efeitos também são atenuados rapidamente com a distância, o que é consistente com a ideia de que os spillovers de conhecimento ocorrem principalmente a partir da interação face-to-face entre trabalhadores.

Palavras-chave: Padrões de Localização. Concentração Industrial. Medidas Baseadas em Distâncias. Atenuação. Externalidades de Capital Humano.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

What determines the geographic concentration of economic activities is one of the most fundamental questions in urban economics. Marshall (1890) already identified the intrinsic market factors that explain location patterns and generate local externalities. The existence of large cities is evidence that there are gains in agglomeration and that in general these gains offset the forces of dispersion; otherwise activities would not be concentrated (Puga, 2010; Duranton and Puga, 2014; Thisse, 2018). Due to the greater availability of disaggregated data and the development of techniques (distance-based measures) that allow a more careful analysis of the spatial distribution of firms and workers (see, e.g., Marcon and Puech, 2003; 2009; Duranton and Overman, 2005; 2008), the location patterns and their possible causes have been documented. This evidence, however, almost exclusively pertains to developed countries.

Only recently have these measures been applied in the context of developing countries, where spatial distribution of activities tends to be more unbalanced. This is unfortunate, since not only do the potential gains from agglomeration of activities tend to be relatively more important for these countries (see, e.g., Duranton, 2016a; Barufi *et al.*, 2016; Chauvin *et al.*, 2017; Combes *et al.*, 2013; 2020), but also not well-informed and designed territorial public policies may promote inefficient allocation of their fewer resources. In fact, in accordance with these characteristics, Brakman *et al.* (2016) and Aleksandrova *et al.* (2019) revealed that manufacturing activities are more concentrated in China and Russia, respectively, than in developed countries such as the UK (Duranton and Overman, 2005), Japan (Nakajima *et al.*, 2012), Belgium, France, Germany, Italy, Spain (Vitali *et al.*, 2013), and Canada (Behrens and Bougna, 2015).

More than this, the externality generated by agglomeration itself may have a different spatial scope in cities in developing countries. This may occur because the structure of cities is different in these environments, which can substantially affect the geographic spread of agglomeration externalities. In turn, while understanding this phenomenon in a micro-geographic context is politically relevant and essential to understand their nature, there is little empirical evidence in this respect. To be more precise, using micro-geographic

data, only Li *et al.* (2020a) provided evidence for China suggesting that attenuation is very different among industries and that in general this is faster than in the US.

Besides this limited set of evidence for developing countries, other reasons make the Brazilian manufacturing sector particularly appealing for the analyses. For example, unlike China, historically there has been no restriction on worker mobility, and economic activities are more market oriented in Brazil, which can substantially affect the geographical distribution of activities and also the spatial extent of agglomeration economies. Associated with this internal mobility, as recently highlighted by Chauvin *et al.* (2017), the country has a high urbanization rate (around 85%) in comparison with other developing countries. Brazil is among the countries with the highest levels of income inequality in the world (Fishlow, 1972; Mendonça and Barros, 1995; Narita *et al.*, 2003) and the internal regional inequality in educational levels (Suliano and Siqueira, 2012; Silva and Silveira Neto, 2015) is also high, which can make the external gains associated with education, such as human capital spillovers, more localized. In addition, the spatial location of manufacturing has played an important role in originating the current extremely high level of Brazilian spatial economic inequality, because the country implemented a strong and spatially located import substitution process (Furtado, 1963; Leff, 1972; Baer, 2002). In line with this historical location pattern, most of the country's current territorial policies still focus on the manufacturing sector. This situation makes obtaining continuous and spatially consistent measures of concentration of activities and a better understanding of the spatial scope of agglomeration externalities fundamental.

This dissertation contributes fills part of this gap in the empirical literature by exploring the issue of agglomeration economies from a microgeographic perspective in Brazil. For this, we use a unique geocoded database for manufacturing industries in Brazil. Each of the following chapters has its own research problem, but although they are independent, they deal with the same general theme: economies of agglomeration.

In chapter 2 we implement the nonparametric approach developed by Duranton and Overman (2005) to better understand the location patterns of Brazilian manufacturing activities. The DO Index is a distance-based measure and therefore is not susceptible to the modifiable areal unit problem (MAUP), common in traditional concentration measures. Basically, the DO Index determines the distribution of bilateral distances between firms and compares this distribution with a set of randomly distributed bilateral distances. Therefore, an industry can be defined as significantly localized or dispersed if the distribution of bilateral distances observed deviates from the random pattern. Our results show that 89.9% and 91% of manufacturing at the 3-digit CNAE level had statistically

significant localization for 2006 and 2015, respectively, and that these patterns remain high when we consider 4-digit codes in weighted and unweighted versions of the measure. High-tech industries have location patterns at short distances, being located mainly in large urban areas, while low-tech industries are located at large distances. Consistent with Marshallian agglomeration forces and with transport cost, proxies for labor pooling, knowledge spillovers and transport costs are related to measures of geographic localization, indicating that these factors are associated with plants' location patterns. More than this, knowledge spillovers are more important for high-tech industries, a result that conforms very well with the concentration pattern observed for this type of industry.

Chapter 3 provides evidence about the spatial scope of agglomeration economies, focusing on the localization effects (own-industry employment), on the number of births per square kilometer and their associated employment. Unlike previous studies, such as Rosenthal and Strange (2003), we use a set of tools and the characteristics of our data to deal with spatial sorting problems and other potential sources of endogeneity. Initially, to better understand the location pattern of new establishments, we use the localization and colocalization measures proposed by Duranton and Overman, (2005; 2008). At this stage we are able to differentiate industries according to the pattern of geographic proximity between new and existing establishments. To deal with sorting, we use cells of 1 square kilometer defined exogenously from the territorial limits of Brazil. The idea is that very small geographic areas are outside the set of locational choice of the new establishments, mainly within and near the large cities, where land use is more intensive and different sectors dispute the use of geographic space. Other remaining sources of endogeneity are controlled by within-city fixed effects and a broad set of controls at the cell level and also at city level for economic environment, previously existing transportation infrastructure, geographic characteristics, and local development policies around the place chosen by the new establishment. We also use a control function approach with a shift-share instrumental variable to address the problem of endogenous explanatory variables in nonlinear models. In this nonparametric analysis, we find there are patterns of colocalization between entrants and existing establishments and that these patterns occur mainly at short distances. For these industries, we find that the localization economies generated by the own-industry employment at different geographic distances from the location chosen by the new establishment are attenuated rapidly with distance (around 5 km). These results are clearly in line with the high concentration of manufacturing industries, as presented in chapter 2, and also with the free geographic mobility of workers in the country.

In chapter 4 we investigate the spatial extent of human capital spillovers on the wage earnings within Brazilian cities. For this, we use a geocoded employer-employee panel dataset, exogenously determined grid and different spatial distance bands. First, to better understand how the spatial attenuation of human capital externalities can be incorporated into the firm optimization problem, we generalize Moretti (2004a)'s theoretical framework by introducing a continuous space, therefore allowing the effects to vary with geographic distance from the individual's workplace. Empirically, we deal with the spatial sorting and other sources of endogeneity in the wage-human capital spillover relationship with controls for observable and unobservable individual and establishment characteristics and instrumental variables based on the exogenous expansion of higher education in Brazil. Specifically, our identification strategy uses the large shifts in national education policy between 1991 and 2004 as an exogenous source of variation in the number of college-educated people across different distance bands within Brazilian metropolitan regions to identify the effect of the concentration of college-educated workers in the distance bands that surround the individual's workplace on the individual wages (our proxy for labor productivity). Our main results indicate that distance still seems to be more important in the developing country context. There is a spatial pattern both in the economic mass externalities and in the knowledge spillovers generated by the proximity of college educated workers. These are stronger up to 1 km and from there they are attenuated up to 10 km. These results are robust even when we combine different controls, such as worker-plant and worker-city matches, and conform very well with the results obtained in the two previous chapters, namely the high levels of concentration of manufacturing industries in Brazil when compared to other developing and developed countries; the high concentration at short distances of high-tech industries relative to low-tech ones; the importance of proximity to other plants in the same industrial sector; and the attenuation of localization economies with distance.

Chapter 2

Manufacturing location patterns in Brazil

2.1 Introduction

In the last two decades, the use of distance-based measures of spatial concentration have allowed obtaining new detailed evidence about location patterns of economic activities, mainly for developed countries (see, e.g., Duranton and Overman, 2005, 2008; Behrens and Bougna, 2015). Despite these recent advances, at least two issues remain underexplored: the measuring of spatial agglomeration in developing countries and the role played by economic variable determinants of agglomeration economies. The first is associated with the lack of detailed microgeographic data, the second with the complicated challenges of identifying specific effects among simultaneous agglomeration forces. Nevertheless, both areas are essential for understanding the spatial economies of developing countries.

Because of the scarcity of appropriate data and computational limitations, studies that use distance-based metrics are still scarce. A set of recent evidence for a small group of European, Asian and American countries showed that industrial activity exhibits specific location patterns. These findings suggest, therefore, that high level concentration of manufacturing can be observed in different countries of the world. Duranton and Overman (2005), e.g., analyzed location patterns for the manufacturing industry in the UK and showed that 52% of industry had non-random localization. The main findings of Nakajima *et al.* (2012) indicated that 50% of the manufacturing industries in Japan had spatial location patterns. Vitali *et al.* (2013) presented evidence of industrial concentration patterns for six European countries: Belgium, France, Germany, Italy, Spain and the UK. Their main findings suggested that for all countries studied, traditional industries had significant localization patterns. For Germany, Koh and Riedel (2014) also found evidence that the manufacturing (71%) and the service (97%) sector, exhibit significant geographic localization. Also in line with these results, Behrens and Bougna (2015) found that 40% to 60% of manufacturing industries in Canada were geographically localized, depending on the industrial sector and the year studied.

For developing countries, the available evidence is even scarcer. To be precise, only

Brakman *et al.* (2016) and Aleksandrova *et al.* (2019) have analyzed the location of manufacturing in developing countries, in China and Russia, respectively. The authors found that around 80% of industries at 4-digit level in China and 3-digit level in Russia were significantly localized, indicating more pronounced patterns than some developed countries. In the Brazilian regional context, the available evidence about manufacturing spatial concentration is based exclusively on measures that are sensitive to the modifiable areal unit problem (MAUP) (see, e.g., Silveira Neto, 2005; Resende and Wyllie, 2005; Lautert and Araújo, 2007; Rocha *et al.*, 2019, Ferreira *et al.*, 2019). Within urban spaces, Silva *et al.* (2019) presented evidence based on Duranton and Overman (2005)'s metric (hereafter DO index) for the Recife Metropolitan Region (RMR). The spatial limitation of this work was obvious; these patterns of location in the RMR did not necessarily reflect those observed on a national scale.

Here we aim seeks to reduce this gap in the literature. We provide evidence about the patterns of location of Brazilian manufacturing activities using detailed microgeographic panel data and the DO index. Our geocoded database for manufacturing plants from 2006 to 2015 includes a sample of approximately 2.8 million of establishments, representing 96% of all manufacturing activity, and on average 7,033,906 jobs per year. We also explore how industrial location patterns change according to technological intensity. High-tech industries have a larger share of workers with college degrees, invest more in R&D and innovate together with other establishments, characteristics that can favor the location of these industries in large urban centers. Furthermore, we investigate what economic factors are associated with the location patterns observed using proxies for Marshallian agglomeration forces, transport cost, natural advantages associated with proximity to inputs and market structure, while controlling for observable and unobserved characteristics fixed in time specific to each industry.

In addition to the lack of evidence, other economic and technical factors make the investigation for Brazilian manufacturing particularly appropriate. Among the economic reasons, we highlight first that unlike other countries such as China, Brazil is historically more market oriented and the inter-regional mobility of workers is higher, which may substantially affect agglomeration patterns. Second, the heterogeneous spatial distribution historically observed in Brazil, its persistence through time, and the importance of industrialization to explain regional inequalities (see, e.g., Furtado, 1963; Leff, 1972; Baer, 2002). Within this context, regional development policies have historically been associated with incentives to manufacturing activities. Third, we observe that since the first decade of this century, there has been a reduction in per capita household income inequality (see,

e.g., Hoffmann, 2006). By analyzing the period 2006-2015, we investigate if there was any change in manufacturing location patterns consistent with this inequality reduction. Technically, different from investigations based traditional spatial concentration measures (e.g., Gini, EG), our strategy is not affected by changes in municipal boundaries, something not so rare in Brazil.[1]

We can summarize our key results as follows: (i) 89.9% and 91% of 3-digit Brazilian manufacturing plants were significantly localized in 2006 and 2015, respectively – higher than those documented for other countries in transition such as China (Brakman *et al.*, 2016) and Russia (Aleksandrova *et al.*, 2019), as well as for developed countries (Duranton and Overman, 2005; Koh and Riedel, 2014; Behrens and Bougna, 2015), and when we consider the employment weighted version of the DO index at the 3-digit and 4-digit levels (weighted and unweighted), the results are also high; (ii) location patterns vary greatly depending on the technological level of the industry, where high-tech industries have location patterns at short distances, being located mainly in the large urban areas of the Southeast region, while low-tech industries are located at great distances; (iii) consistent with Marshallian agglomeration forces, proxies for labor pooling and knowledge spillovers are related to measures of geographic localization, indicating that these factors are associated with plants' location patterns; (iv) the evidence is weak for natural advantages associated with proximity to inputs; and (v) competition can act as a force contrary to industrial localization, favoring dispersion.

The remainder of the paper is structured as follows. In section 2.2 we present an exploratory overview of the spatial distribution of industries in Brazil and we describe the main details about our database. In section 2.3 we present the methodology and results from the DO index. In section 2.4 we explore the conditions that affect manufacturing localization. The last section contains our final comments.

2.2 Data and a snapshot of Brazilian manufacturing, 2006-2015

2.2.1 Data

Our main source of data is the Annual Report of Social Information (*Relação Anual de Informações Sociais*, or RAIS), which all formally organized companies must send to the

[1]For example, between the demographic censuses of 2000 and 2010 conducted by the Brazilian Institute of Geography and Statistics (IBGE), 57 municipalities were created by breaking away from existing ones (there are no unincorporated areas in Brazil).

Ministry of Labor and Employment. The resulting database provides a very rich source of data on the formal labor market. Information on firms and workers is available. For firms, the information covers address, number on the National Registry of Legal Entities (CNPJ), National Classification of Economic Activities (CNAE), which is compatible with the International Standard Industrial Classification of all Economic Activities (ISIC), revision 4, date of opening and closing of activities (if applicable), number of workers, size of the establishment, and legal nature of the establishment. For workers, information on schooling, age, wage, race, the plant in which are employed, among other important characteristics, is available.

Using geocoding techniques, we obtained a unique microgeographic database for 2,775,799 plants, and on average, 7,033,906 jobs per year. Of all the plants identified in the RAIS database, our sample represents is 96% of the total (see details in Table A.1 in Appendix A.1). When considering employment, the geocoded percentage is higher, 96.82% of the total. Note also that when we disaggregate by year, our percentages of geocoded plants and employment are never less than 95.5% and 96.6%, respectively. Our database accurately characterizes the distribution of the manufacturing activities in the country without having to use an a priori definition of geographic space.

The set of evidence about the location patterns of plants of different manufacturing activities was obtained by considering the 3-digit level of sector desegregation using the official CNAE 2.0.[2] This is the common level of sector desegregation used in similar studies (see, e.g., Duranton and Overman, 2005; Aleksandrova *et al.*, 2019) and comprises 285 groups of different economic activities (including agriculture, manufacturing, and services). As a robustness check, we also present our baseline results considering 4-digit coding of sector desegregation. After the constraints imposed on the database, we finally work with 103 manufacturing activities (sectors). Note that we have a panel with 103 cross-sectional and 10 year data units between 2006 and 2015, which allows us to control for unobserved and fixed effects in time of each sector at the 3-digit level. We will analyze the results of only industries with at least 10 plants each year.

2.2.2 Snapshot of Brazilian manufacturing

With 8,510,820.623 km^2 and a population of 190,755,799, according to data from the 2010 Demographic Census provided by the Brazilian Institute of Geography and Statistics

[2]Further details are available at: <https://concla.ibge.gov.br/busca-online-cnae.html?view=secao&tipo=cnae&versaoclasse=5&secao=C>.

(IBGE), Brazil is a preponderantly urban country (84.36% of population). The process of structural change of the rural to urban population movement occurred together with the different phases of industrialization of the country's economy. As the IBGE data show, since the 1960s the urban population has been growing, while the rural population has declined since the 1970s. These characteristics show the great importance of urban centers as dynamic environments. In fact, manufacturing activity is clearly denser in some regions, with emphasis on the Southeast[3] (46.18%). In a more disaggregated way, the state of São Paulo concentrates 26.81% of the manufacturing plants in the country, and when considering urban contexts, the São Paulo Metropolitan Region (SPMR) is the largest and most important metropolitan region in the country, concentrating 11.66% of the plants in 2015 (see Table A.4 in Appendix A.1). This indicates that the forces in favor of the concentration of firms and workers in large cities outweigh the forces for dispersion puga2010,duranton2014.

Also reflecting the spatially concentrated and market-oriented development process observed in Brazil, the Southeast region has the largest shares of employment (52.8% and 50% in 2006 and 2015, respectively) in manufacturing. Together with the South region, it represents more than 75% of the total jobs and also manufacturing plants. Other aggregate indicators also clearly show this pattern. For example, data from the Regional Accounts of Brazil provided by the IBGE for 2016 show that the Southeast region has 55.4% (around $ 115 million) of the manufacturing value added in the country. This is no surprise, since the country's largest and most dynamic urban centers are located in the Southeast region (e.g., the SPMR is the largest of them).

The industrial characteristics are also quite heterogeneous when considering the different levels of technological intensity. Table 2.1 summarizes industry-level details for two threshold years of our database, including the technological classification of the industry, average plant size by industry, share of workers with college degrees and the share of establishments that have implemented some product innovation in partnership with other companies. There is clear substantial differentiation by technological level. First, the technology-intensive industries (here we consider those classified as medium-high and high-tech) have a higher share of workers with college degrees[4] (on average, 20.6% in 2015) than the low-tech industries (on average, 8.13% in the same year). Second, these industries are characterized by high interaction in production innovation in partnership with other companies (on average, 1.77%) when compared to low-tech industries (on

[3]Brazil has five official regions: South, Southeast, Midwest, Northeast and North.
[4]Bachelor's, master's and doctorates.

average, 0.46%), as shown by detailed data from the Innovation Survey[5] (PINTEC) made available by the IBGE. Furthermore, according to PINTEC data for 2014, the R&D expenditures of high-tech industries represent 41.25% of the total value (which also include the expenditures of extractive industries, the electricity and gas sector and some selected service sectors) of the expenditures made in internal R&D activities of establishments that implemented innovations. When considering only the expenditures of the manufacturing firms, this percentage is 58.03%, while the firms belonging to these industries represent only 12.18% of the country's manufacturing. Therefore, a group of innovative firms are responsible for more than half of the investments in R&D. These characteristics suggest that for this type of industrial activity, geographical proximity can be an important determinant of productivity, since it is directly associated with knowledge spillovers.

Table 2.1 Breakdown of plants by CNAE 2-digit codes

CNAE 2-digt/Industry Name		# of 3-digt	Tech level	# of plants		avg. emp.		% college		% inovation	
				2006	2015	2006	2015	2006	2015	2006	2015
10	Food processing	9	low	32445	42935	36.08	35.62	3.68	6.38	0.32	0.85
11	Beverage production	2	low	2164	2301	48.13	56.42	9.94	12.42	–	0.74
12	Tobacco products manufacturing	2	low	208	218	76.15	63.59	13.57	19.09	2.88	–
13	Textile products manufacturing	5	low	8517	9930	34.34	26.23	3.12	5.16	0.40	0.24
14	Apparel manufacturing	2	low	41467	49051	13.75	12.53	1.46	3.81	0.15	0.28
15	Leather and leather products mfg	4	low	11483	10826	33.68	31.54	1.18	2.53	0.41	0.81
16	Wood products manufacturing	2	low	14951	13857	15.25	12.72	1.80	3.64	0.29	0.04
17	Paper manufacturing	4	low	4036	4152	38.26	42.71	7.94	12.45	0.35	0.17
18	Printing & related support activ.	3	low	10293	12942	9.84	8.75	5.87	10.67	0.15	0.24
19	Petroleum & biofuels mfg	3	m-low	441	562	221.16	269.15	6.52	21.25	1.13	0.71
20	Chemical manufacturing	8	m-high	7910	8559	30.22	31.50	15.70	21.97	0.99	0.93
21	Pharmaceuticals products mfg	2	high	1130	801	73.90	128.82	27.06	39.70	6.90	2.87
22	Plastics & rubber products	2	m-low	13200	13336	29.45	30.93	4.61	7.94	0.60	0.82
23	Nonmetallic mineral products	5	m-low	19320	26461	16.78	16.15	3.57	5.59	0.29	0.25
24	Metallurgy	5	m-low	4680	3772	49.37	56.68	8.60	13.64	0.47	0.42
25	Metal products mfg	6	m-low	26587	37090	15.45	12.15	4.05	6.57	0.21	0.32
26	Computer & electronic products	8	high	2847	3276	49.35	41.59	12.41	18.78	0.42	2.59
27	Electrical machinery mfg	6	m-high	3722	4353	45.99	45.55	9.93	14.47	0.64	1.40
28	Machinery manufacturing	6	m-high	10572	13534	27.14	26.70	9.77	15.36	0.83	1.29
29	Motor vehicle manufacturing	5	m-high	4362	6047	87.61	70.47	10.61	17.78	1.33	1.54
30	Transport exc. motor vehicles	5	m-high	813	1154	82.50	85.16	13.88	16.13	1.11	1.73
31	Furniture manufacturing	1	low	14442	20451	14.84	12.52	2.32	4.46	0.09	0.78
32	Miscellaneous manufacturing	6	low	6337	12341	16.58	11.96	5.67	8.86	0.21	0.49
33	Maintenance of machinery	2	m-low	6862	18524	12.80	9.53	5.78	8.73	0.19	0.05
		103		248789	316473	25.14	22.705	5.77	9.62	0.37	0.44

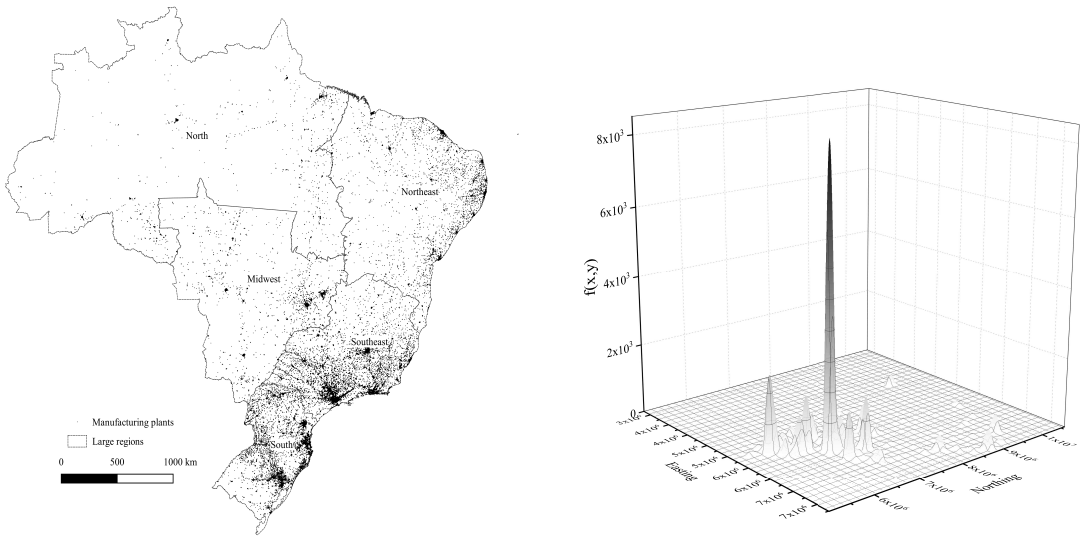
Notes: Cavalcante (2014)'s technological classification based on the compatibility of CNAE with the OECD technological classification. The college variable represents the share of workers with college degrees (including postgraduate). The innovation variable represents the share of establishments that have implemented some product innovation in partnership with other companies. Source: Authors' computations using information from RAIS and PINTEC (Innovation Survey) database provided by the Ministry of Labor and Employment and IBGE, respectively.

The pattern of geographic concentration can be seen in Figures 2.1 (a) and (b), which present a snapshot of geographic distribution of manufacturing industries plants in the country (boundaries refer to large region limits) and a surface that represents the bivariate kernel density function[6] in 2015, respectively. Note that the figure highlights the Southeast

[5]PINTEC is a triennial survey, so the data presented for 2006 and 2015 correspond to the data released for the periods 2003-2005 and 2012-2014.

[6]Quartic kernel form.

and South regions as the most densest, i.e., they are the most industrialized regions, and within these regions, patterns of location can be observed, with denser areas close to the large urban centers and on the coast. In Figure 2.1 (b) the main distortion of the spatial grid represents the SPMR, followed by the other major urban centers in the country, such as Rio de Janeiro, Belo Horizonte and Porto Alegre metropolitan regions. Accordingly, 33% and 19.8% of the high and low-tech industries are located in the five largest metropolitan regions[7] of the country (in order: São Paulo, Rio de Janeiro, Belo Horizonte, Porto Alegre and Recife), respectively (see Table A.5 in Appendix A.1).



(a) Brazil with large regions

(b) Bivariate kernel density

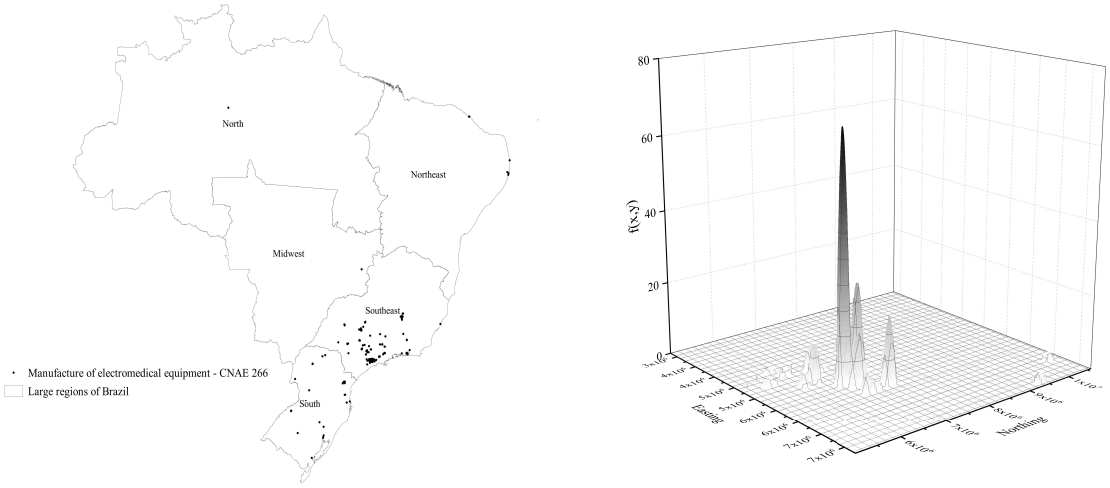
Figure 2.1 Location of all manufacturing plants in Brazil - 2015

Just as a preliminary illustration, we separate an example of high-tech and a low-tech industrial sectors[8] in Figures 2.2 and 2.3 for *manufacture of electro-medical and electrotherapeutic equipment* (CNAE 266) and *manufacture of other food products* (CNAE 109), respectively. The latter sector encompasses a wide variety of products and is characterized by low use of skilled labor. There are clearly differentiated location patterns, where plants of low-tech industries appear clearly much more sprawled. As shown by the surface in Figure 2.3 (b), besides the distortions in the large urban centers, it is also possible to observe that the plants of this industry are also located near or outside them. Note, however, that although they are more spread out, this industry follows the pattern

[7] According to the 2010 Demographic Census made available by IBGE.

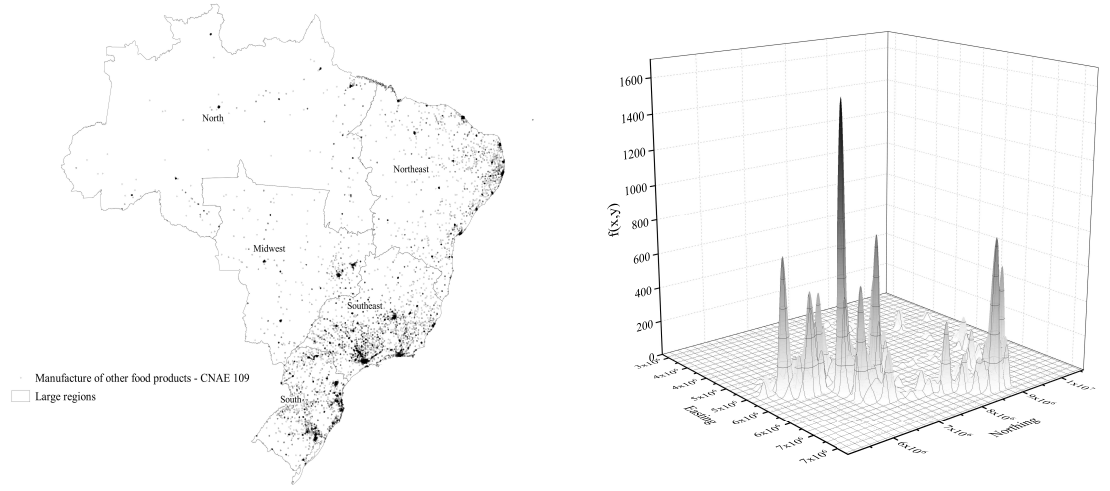
[8] The location of all high and low-tech plants can be seen in Figure A.1 in Appendix A.1

of stronger general location on the coast. For the low-tech sector, 39.7% of the total plants are located in the Southeast region, while for the high-tech sector, 79.8% of the total plants are in the same region. With respect to the urban context, for the low-tech sector 15.2% of the total plants in the Southeast are in the SPMR, in contrast with the 36% in the high-tech sector.



(a) Brazil with large regions (b) Bivariate kernel density

Figure 2.2 Location of plants making electro-medical equipment in Brazil - 2015



(a) Brazil with large regions (b) Bivariate kernel density

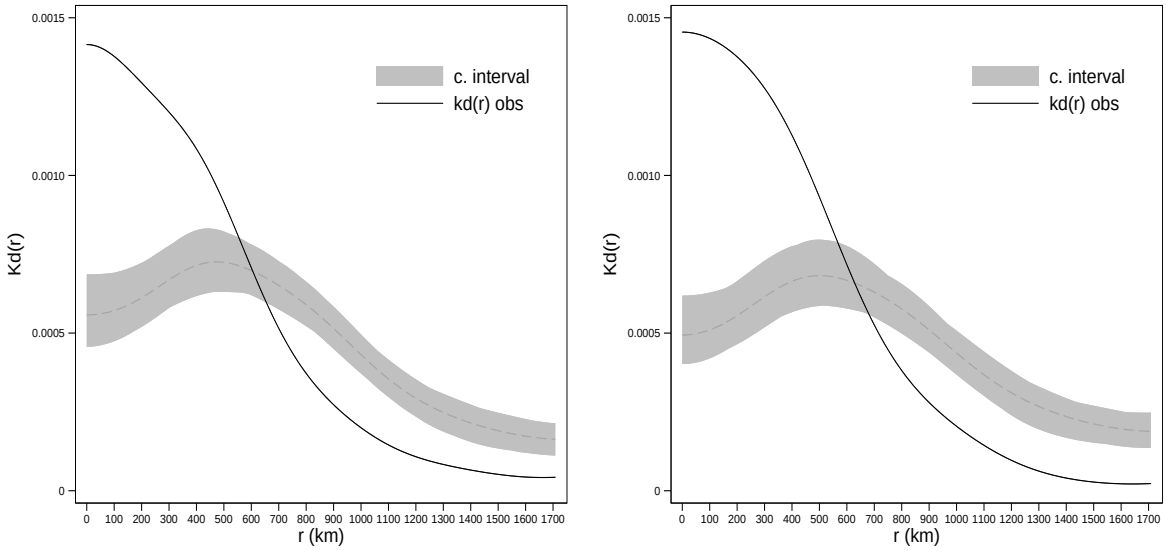
Figure 2.3 Location of food processing plants in Brazil - 2015

2.3 Location of Brazilian manufacturing

2.3.1 Initial results

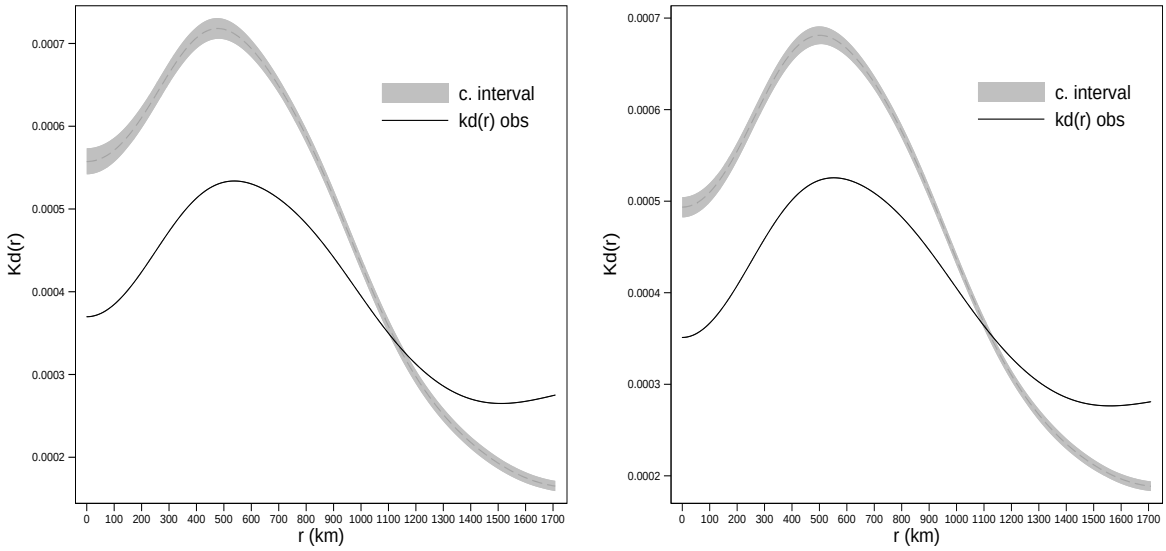
We use a distance-based measure to analyze the spatial distribution of manufacturing establishments. As outlined above, more traditional measures of concentration of productive activity are susceptible to MAUP. The index of Duranton and Overman (2005) overcomes this problem. The general idea of this index is to estimate the bilateral distance distribution between establishments – or, in its weighted version, of all employees in an industry – and then compare the estimated distribution with a set of randomly distributed bilateral distances. An industry can be classified as localized if one observes a higher K-density than that of randomly drawn distributions, or dispersed if observing a lower K-density than that of randomly drawn distributions. One can also measure the strength of localization and dispersion, Γ_m and Ψ_m, for each industry m, respectively, by the area between the observed distribution and the upper- and lower-bounds of the confidence bands. Considering localized industries as an example, one can interpret Γ_m as "excess probability" of finding another firm in the same industry closer than some distance r after controlling for the reference distribution at the 5% risk level (Behrens and Bougna, 2015, p.50). More details are provided in Appendix A.2.

To exemplify and understand the logic of the DO index, we illustrate the possible patterns with the help of Figures 2.4 and 2.5. In each of the figures, the black solid line represents the observed value of the DO index, $\hat{\mathrm{K}}d_{obs}(r)$, for the selected industry; the upper and lower bounds, $\hat{\mathrm{K}}d_{hi}(r)$ and $\hat{\mathrm{K}}d_{lo}(r)$, respectively, are represented by the extremes of the hatched area that determines the confidence interval containing 95% of counterfactual distributions. Therefore, when $\hat{\mathrm{K}}d_{obs}(r)$ is within this range, we cannot reject the null hypothesis (at the 5% level) that the specific industry is randomly distributed in space. When $\hat{\mathrm{K}}d_{obs}(r)$ is above the upper confidence band, the distribution of bilateral distances observed among the companies belonging to the industry in question exceeds the random pattern and this is interpreted as localization. When this occurs over short distances – remember that we are analyzing the whole country, but our maximum range is 1708.11 km (see Appendix A.2) – we say that plants in this particular industry are located at short distances (see Figure 2.4 (a) and (b)). At the other extreme, i.e., when $\hat{\mathrm{K}}d_{obs}(r)$ falls below the lower confidence band, bilateral distances between plants are underrepresented relative to the random pattern, and this is interpreted as dispersion (see Figure 2.5 (a) and (b)).



(a) 2006 (b) 2015

M. of electro-medical and electrotherapeutic equipment - CNAE 266

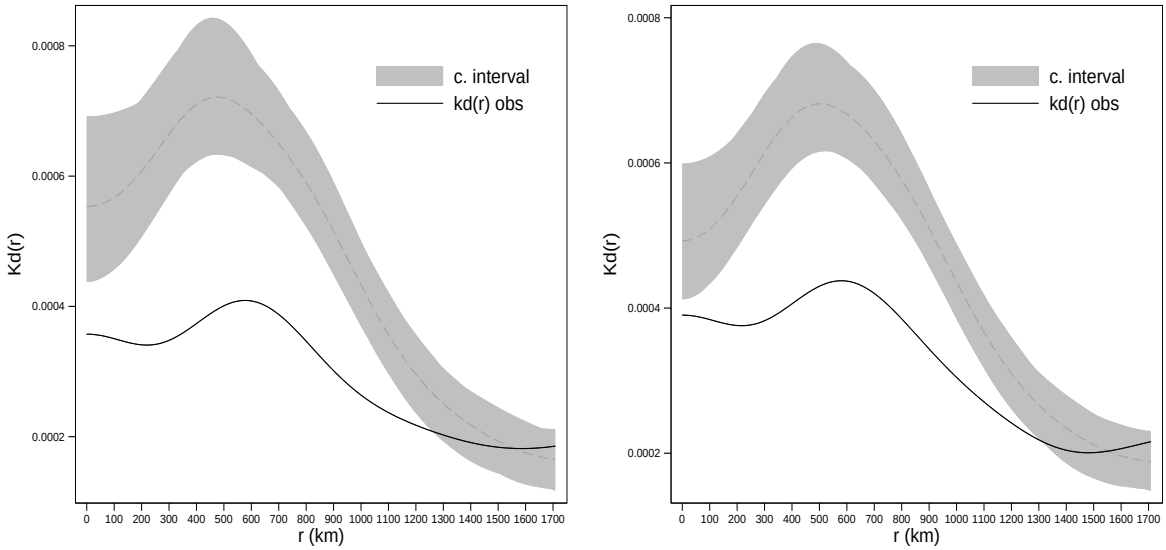


(c) 2006 (d) 2015

M. of other food products - CNAE 109

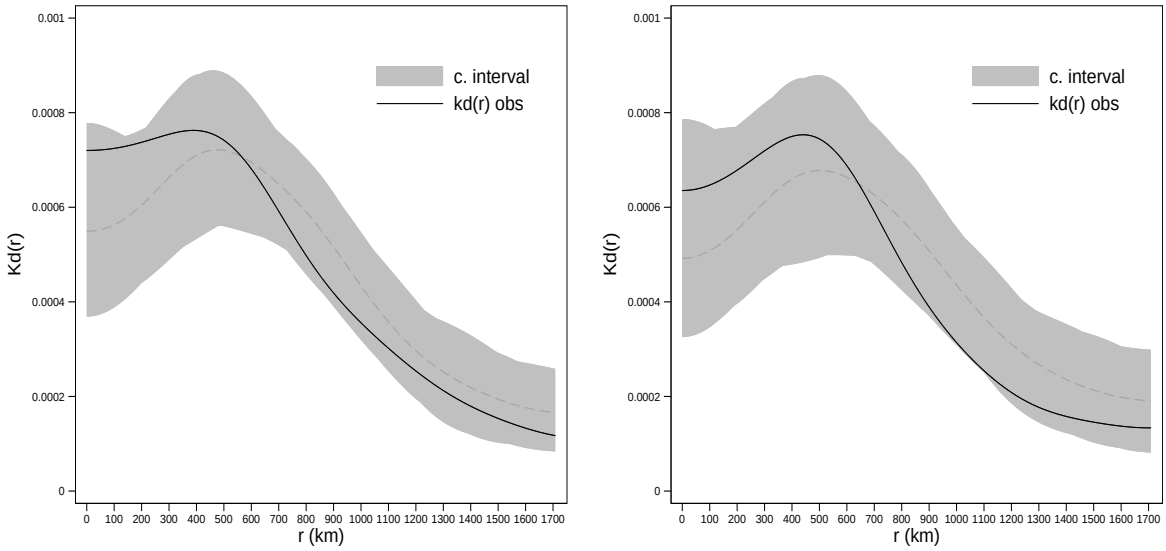
Figure 2.4 K-density estimates for selected manufacturing sectors (3-digit CNAE) located at short (a and b) and long (c and d) distances in 2006 and 2015

Figures 2.4 (a) and (b) for 2006 and 2015, respectively, indicate that the *manufacture of electro-medical and electrotherapeutic equipment* - CNAE 266 is located at short distances. Note that the peak in the K-density occurs at very short distances, indicating that the industry is overrepresented at short distances. This result confirms the patterns illustrated



(a) 2006 (b) 2015

Preservation and manufacture of fish products - CNAE 102



(c) 2006 (d) 2015

M. of pulp and paper - CNAE 171

Figure 2.5 K-density estimates for selected manufacturing sectors (3-digit CNAE) dispersed (a and b) and random (c and d) in 2006 and 2015

previously through the map in Figure 2.2, more specifically, the main cone on the surface of Figure 2.2 (b), which denotes São Paulo Paulo Metropolitan Region (SPMR). The results can be better understood by looking at the technological features of the industry.

This is a high-tech[9] industry included in a more aggregate group of industries (Computer & electronic products, 2-digit CNAE 26) characterized by the intensive use of technology[10] and employment of skilled workers (as previously shown in Table 2.1). The combination of these features with overrepresentation at short distances is consistent with the arguments associated with human capital spillovers in metropolitan areas.[11]

On the other hand, Figures 2.4 (c) and (d) show, for the same periods, an example of long-distance location (more than 1,170 kilometers). The *manufacturing of other food products*[12] - CNAE 109 is characterized by low technology, and therefore generally does not need high value-added production inputs or more sophisticated machines, and employs low skilled workers. This location pattern was previously shown in Figure (2.3); more specifically, this distance is in accordance with the distances between the main cities in the Southeast-South (e.g., the distance between São Paulo and Porto Alegre is 1,150 km) and Southeast-Northeast (e.g., the distance between São Paulo and Recife is 2,600 km). These characteristics are also observed in the other food and beverage manufacturing sectors. As part of the agribusiness production chain, the food industry is directly related to agricultural production and for these reasons tends to be more scattered. The geographic and climatic characteristics of Brazil favor the location of these industries in all regions of the country, that is, the availability of natural resources favors the pattern of location over long distances, and certainly is more important for these industries than for the high-tech sectors. Note also that our results are in accordance with the inverse relationship between transportation costs and geographic proximity. The location of large plants in the sector is possibly oriented by transport costs, and in the absence of good transport infrastructure, the costs increase, favoring scattering.

Figures 2.5 (a) and (b) show the pattern of dispersion of *preservation and manufacture of fish products* - CNAE 102 for 2006 and 2015, respectively. This industry is part of the large food manufacturing group (CNAE 10), which is labor-intensive, using low-technology,

[9]According to technological classification proposed by Hatzichronoglou (1997) for OECD countries and adapted for Brazil by Cavalcante (2014) through compatibility with CNAE 2.0, which is based on the relationship between expenditures on R&D and added value, intermediate and capital goods.

[10]Reflecting this pattern, according to data from the Innovation Survey (PINTEC) for 2014 provided by IBGE, the large group (CNAE 26) is the fourth largest investor in internal R&D activities among the sectors surveyed at the 2-digit manufacturing level in the country. It is behind only the sectors of Petroleum & biofuels manufacturing (CNAE 19), which is a strategic sector and receives major investments from Petrobras; Chemical manufacturing (CNAE 20); and Motor vehicle manufacturing (CNAE 29).

[11]About human capital spillovers in metropolitan areas, see, e.g., Ciccone and Hall (1996), Moretti, (2004a; 2004c), Fu (2007), Duranton (2016b), and Dingel *et al.* (2019).

[12]With the exception of meat products, canned fruit, vegetable and animal oils and fats, dairy products, sugar refining and coffee grinding, the other food products sectors includes all other segments.

and traditionally more dispersed than high-tech industries. As an example of the random location pattern, Figures 2.5 (c) and (d) show that for the *manufacture of pulp and paper* - CNAE 171, we do not reject the null hypothesis of randomness. This is also a low-tech industry and is a branch of Brazilian agribusiness. Like the food industry, it is also intensive in natural resources. The results obtained for these sectors, in general, present similar patterns when compared with those obtained for developed countries such as Japan (Nakajima *et al.*, 2012), some European countries (Belgium, Germany, Italy, France, Spain and UK in Vitali *et al.* (2013)) and Canada (Behrens and Bougna, 2015), another geographically large country. In comparison with the results obtained for other countries in transition, such as China (Brakman *et al.*, 2016) and Russia (Aleksandrova *et al.*, 2019), our results are also generally similar. For example, in the two aforementioned countries, the manufacture of electronic equipment appears among the most localized while the manufacture of food is dispersed.[13]

2.3.2 General results

We now examine the general patterns of location. Table 2.2 outlines the set of evidence on geographic distribution patterns for all manufacturing sectors. We consider only industries with 10 or more plants.[14] Our results indicate that 89.9% and 91% of the manufacturing sectors analyzed differ significantly from randomness at the 5% level of significance in 2006 and 2015, respectively. These patterns clearly remain high when considering the weighted version (82.83% and 80%) of the DO index and the unweighted (83.33% and 86.67%) and weighted (65.48% and 70.98%) 4-digit sector disaggregation.[15] The patterns observed are higher and more general than those obtained for manufacturing when compared with previous studies using traditional (not distance-based) geographic concentration measures for Brazil (see, e.g., Azzoni, 1986; Silveira Neto, 2005; Resende and Wyllie, 2005; Lautert and Araújo, 2007; Silva and Silveira Neto, 2009), and suggest that the manufacturing presents heterogeneous patterns of geographic distribution which persists across years. In the two years analyzed, only four industries are classified as dispersed (unweighted 3-digit

[13] As robustness check, we also compare our results obtained from K-density weighting, and clearly according to the pattern described for the location of low-tech industries, the sector of *manufacturing of other food products* - CNAE 109 appears dispersed (see Figure A.3 in the Appendix A.3).

[14] As in Duranton and Overman (2005). After the restrictions, the sample contains 99 and 100 three-digit industries (out of 103) in 2006 and 2015, respectively, and 252 and 255 four-digit industries (out of 258) in the same period.

[15] Note that although the share of total industry located at 4-digits approximates that observed at 3-digits, there are differences because the aggregation tends to mix sub-industries that exhibit different location patterns (see Duranton and Overman, 2005; Behrens and Bougna, 2015).

version) while the great majority are classified as localized. This evidence indicates that despite the changes in the spatial dimension of Brazilian industrial development, marked by expansion to other regions of the country, especially the Midwest region (Costa and Biderman, 2016; Rocha *et al.*, 2019), this process was accompanied by increased internal heterogeneity of the regions, favoring the emergence of productive clusters (Pacheco, 1999; Lima and Simões, 2010).

Table 2.2 Summary of location patterns for manufacturing

	Unweighted				Employment weighted			
	2006		2015		2006		2015	
3-digit industries	# of ind.	%	# of ind.	%	# of ind.	%	# of ind.	%
Localized	89	89.90	91	91	82	82.83	80	80
Dispersed	4	4.04	4	4	9	9.09	12	12
Random	6	6.06	5	5	8	8.08	8	8
$\bar{\Gamma}\|_{\Gamma_m>0}$	0.0247	–	0.0259	–	0.0246	–	0.0279	–
$\bar{\Psi}\|_{\Psi_m>0}$	0.0232	–	0.0196	–	0.0112	–	0.0070	–
	99[a]	100	100[b]	100	99	100	100	100
4-digit industries								
Localized	210	83.33	221	86.67	165	65.48	181	70.98
Dispersed	16	6.35	14	5.49	38	15.08	34	13.33
Random	26	10.32	20	7.84	49	19.44	40	15.69
$\bar{\Gamma}\|_{\Gamma_m>0}$	0.0244	–	0.0267	–	0.0241	–	0.0253	–
$\bar{\Psi}\|_{\Psi_m>0}$	0.0114	–	0.0133	–	0.0080	–	0.0048	–
	252[c]	100	255[d]	100	252	100	255	100

Notes: See the Appendix A.2 for details on how to compute Γ_m and Ψ_m. The values of $\bar{\Gamma}|_{\Gamma_m>0}$ and $\bar{\Psi}|_{\Psi_m>0}$ are averages for all significantly localized industries and for all significantly dispersed, respectively. After the restrictions imposed by the minimum of 10 plants in each sector: [a] four and [b] three 3-digit sectors were discarded (out of 103) and [c] six and [d] three 4-digit sectors were discarded (out of 258). The confidence bands are computed using 1,000 bootstrap replications. In the top panel we present a summary of the results considering the classic and weighted versions of the DO index and the 3-digit level of sectoral disaggregation, while the lower panel shows the summary of the results when considering the classical and weighted versions and sectoral disaggregation at the 4-digit level. Source: Prepared by the author based on estimates.

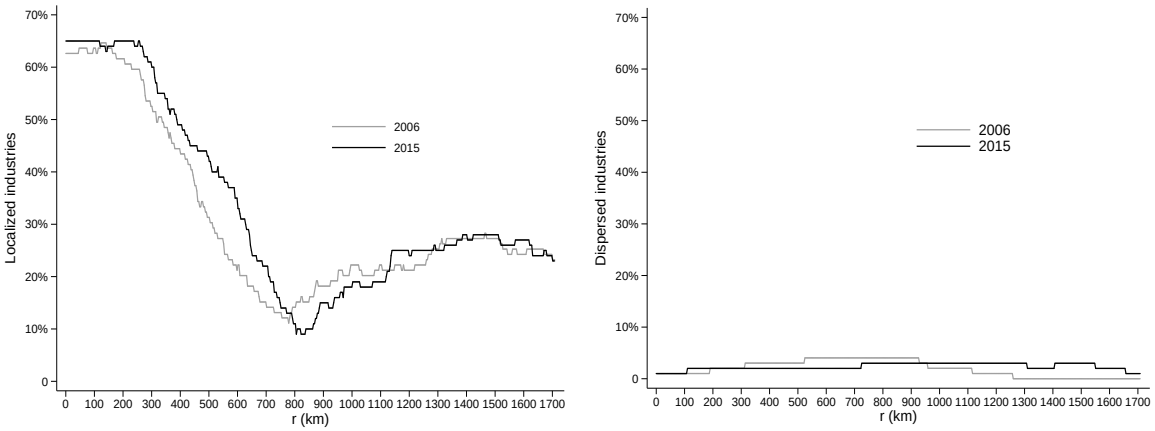
Consistent with the historical process of industrial development and the greater mobility of factors in Brazil, our evidence indicates that Brazilian manufacturing has stronger localization patterns than those found for other developing countries like China (Brakman *et al.*, 2016) and Russia (Aleksandrova *et al.*, 2019), and much stronger than those found for developed countries like UK (Duranton and Overman, 2005), Japan (Nakajima *et al.*, 2012), Belgium, France, Germany, Italy, Spain (Vitali *et al.*, 2013), and Canada (Behrens and Bougna, 2015). For example, Brakman *et al.* (2016) shows that 81% of manufacturing is localized in China while Aleksandrova *et al.* (2019) show that 80% follows the same pattern in Russia. In developed countries, Duranton and Overman (2005) for UK found that 52% of manufacturing to be localized, while Nakajima *et al.* (2012) for Japan found

a percentage of 50%, and the percentages for Germany and Canada are around 71% and 60% according to Koh and Riedel (2014) and Behrens and Bougna (2015), respectively.

In a decade, when we consider the classic version of the DO Index (both 3-digit and 4-digit), Brazilian manufacturing presented stronger localization patterns, especially when considering more disaggregated industries, 11 of which became localized in 2015. Furthermore, the strength of localization – as measured by the average across all localized sectors, $\bar{\Gamma}|_{\Gamma_m>0}$ – is greater in 2015. However, in the weighted version at the 3-digit level, we observe a subtly inverse process – a lower number of localized industries –, although the strength of localization is greater in 2015 and at the 4-digit level, the number of industries localized is higher. This evidence is not available in studies using traditional measures of spatial concentration of employment (such as the EG index). In general, except in technology intensive sectors, the authors mentioned have argued that the EG index shows a decreasing trend during recent decades, which supports the spatial deconcentration argument. Our approach is more general and can identify patterns of spatial location that occur at short or long distances separately for plants and employment. Therefore, our results indicate that although manufacturing employment has tended to deconcentrate in recent decades (Resende and Wyllie, 2005; Lautert and Araújo, 2007; Rocha *et al.*, 2019), this process occurred simultaneously with an increase in the number of manufacturing sectors with statistically significant localization patterns.

By exploring the advantages of the DO index, we can obtain the distances at which the localization or dispersion occurs. Figures 2.6 (a) and (b) illustrate the share of localized and dispersed industries between 0 and 1700 kilometers, respectively. Figure 2.6 (a) indicates that the number of industries that are located at 0-200 km remained relatively stable (around 62%) in 2006 and from there decreased sharply with the distance up to approximately 800 km, when the number of industries located at long distances leave the upward curve. The reason why this pattern is observed at short distances is associated with the fact that most industries are located in or near large urban centers (the 0-200 km range is at the scale of Brazilian metropolitan regions). The pattern observed at intermediate distances (300-700 km) is associated with location between large urban centers and with long distances associated with the location of the industry between the large regions of the country. Additionally, another important feature of these results is that long distance localization is mainly associated with low-tech industries. We will return to that point later when investigating the differences between low- and high-tech industries' location patterns. In 2015, this pattern repeated but with a shift of the curve, i.e., a larger number of industries located over shorter distances (around 65%). Figure 2.6

(b) gives the share of dispersed industries. Only 1% of industries are dispersed within 0-100 km and 3% within 200-400 km.



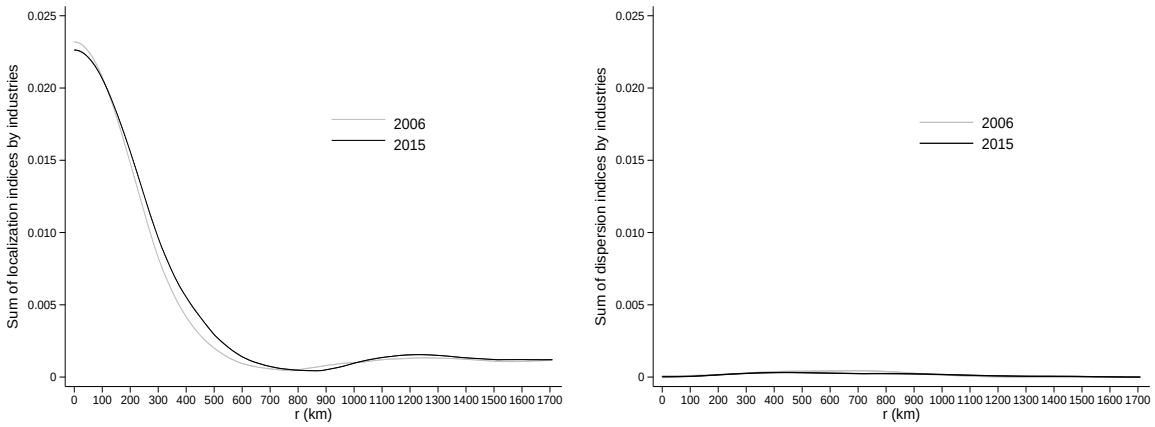
(a) Significant localization by distance

(b) Significant dispersion by distance

Figure 2.6 Share of localized and dispersed industries (3-digit), 2006 and 2015

Although the deviations from randomness are presented in Figure 2.6, the extent of these deviations is not (see, e.g., Nakajima *et al.*, 2012; Brakman *et al.*, 2016). So, to measure the extent of localization across all industries for each distance, we use summation over industries as measure of the extent of deviation at any given distance, i.e., $\Gamma(r) = \sum_m \Gamma_m(r)$ for localization and $\Psi(r) = \sum_m \Psi_m(r)$ for dispersion. Figures 2.7 (a) and (b) report this set of information for 2006 and 2015. As shown in (a), the extent of localization is greater at shorter distances, around 0-100 km. On the other hand, much less pronounced, the extent of dispersion is greater between 600-800 km, as shown in (b). Note also that no major changes occurred between 2006 and 2015. Although the share of localized industries was higher in 2015, as shown in Figure 2.6 (a), the intensity of agglomeration was smoother in 2015. This information suggests a change in the intensity of the localization patterns over the years (we present evidence about the comparison of K-densities across years in Appendix A.3).

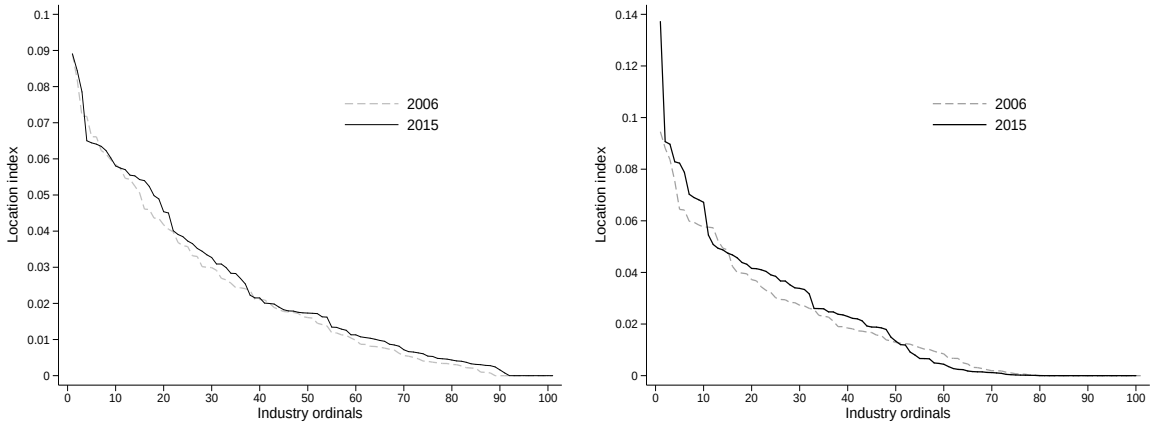
Last, we rank the industries in descending order of localization indices. Figures 2.8 (a) and (b) present this ranking for unweighted and weighted versions in 2006 and 2015, respectively. In both cases, the distribution is clearly skewed, with a group of 10 industries with higher levels of localization, and there are no major changes in the localization patterns between 2006 and 2015. This suggests the strength of the forces that favor agglomeration in the country. Furthermore, Figure 2.8 provides an interesting overview of



(a) Localized (b) Dispersed

Figure 2.7 Extent of localization and dispersion (3-digit), 2006 and 2015

small changes in the localization pattern during the period studied. Note that although the number of localized industries was lower in 2015 in the weighted version of the DO index (as shown in Table 2.2), the localization increased at the very top and decreased at the bottom of the distribution (Figure 2.8 (b)), suggesting that the trend of spatial deconcentration of employment does not affect all industries in the same way.



(a) 3-digit unweighted (b) 3-digit weighted

Figure 2.8 Rank-order distributions of location indices for manufacturing sectors

2.3.3 Sectoral scope

As previously mentioned, the high-tech industries present different localization patterns from the low-tech industries, as they are more strongly localized at short distances (e.g., sectors presented in Figure 2.4). So, we now analyze these two types of industries separately. Initially, to identify the industries presented in Figure 2.8 (a), Table 2.3 presents the rankings of the 20 most localized industries in 2006 and 2015, and all industries classified as dispersed and randomly distributed. In both years, 70% of the most localized industries are technology intensive (e.g., *manufacture of electro-medical and electrotherapeutic equipment* - CNAE 266, *manufacture of parts and accessories for automotive vehicles* - CNAE 294, *manufacture of measuring, testing and control devices and instruments* - CNAE 265 and *manufacture of railway vehicles* - CNAE 303).

When comparing our results to those obtained for other developing countries with large territorial extensions, such as China and Russia, the industrial location patterns are similar. As Brakman *et al.* (2016) shows, for China, for example, “Machinery Manufacturing" and “Textile" sectors are among the most localized, while Aleksandrova *et al.* (2019) shows that these sectors are also among the most localized in Russia. Furthermore, traditional industries such as food processing are more dispersed. In fact, as indicated in Table 2.3, among the 20 most localized, there is no food industry. Labor-intensive industries (low-skilled) such as the *manufacture of food products* (CNAE 10) have weaker concentration patterns, pointing to a trend of spatial deconcentration, as studies of the spatial concentration of employment in Brazil show based on discrete concentration measures (see, e.g., Silveira Neto, 2005; Resende and Wyllie, 2005; Lautert and Araújo, 2007; Almeida and Rocha, 2018). Our results indicate that 89% of the food manufacturing sectors presented statistically significant location patterns over long distances in 2006 and 2015. These results are associated with the spatial structure of the labor market. The spatial distribution of education is very heterogeneous. Poorer regions have fewer skilled workers (Suliano and Siqueira, 2012; Silva and Silveira Neto, 2015). On the other hand, the spatial distribution of low-skilled labor is more homogeneous, which favors the lower levels of concentration of industries intensive in this type of workers.

To illustrate the differences between the localization patterns between the low- and high-tech industries, analogous to Figure 2.6, we present in Figures 2.9 (a) and (b) the share of high- and low-tech localized industries across all distances, respectively. There is a clear difference: while high-tech plants have a strong localization (around 78%) at short distances (0-200 km), low-tech plants have a smaller share (around 50%). Furthermore,

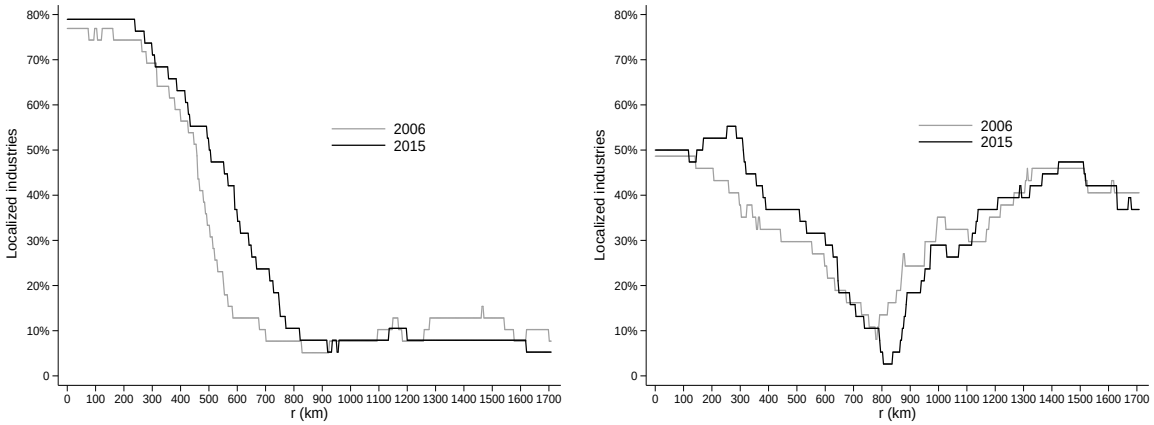
Table 2.3 Twenty most localized, dispersed and random industries in 2006 and 2015

CNAE 3 digit	Industry name	Tech level	Localization/Dispersion ranking			
	Most localized		$\Gamma_{m,2006}$	2006	$\Gamma_{m,2015}$	2015
266	M. of electromedical and irradiation equipment	high	0.0718	3	0.0891	1
304	M. of aircraft	high	0.0436	18	0.0785	2
294	M. of parts and accessories for motor vehicles	m-high	0.0718	4	0.0650	3
263	M. of communication equipment	high	0.0461	16	0.0644	4
286	M. of machinery for industrial uses	m-high	0.0584	10	0.0641	5
265	M. of measuring, testing and control instruments	high	0.0661	5	0.0635	6
154	M. of parts for footwear, of any material	low	0.0820	2	0.0623	7
133	M. of knitted and crocheted fabrics	low	0.0357	25	0.0603	8
261	M. of electronic components	high	0.0573	11	0.0580	9
132	Weaving, not knitted or crocheted	low	0.0661	6	0.0575	10
279	M. of electrical equipment not otherwise specified	m-high	0.0597	9	0.0570	11
284	M. of machine tools	m-high	0.0623	7	0.0555	12
274	M. of lamps and other lighting equipment	m-high	0.0417	20	0.0553	13
153	Footwear manufacturing	low	0.0546	12	0.0543	14
322	M. of musical instruments	low	0.0460	17	0.0540	15
282	M. of general-purpose machinery and equipment	m-high	0.0525	14	0.0524	16
281	M. of engines and transmission equipments	m-high	0.0506	15	0.0498	17
303	M. of railway vehicles	m-high	0.0890	1	0.0489	18
209	M. of miscellaneous chemical products	m-high	0.0369	23	0.0454	19
273	M. of equipment for distribution of electrical energy	m-high	0.0543	13	0.0450	20
	Dispersed		$\Psi_{m,2006}$		$\Psi_{m,2015}$	
102	Preservation and manufacture of fish products	low	0.0575	1	0.0423	1
301	Shipbuilding	m-low	0.0124	3	0.0293	2
267	M. of optical and cinematographic equipments[a]	high	–	–	0.0063	3
272	M. of batteries and electric accumulators[a]	m-high	–	–	0.0005	4
122	M. of tobacco products[b]	low	0.0196	2	–	–
192	M. of oil products[b]	m-low	0.0036	4	–	–
	Random					
171	M. of pulp and other pulp for papermaking	low				
211	M. of pharmaceutical products	high				
292	M. of trucks and buses	m-high				
204	M. of man-made fibres	m-high				
183	Reproduction of recorded materials on any medium	m-low				
193	M. of biofuels[c]	m-low				
252	M. of tanks, metal containers and boilers[c]	m-low				

Note: Γ_m and Ψ_m are computed at 1708.11 kilometer distance. Column 3 presents the levels of technological classification by manufacturing sectors: low, medium-low (m-low), medium-high (m-high), and high. We consider as technology-intensive industries those with m-high or high levels. "M." is manufacture. Column 7 show a ranking of Γ_m in 2015 and in column 5 we present the industry's position in the 2006 ranking, allowing comparison of the changes in the localization pattern of each sector. [a] are industrial sectors classified as dispersed only in 2015, while [b] are those dispersed only in 2006 and [c] are the sectors that presented random location pattern only in 2006. Source: Prepared by the author based on estimates.

unlike high-tech industries, low-tech plants also exhibit a similar pattern over long distances (around 48%). Note also that for low-tech industries there is a strong reversal of the curve near 800 km both in 2006 and 2015 (Figure 2.9 (b)), which is not observed for high-tech plants, especially in 2015 (Figure 2.9 (a)). As previously presented, the pattern of location

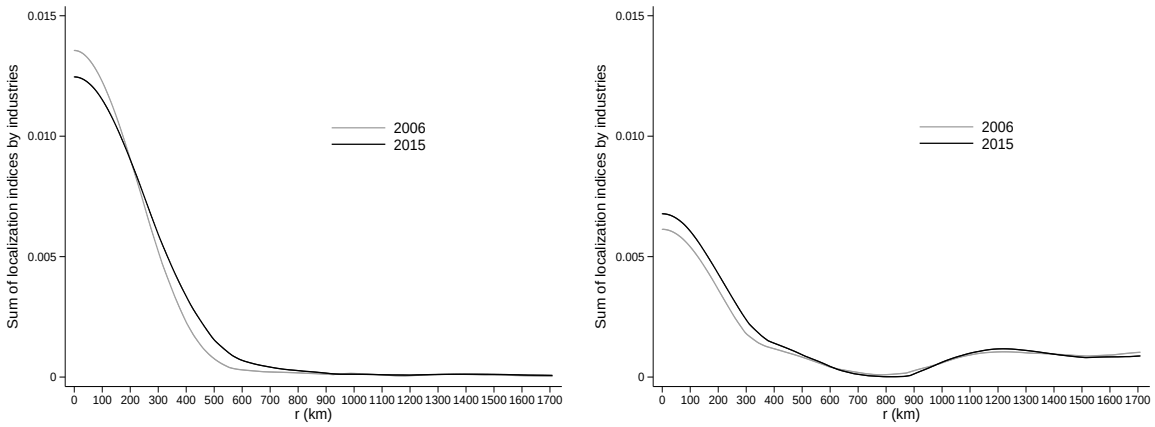
of high-tech industries within a radius of 0-700 km represents the localization of these plants in major Brazilian cities. This pattern at short distances remains when examining the data in a more aggregated form, at 2-digit level, as in Table A.6 in the Appendix A.3, where one can see that the average maximum distances for high-tech industries (e.g., *transport equipment except motor vehicle manufacturing* - CNAE 30 with 472.43 km and *computer & electronic products* - CNAE 26 with 506.97 km) are shorter than those observed for low-tech industries (e.g., *food manufacturing* - CNAE 10 with 1655.88 km and *beverage production* - CNAE 11 with 1708.11 km).



(a) High-tech (b) Low-tech

Figure 2.9 Shares of localized industries by technology group (3-digit), 2006 and 2015

These differences are even clearer when analyzing the extent of localization across high and low-tech industries for each distance, as shown in Figures 2.10 (a) and (b), respectively. Initially, when we compare the side (a) with the side (b), notably the high-tech plants are more strongly located at short distances, with the highest values between 0-100 km, suggesting that geographical proximity is more important for these sectors. This evidence is clearly in accordance with the map previously presented in Figure 2.2 – and in Table A.5 in the Appendix A.1 – about location patterns of high-tech industries in metropolitan regions, although here we are considering all high-tech sectors. Furthermore, as observed for the share of localization of low-tech industries, the localization also increases from 900 km, in clear contrast to the pattern observed for high-tech industries.



(a) High-tech (b) Low-tech

Figure 2.10 Localization indices by technology group (3-digit), 2006 and 2015

2.4 Conditioning of manufacturing localization

We have observed so far that some industries present patterns of geographic location at shorter distances (high-tech industries) while others (low-tech industries) present patterns at longer distances. What are the potential drivers of this agglomeration of industries? To provide an association of this phenomenon to the local externalities, we use the location index as dependent variable in a multivariate regression model as a function of economic forces potentially associated with the location pattern of manufacturing activity. Briefly, these economic arguments for understanding firms' location patterns derive from traditional location factors about external economies (via sharing, matching and learning)(Marshall, 1890; Duranton and Puga, 2004), transport cost (Krugman, 1991b), natural advantages associated with proximity to inputs (Rosenthal and Strange, 2001; Ellison *et al.*, 2010), local market structure (Glaeser *et al.*, 1992; Combes, 2000) and scale economies. Given the longitudinal structure in which our data are organized, we explore the observed and unobserved characteristics fixed in time specific to each CNAE 3-digit industry. Most studies that have performed this type of analysis have used only cross-section data. Formally, the general specification is given by:

$$\Gamma_{mt} = \alpha + \mathbf{X}_{mt}\beta + \mathbf{Z}_{mt}\theta + \alpha_m + \eta_t + \epsilon_{mt} \tag{2.1}$$

where Γ_{mt} is the location index for industry m and year t; $\mathbf{X}_{mt}$ is a matrix formed by the explanatory variables discussed below; $\mathbf{Z}_{mt}$ is a matrix formed by the control variables;

α_m are sector-specific fixed effects in time; η_t is a time-specific fixed effect; and ϵ_{mt} is an error term.

The vector of parameters of interest is β, which captures the effects of Marshallian agglomeration forces, transport cost, natural advantages and competition on industry localization. The first local factor potentially associated with location patterns of manufacturing activity is based on Krugman (1991a)'s model. That model shows that labor pooling can lead to spatial concentration. Consider, for example, that many firms are locally concentrated because of the abundant supply of local labor. Due to the large number of firms, the demand for labor at the industry level tends to be more stable than demand at the firm level, this occurs because while some firms are hiring, others are firing. Since aggregate demand for labor remains relatively stable, local wages tend not to vary much. So, the individual firm can hire more without major changes in labor costs. The idea is that through the concentration of workers, firms can benefit by sharing the risks (see Krugman, 1991a; Duranton and Puga, 2004; Combes and Duranton, 2006; Ellison *et al.*, 2010). Following the strategy of Overman and Puga (2010), we include in our regressions a direct measure for labor pooling that captures how much idiosyncratic volatility is faced by individual establishments in each sector. From the way it is constructed, this variable captures risk sharing effects of labor pooling (see details in Appendix A.3).

Also in line with one of the fundamental arguments of agglomeration economies, education level can be related to patterns of location (Ciccone and Hall, 1996; Moretti, 2004a, 2004c; Fu, 2007; Duranton, 2016b; Dingel *et al.*, 2019). As Greenstone *et al.* (2010) pointed out, knowledge spillovers can occur by sharing knowledge and skills through the formal and informal interaction of workers. Through the concentration of skilled workers, firms and workers can benefit from technological spillovers (Storper and Venables, 2004; Kolko, 2010; Lychagin *et al.*, 2016) and the urbanization of high human-capital industries is evidence of the role that density plays in facilitating the flow of ideas (Glaeser and Kahn, 2001). We include the percentage of workers with at least one college degree in the sector to capture that effect on the location index.

As established by the model presented by Venables (1996) for sharing intermediate inputs, firms tend to be more productive when located close to each other by the ability to share the same suppliers of intermediate inputs. So, the location of input providers can importantly affect the location decision of firms and favor agglomeration (see, e.g., Holmes, 1999; Rosenthal and Strange, 2001; Billings and Johnson, 2016). To include these important effects in our analysis of manufacturing localization, we use detailed information from Brazilian input-output tables to construct measures of the strength of

interactions between industries. To be more precise, the input-output coefficients are computed as the sum of shares relative to total manufacturing inputs and outputs of each industry. Official input-output tables provided by the IBGE are not available for all years of our sample and are also not aligned with CNAE 2.0 classifications. To deal with the absence of annual data, we use the input-output tables estimated based on the method of Guilhoto and Sesso Filho (2005, 2010) from the Brazilian National Accounts data, made available by the Nucleus of Regional and Urban Economics (NEREUS) of the University of São Paulo (USP) (details of cross-referencing of tables and CNAE 2.0 classes and adjustments can be found in the Appendix A.3). As shown by the authors, for the years when the official tables are available (only 2010 and 2015 in the period of our sample), the estimated tables are not statistically different from the official ones and are therefore a good alternative.

According to the traditional arguments of location models based on transport costs (Krugman, 1991b; Fujita *et al.*, 1999; McCann, 2013), this is an important determinant of firms' location in space. When inputs are far from their market, firms will weigh the distance between customers and suppliers based on the transport costs (Marshall, 1890). Despite their fundamental theoretical role in spatial location models, empirically there is little evidence about how transport costs drive the spatial location patterns of industries (Behrens *et al.*, 2018). Unlike most studies about related subjects, which use proxies based on distances or infrastructure for transport costs (e.g., Chandra and Thompson, 2000; Duranton *et al.*, 2014), Behrens *et al.* (2018) and Behrens and Brown (2018) used a direct measure obtained from the Origin-Destination Survey for Canada. The authors reported that measuring transport costs directly helps better understand how they affect geographic patterns of economic activity. Our measure, although not as sophisticated as those presented by the aforementioned authors, is obtained directly from the freight cost data available from the Annual Industry Survey (PIA) of the IBGE. Basically, this variable is given by relative freight costs, i.e., freight costs divided by the value of industrial production (real values).

Additionally, we also use proxies to capture the effects of spatial heterogeneity and competition on industrial concentration. First, some regions simply possess better natural environments for certain industries – a classic example is natural resource-based industries such as manufacture of oil products and biofuels. Spatial concentration can occur based on these natural cost advantages (see, e.g., Ellison and Glaeser, 1999; Ellison *et al.*, 2010). Like Rosenthal and Strange (2001), we use the industry-specific ratio of energy and water expenses to the value of production to measure energy input cost and water-related costs

(both in real values), respectively, for including the importance of natural advantages associated with proximity to inputs. Second, traditional arguments of externality models indicate that innovative companies realize that some of their ideas will be imitated on by their neighbors without compensation, so local competition reduces the returns to the innovator. On the other hand, it also increases pressure to innovate. Thus, the impact of competition on information spillovers is ambiguous and difficult to separate empirically (see, e.g., Glaeser *et al.*, 1992; Combes, 2000). We include this effect in our empirical model by adapting the local competition indicator used by Glaeser *et al.* (1992) to our sectorial data, i.e., number of firms per worker in this industry relative to the number of firms per worker in all manufacturing industries in Brazil. A high value means that the industry has more firms relative to size in all manufacturing sectors.

Finally, using the PIA database, we obtained control variables. The first control variable used is associated with rental expenses. As highlighted by Dekle and Eaton (1999), the competition for scarce land in large cities provides a centrifugal force to offset centripetal agglomeration effects. These authors pointed out that for a firm to locate in a region with high rent costs, the region should yield productivity benefits to the firm that are higher than those of regions with lower rents. So, we use the industry-specific ratio of rental expenses to the value of production to control this effect. Our last control is associated with tax expenses. Adapting Baldwin and Krugman (2004)'s argument to the geographical context within a country, other things being equal, producers will move to whichever region has the lowest tax rates. This issue is especially important in Brazil due to fiscal disputes[16] among states and municipalities (Nascimento, 2008) and because the country has one of the highest tax burdens in the world (Afonso *et al.*, 2005). So, we use the industry-specific ratio of tax expenses to the value of production to control this effect.

Table 2.4 summarizes the results of regressions of location measures at industry-level as a function of proxies for labor pooling, knowledge spillovers (college degree), input sharing, transport costs, natural advantages associated with proximity to inputs, and competition when we control for other time-varying local factors, scale economies captured by the industrial fixed effects, and time-specific fixed effects. Given the limitations imposed by our input sharing variable, we estimate the specifications with and without this variable. The coefficient associated with labor pooling variable shows positive and significant values in all specifications, with and without our input sharing variable or control variables, after including time-specific fixed effects, i.e., on average, plants that face more idiosyncratic

[16]When states or municipalities compete to offer greater “comparative advantages" to private initiative by offering investment incentives.

shocks relative to their industry are more spatially localized (Overman and Puga, 2010). These results indicate a positive association between plant-industry hiring variance and industry concentration according to the classic Marshall (1890) arguments formalized by Krugman (1991a). This mechanism that favoring industrial concentration is especially important for Brazil due to the high regional mobility and the profile of migrant workers. In general, migrants are young and more qualified, and seek the amenities and greater employment opportunities in the large urban centers of the Southeast region (Barbosa *et al.*, 2010). This makes large cities where the labor market is more competitive, specialized and dense act as polarizing centers of firms, reinforcing the agglomeration and allowing better adaptation of firms to productivity shocks. This positive correlation was also found in previous studies of other countries (see, e.g., Rosenthal and Strange, 2001; Overman and Puga, 2010) and for Brazil (see, e.g., Almeida and Rocha, 2018; Ferreira *et al.*, 2019) using the EG index. Furthermore, in columns 4-6 we include the interaction of the labor pooling variable with a technology dummy that assumes value equal to one if the sector is high-tech and zero otherwise. The idea is to explore whether this type of externality is more important for high-tech industries. Although the interaction coefficient is positive and significant in column 4, it is not significant in the others specifications, suggesting it is not clear if this mechanism is associated with the pattern previously observed for the high-tech industries.

The coefficient estimated for the percentage of college degree in individual industries is not statistically significant, but the coefficient associated to interaction high-tech×college is positive and significant. This evidence indicates that, on average, high-tech industries are more localized when their workers are more qualified, which suggests the presence of knowledge spillovers. As documented in the literature (see, e.g., Moretti, 2004a, 2004c; Greenstone *et al.*, 2010), plants may be more productive when located in cities with a larger share of skilled workers, which favors industrial concentration. As shown in Table 2.1, high-tech industries have a larger share of college educated workers in the total, and these industries are more often located at short distances in comparison with low-tech industries (Figures 2.9 and 2.10). In this context, the gains through the learning effect appear as a force favoring the location at short distances, consistent with the location of these industries in and between large urban centers (see Figure 2.2).

For transport costs, all coefficients are negative and strongly significant, indicating that, on average, high transport costs are associated with lower geographic concentration – in line with findings about transport costs for Canada (Behrens *et al.*, 2018; Behrens and Brown, 2018) and Russia (Aleksandrova *et al.*, 2019). Note also, for the last result, that

Table 2.4 Conditions affecting manufacturing location

	Dependent variable: $\ln(\Gamma_m + 1)$					
	(1)	(2)	(3)	(4)	(5)	(6)
Labor Pooling	0.0033**	0.0033**	0.0036**	0.0029**	0.0029**	0.0032**
	(0.0015)	(0.0015)	(0.0014)	(0.0015)	(0.0015)	(0.0014)
College	0.0051	0.0052	-0.0011	-0.0001	0.0004	-0.0056
	(0.0084)	(0.0084)	(0.0081)	(0.0086)	(0.0086)	(0.0084)
Transport costs	-0.0534**	-0.0597**	-0.0556**	-0.0549**	-0.0597**	-0.0554**
	(0.0243)	(0.0245)	(0.0232)	(0.0242)	(0.0244)	(0.0232)
Transport costs^2	-0.0079***	-0.0084***	-0.0082***	-0.0080***	-0.0083***	-0.0080***
	(0.0028)	(0.0028)	(0.0027)	(0.0028)	(0.0028)	(0.0027)
Water	0.0052*	0.0048	0.0052*	0.0049	0.0047	0.0050*
	(0.0031)	(0.0031)	(0.0030)	(0.0031)	(0.0031)	(0.0030)
Energy	-0.0030	-0.0042	-0.0046	-0.0023	-0.0033	-0.0036
	(0.0059)	(0.0060)	(0.0058)	(0.0060)	(0.0060)	(0.0059)
Competition	-0.0247***	-0.0258***	-0.0219***	-0.0248***	-0.0256***	-0.0223***
	(0.0063)	(0.0063)	(0.0061)	(0.0063)	(0.0063)	(0.0061)
Input sharing			0.0058			0.0023
			(0.0068)			(0.0071)
Labor Pooling ×				0.0132*	0.0123	0.0102
High-tech				(0.0079)	(0.0079)	(0.0075)
College × High-tech				0.0355***	0.0325**	0.0264**
				(0.0127)	(0.0129)	(0.0129)
3-digit FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Control variables	No	Yes	Yes	No	Yes	Yes
F-statistics	5.0649	4.7257	5.3010	5.0810	4.6775	5.0699
R-squared	0.0911	0.0961	0.1216	0.1025	0.1054	0.1286
Observations	871	871	799	871	871	799

Notes: All explanatory and control variables are standardized. Heteroskedastic robust errors are given in parentheses. Significance levels: * $p \leq 0.10$; ** $p < 0.05$; *** $p < 0.01$. Source: Prepared by the author based on estimates.

we consider possible non-linearities, where the coefficient associated with the non-linear component is negative, reinforcing the idea of an inverse relationship between transport costs and relative geographic concentration. This mechanism is expected to act strongly in Brazil given the characteristics of the country's transport infrastructure. Since the 1990s, the highway mode has accounted for more than 60% of the total cargo transport in the country (Neto *et al.*, 2011). When comparing this participation with that of other developed countries of continental dimension such as USA (26%) and Australia (24%) and in transition such as China (8%) (Bartholomeu and Caixeta Filho, 2008), the effects are evident of excessive dependence on highway cargo transport. We recognize that much of this dependence is associated with the agricultural sector, but we also expect it to be an important factor for manufacturing industries, influencing their spatial location pattern.

The coefficients associated with water expenses, are positive and significant only in columns 1, 3 and 6, which indicates that, on average, the larger the participation of this

resource in the intermediate consumption, the more localized the industry will be. This may indicate that proximity to natural resources acts as a force favoring the concentration of industries. The coefficients associated with energy expenses do not provide clear evidence of the importance of this resource for industrial concentration, and are not significant in all specifications. Last, the coefficients associated with our competition proxy are negative and significant in all specifications. These results suggest that the effects of competition act in the contrary direction of localization when we control for specific industry characteristics.

The estimated coefficients of our proxy for input sharing are not significant. Our input sharing measure is the sum of the technical coefficients, and like in Aleksandrova *et al.* (2019), we work only with the manufacturing portion of the input-output tables, i.e., excluding services, primary industries, private consumption, government items, and imports/exports. Similar measures have been used in other studies on the subject in the literature (see, e.g., Rosenthal and Strange, 2001; Ellison *et al.*, 2010). We believe that the lack of statistical robustness is possibly associated with the lack of detailed three-digit data and the small variation, so this cannot be interpreted as weak evidence of input sharing in Brazil. As previously presented, we adapted the available data from the input-output tables to the CNAE 2.0 level, but unfortunately there is no 3-digit match, so we repeated the 2-digit data in the 3-digit sectors. With this aggregation, certainly some characteristics about the demand for intermediate inputs of the 3-digit sectors are lost, which can influence the statistical significance of the coefficients, besides reducing the variation. Furthermore, the mechanism generated by input sharing is best exploited when one is interested in co-agglomeration of industries.

2.5 Concluding remarks

In this study, we present a detailed and comprehensive analysis of the location patterns of manufacturing industries in Brazil, and also provide suggestive evidence of the main mechanisms associated with the observed patterns. This analysis reduces the shortage of evidence on the location patterns of manufacturing industries in developing countries that are still scarce in the literature – available only for China (Brakman *et al.*, 2016) and Russia (Aleksandrova *et al.*, 2019) – and provides insights into a wide range of factors – Marshallian agglomeration forces, transport cost, spatial heterogeneity, market structure, and scale economies – associated with the spatial localization of industries. Our evidence is

supported by distance-based measures obtained from microgeographic data for a ten-year period.

In general terms, we can highlight some aspects about the location patterns of manufacturing in Brazil. First, when we consider the level of sectoral disaggregation at the CNAE 3-digit level, our results indicate that 89.90% and 91% in the unweighted version and 82% and 80% in the weighted version of the industry is statistically localized for 2006 and 2015, respectively. In a more disaggregated way, when we consider the CNAE 4-digit level, 83.33% and 86.67% in the unweighted version and 65.48% and 70.98% in the weighted version of the industry are statistically located for the same period. This evidence indicates that Brazilian manufacturing presents higher levels of spatial location than other countries in transition (Brakman *et al.*, 2016; Aleksandrova *et al.*, 2019), and much higher when compared to developed countries (Duranton and Overman, 2005; Barlet *et al.*, 2008; Koh and Riedel, 2014; Behrens and Bougna, 2015). This is consistent with the historical process of Brazilian market-oriented development, with high inter-regional mobility of workers and with government incentives for industry spatial targeting. Moreover, these findings do not indicate a clear trend towards changes in the spatial location patterns of Brazilian industry in both the unweighted and weighted versions, although we observed in this period a reduction in income inequality, which indicates that the agglomeration forces are stronger in Brazil.

Second, when we consider the technological heterogeneity of the industries, we find that high-tech industries are more located at short distances (78% of industries) than low-tech ones (50%). This pattern is consistent with the geographic configuration of large Brazilian urban areas and indicates that the geographic externalities generated from geographic proximity can act more strongly in determining the spatial location of these industries. High-tech industries employ a larger share of college educated workers and are also more innovative, characteristics that are better exploited in large cities. Furthermore, the spatial distribution pattern of low-tech industries (e.g., *manufacturing of other food products* - CNAE 109, located over long distances) is consistent with the more homogeneous spatial distribution of low-skilled workers than high-skilled workers, present in large cities.

Last, we present evidence about the relationship between patterns of location (localization index, Γ_m) and Marshallian agglomeration forces, transport cost, spatial heterogeneity and, market structure when we control for the observed and unobserved characteristics specific to industries fixed in time and other control variables. In general, we find evidence that labor pooling, and in particular to high-tech industries, the percentage of workers

with college degrees, generate positive effects, favoring the location of plants close to each other. For different specifications, the effects remained statistically significant, indicating that when plants are located in clusters – as shown above – they can benefit from sharing local labor and be affected by learning effects and knowledge spillovers. The results also indicate that transport costs are important to industrial location, where industries with high transport costs should, on average, be relatively more dispersed. Our water expense proxy for spatial heterogeneity indicates that being close to natural resources is associated with the location of industries, although this evidence is statistically weak. There is no clear evidence that energy expenses are associated with spatial location. When controlling for the specific characteristics of industries, there is evidence that the greater the competition, on average, the more dispersed the industries tend to be. This may suggest that competition acts as a dispersion force.

Our evidence provides new insight into the patterns of manufacturing location not available in previous studies of Brazil and allows a more detailed analysis by overcoming limitations of more traditional concentration measures. Our findings generate more in-depth knowledge about the location decisions and the associated mechanisms, which can serve as support for the formulation of regional and urban development public policies in the country. In addition, our findings also suggest different ways to continue exploring these questions in the country. For example, by identifying industry clusters, we perceive different patterns associated with the technological levels of companies. This is an issue that can be explored in more detail through the analysis of the location patterns of individual sectors. Also, can explore the co-agglomeration patterns among manufacturing industries can be investigated. More detailed analyses of location patterns by size and export status are also welcome. Finally, further investigation of the causes industrial spatial concentration is needed. All these points are on our research agenda.

Chapter 3

The spatial scope of agglomeration economies in Brazil

3.1 Introduction

Agglomeration economies are one of the reasons why cities offer better jobs and provide attractive environments for more productive firms and new establishments (Marshall, 1890; Carlton, 1983; Head *et al.*, 1995). Together with other local factors,[1] a better understanding of the spatial scope of these externalities sheds light on why some places are more entrepreneurial than others (Glaeser and Kerr, 2009; Duranton, 2015). Given these issues and the political relevance of the theme to regional and urban development, the relationship between new firm location choices and the agglomeration economies has been widely studied (see, e.g., Arauzo-Carod *et al.*, 2010, for a survey). Most of these studies, however, use aggregated geographic data and do not capture microgeographic patterns that occur within cities, for example at the neighborhood level, because they assume implicitly that the agglomeration economies operate homogeneously within cities.

This paper addresses an important question still little studied, especially in developing countries, namely the spatial scope of agglomeration economies. As raised by Rosenthal and Strange (2003), what is the geographic and industrial scope of agglomeration externalities? Empirical evidence for developed countries shows that the agglomeration economies tend to be attenuated with distance. Using creation of and employment by new establishments, Rosenthal and Strange (2003) found that agglomeration economies, particularly the localization effects (own-industry employment), are attenuated around 10 km in the US. Based on other outcome variables, such as wages (see, e.g., Fu, 2007; Rosenthal and Strange, 2008; and more recently Håkansson and Isacsson, 2019 for the US and Sweden) or TFP (such as Andersson *et al.*, 2019 for the Sweden), the results are generally similar. In the context of developing economies, only Li *et al.* (2020a) provide evidence about the spatial scope of agglomeration economies (for China). The authors found that the effects

[1] Such as local industry structure, demographics, scale economies, and cost advantages associated with city characteristics (see, e.g., Glaeser, 2007; Glaeser and Kerr, 2009; Glaeser *et al.*, 2010a).

of localization economies are attenuated more rapidly than in developed countries. These results suggest that the spatial scope of agglomeration economies is different in developing countries, which can be related to the quality of urban infrastructure.

Besides the scarcity of detailed evidence on the subject, some economic characteristics also make study of the spatial scope of agglomeration economies particularly interesting in Brazil. For example, unlike China, historically there has been no restriction on worker mobility, and economic activities are more market oriented in Brazil, which can substantially affect the geographical distribution of activities. But little is known about this phenomenon. Previous works are based exclusively on between-city variation in the data (see, e.g., Barufi *et al.*, 2016; Chauvin *et al.*, 2017). Still from this perspective, recent evidence shows that manufacturing activity in Brazil is more concentrated than in other developing countries, such as China and Russia, and much more concentrated than in developed countries.[2] In addition, a better understanding of the spatial scope of these externalities is of unquestionable political relevance. Local development public policies to attract new establishments have been on the agenda of local governments for decades in developing countries, particularly in Brazil (Leff, 1972; Hansen, 1987; Tatsch *et al.*, 2015). Most of these policies, however, are not based on detailed studies of intrinsic market factors such as economies of agglomeration. Instead, they are only supported by a broad range of political interests, and are not always economically efficient (Varsano, 1997; Paes and Siqueira, 2005; 2008).

Here we seek to fill part of this gap in the literature. For this, we use a geocoded employer-employee database of Brazilian manufacturing activities. Initially, to better understand the pattern of location of manufacturing entrepreneurship, we use the non-parametric approach developed by Duranton and Overman (2005; 2008) to document both location and colocation patterns of new manufacturing plants, considering in particular the location of new entrants versus existing establishments. This preliminary data investigation provides insight about geographic proximity between entrants and incumbents establishments and identifies the industrial sectors for which these patterns are most prominent. In turn, these patterns may be associated with the entrants' locational choice due to local externalities generated by proximity to existing establishments.

In this context, taking advantage of characteristics of our database, we use exogenously defined microgeographic areas instead of the official administrative areas to examine the spatial extent of agglomeration economies on the location decisions of new establishments and on the employment levels that they choose. Specifically, we estimate the local

[2]See the chapter 2 of this dissertation.

determinants of the number of firm births per square kilometer and their associated employment levels as functions of the own-industry employment and other economic environment characteristics. Although our focus is on new establishments, that choose the locations by taking the existing economic environment as given, unobserved characteristics that affect both existing business concentrations and attract new establishments make our estimates inconsistent.

To address these potential concerns, we use different tools that involve both the wealth of detail in our data and techniques to deal with the presence of endogenous explanatory variables in nonlinear models. Different from previous studies, such as Rosenthal and Strange (2003) and Li *et al.* (2020a), we have panel data which allows us to control for any observed and unobserved heterogeneities fixed in time in different neighborhoods (or districts) within cities, and therefore allow comparing areas of one square kilometer within the same neighborhood. In this sense, our microgeographic areas of one square kilometer are generally outside of the firm's set of choices because they depend on land availability and use, minimizing potential selection problems. We also include a comprehensive set of control variables for economic environment, previously existing transportation infrastructure, geographic characteristics, and local development policies about the sites chosen by the new establishments. In addition, to address any remaining source of heterogeneity, we use a control function approach with a shift-share instrumental variable that exploits the changes in national employment growth specific to the industry to generate exogenous variation at the microgeographic area level.

The main result here shows that agglomeration economies are attenuated with distance. In particular, the effect of own-industry employment at 1 km is significantly larger than the effect of employment further away, indicating that initial attenuation is rapid. For example, adding 100 workers in the same industry up to 1 km would generate, on average, an increase of 16.8% in the expected number of births and 30% in the expected number of employees. From this same perspective, adding 100 additional employees to the 1-5 km ring would result, on average, an increase of 2.6% in the expected number of births and 10% in the expected number of employees. On the other hand, in nearly all cases for both births and new establishment employment, localization effects disappear after 5 km. The pattern of attenuation with distance remains largely robust to the inclusion of different control variables and the use of instrumental variables, strengthening the reliability of our estimates. Our results for Brazil are consistent with theoretical models of urban areas and previous empirical evidences evidence from other countries.

The paper is organized as follows. Section 3.2 presents our data source and briefly

discusses the location patterns of new establishments obtained from Duranton and Overman (2005; 2008)'s measures. Section 3.3 describes a simple conceptual framework. Section 3.4 presents the empirical approach based on count models of births and new-establishment employment. Section 3.5 discusses and compares the results them with the evidence obtained in other studies. Section 3.6 concludes.

3.2 Data and spatial location of new establishments

3.2.1 Data

We use an exhaustive establishment-level dataset from the Annual Report of Social Information (*Relação Anual de Informações Sociais*, or RAIS), made available annually by the Ministry of Labor and Employment. This database encompasses all formal establishments in Brazil. In this database, each establishment has a unique identifier, the number on the National Registry of Legal Entities (CNPJ). The data include firm address, date of opening (and closing, if applicable), number of active jobs, and the National Classification of Economic Activities (CNAE) version 2.0 (which is compatible with the International Standard Industrial Classification of all Economic Activities (ISIC) revision 4). Using these data for the period 2007-2014, geocoded every year[3], allows us to assess in detail where the new manufacturing establishments in Brazil were spatially located during the period studied. Particularly, we can divide establishments into new entrants and existing ones in each year.

Reflecting the general location pattern of the manufacturing industry (see, e.g., Silveira Neto, 2005; Lautert and Araújo, 2007; Rocha *et al.*, 2019; and the set of detailed evidence presented in the chapter 2 of this dissertation), the new enterprises are concentrated mainly in the Southeast and South regions (as seen in Table B.1 in Appendix B.1). Both in 2007-2008 (75.18%) and 2013-2014 (71.66%) the two regions concentrated more than 70% of the new establishments.[4] This was expected, given that entrepreneurial activity

[3]For each year, more than 99% of new establishments were geocoded. Establishments can change address over the years, so we consider each establishment's birth address to obtain the geographical coordinates. See details in Appendix B.1.

[4]We calculate the number of new establishments every two years. There are technical and economic criteria that justify this division. Technically, as will become clear in the next subsection, some industries have few new establishments annually, which poses a limitation to the use of distance-based measures (Duranton and Overman, 2005; 2008; Klier and McMillen, 2008). Economically, some forces operating in the economic environment around the new plants (e.g., Marshallian agglomeration forces) do not vary significantly from one year to the next.

occurs more frequently in the most dynamic regions, where more business opportunities are present. At a smaller geographic scale, the state of São Paulo represents 23.44% and 20.87% of the total of entrants in the country in the two periods, while the São Paulo Metropolitan Region (SPMR) represents 9.83% and 7.68% of the total of entrants in the country in the same period. Note also that investments are, on average, larger in São Paulo (measured by the number of workers) as a result of the larger market potential.

Based on data broken down by industry, the location patterns of new establishments are also heterogeneous. In this context, another interesting way to visualize this heterogeneity is to look at the location of entrants relative to existing establishments. Figures 3.1 (a) and (b) show these locations for two illustrative industries (entrants in 2013-2014 relative to existing establishments in 2012).[5] Figure 3.1 (a) depicts the entrants (cross) and existing establishments (circle) engaged in the *manufacture of pharmaceutical products* - CNAE 212. Similarly, in Figure 3.1 (b) shows the entrants and incumbents involved in the *manufacture of other food products* - CNAE 109. A careful look suggests that the entrants in CNAE 212 appear to be more concentrated than existing establishments. On the other hand, for entrants in CNAE 109, there does not appear to be a greatly different pattern of spatial distribution between the entrants and incumbents.

While these preliminary remarks are useful for examining visibly clear patterns, more specific relationships, such as colocalization between entrants and existing establishments, are not easily identified. In the next subsection we rigorously explore location and colocation patterns using distance-based measures.

3.2.2 Location patterns of new establishments

To better understand the geographical distribution of manufacturing entrepreneurship,[6] we use the measures developed by Duranton and Overman (2005; 2008) for localization and colocalization, weighted by employment, to assess the location patterns of new establishments for each industry at the 3-digit level. This is a distance-based method, so it is not susceptible to the modifiable areal unit problem (MAUP) common in other

[5]It is interesting to note, as shown in the chapter 2 of this dissertation, that we know the establishments in these two industries are localized relative to overall manufacturing. Therefore, when we evaluate the spatial distribution of the entrants relative to existing establishments, it is important to consider that the latter are already more localized, so even if the entrants present a location pattern similar to the existing establishments, they will still be localized relative to overall manufacturing.

[6]There is evidence that cities with younger plants and more entrepreneurship have higher growth rates, thus suggesting that local entrepreneurship is important for economic development (see, e.g., Faberman, 2011; Glaeser *et al.*, 2015; Fritsch and Wyrwich, 2016). In this context, it is worthwhile investigating in more detail the geographic location patterns of new manufacturing plants.

(a) M. of pharmaceutical products - CNAE 212 (b) M. of other food products - CNAE 109

Figure 3.1 Maps of two illustrative industries

traditional measures of concentration (e.g., Gini index and Ellison and Glaeser (1997)'s index). Specifically, from the continuous localization measure, we can assess, for example, whether new establishments follow the same location pattern, and whether they are localized or dispersed compared to existing establishments in the same industry.[7] On the other hand, from the colocalization measure, we can assess whether new establishments are colocalized relative to existing establishments, i.e., if entrants locate near to (or far from) incumbents.

Both localization and colocalization measures are obtained from bilateral Euclidean distances. In the first, we consider the bilateral distances between all entrants of a specific industry, while in the second, we consider the bilateral distances between each entrant and all existing establishments in the previous period. To illustrate the logic of this measure, we provide some illustrative examples. We begin by presenting examples from localization measurements. The black solid lines in Figures 3.2 (a) and (c) plot the K-density estimates for the entrants in 2013-2014 relative to existing establishments in 2012 in the industries of *manufacture of pharmaceutical products* - CNAE 212 and *manufacture of other food products* - CNAE 109. Graphically, one can detect whether new

[7] As in Duranton and Overman (2008), we define as counterfactual the locations that contain establishments of the same industry only. Details of the index's calculation are provided in Appendix B.2.

establishments are localized relative to existing ones when the K-density lies above its upper confidence band (delimited by the extremes of the hatched area that determines the confidence interval containing 95% of counterfactual distributions). On the other hand, we consider that the entrants are dispersed relative to the incumbents when the K-density lies below its lower confidence band for some distance and never exceeds the upper confidence band. When the K-density is within the confidence interval, we can assume that the entrants do not follow a pattern of spatial distribution different from the existing establishments. As can be seen in Figures 3.2 (a) and (c), the impression from observing the maps in Figures 3.1 (a) and (b) is confirmed. Note, for example, that the entrants in the *manufacture of pharmaceutical products* are localized relative to the industry,[8] while the entrants in the *manufacture of other food products* have location patterns similar to existing establishments.

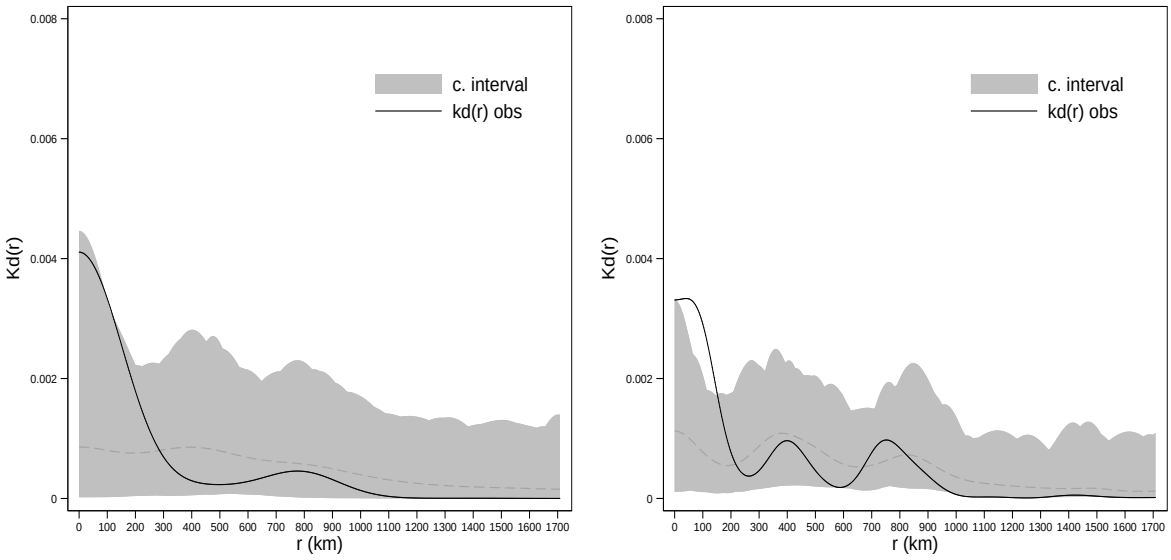
As discussed earlier, another interesting issue is to consider the colocation patterns between entrants and existing establishments. As in Duranton and Overman (2008), we also provide evidence about this. For this study, this pattern is particularly important, since our focus is on the spatial scope of agglomeration economies and therefore is directly related to the proximity between plants in the same industry. Figures 3.2 (b) and (d) plot the K-density estimates of bilateral distances between entrants and all existing establishments for the same industries. Colocation and codispersion can be detected by proceeding as before, i.e., looking at K-density estimates in the black solid lines. Two interesting patterns emerge. First, as examples of colocalization, the entrants in *manufacture of pharmaceutical products* and *manufacture of other food products* are colocalized with existing establishments. Second, the colocalization occurs at short distances.[9]

We can also get an overview for all manufacturing activity. We start with 103 industries at the 3-digit level in each period (2007-2008 and 2013-2014), and as in Duranton and Overman (2008), we drop 16 and 25 industries with fewer than 10 entrants in the first and second periods, respectively. Among the remaining establishments, 14.94% and 12.82% of employment-weighted entrants are localized in each period,[10] as can be seen in Table 3.1. In contrast, 8.05% and 10.26% of entrants are dispersed, while for most industries (around 77% in both periods), entrants do not have statistically different location patterns

[8] An opposite example can be seen in Figure B.3 (c) in Appendix B.2, which shows that the entrants in *forging and metal treatment* - CNAE 253 are dispersed relative to the industry.

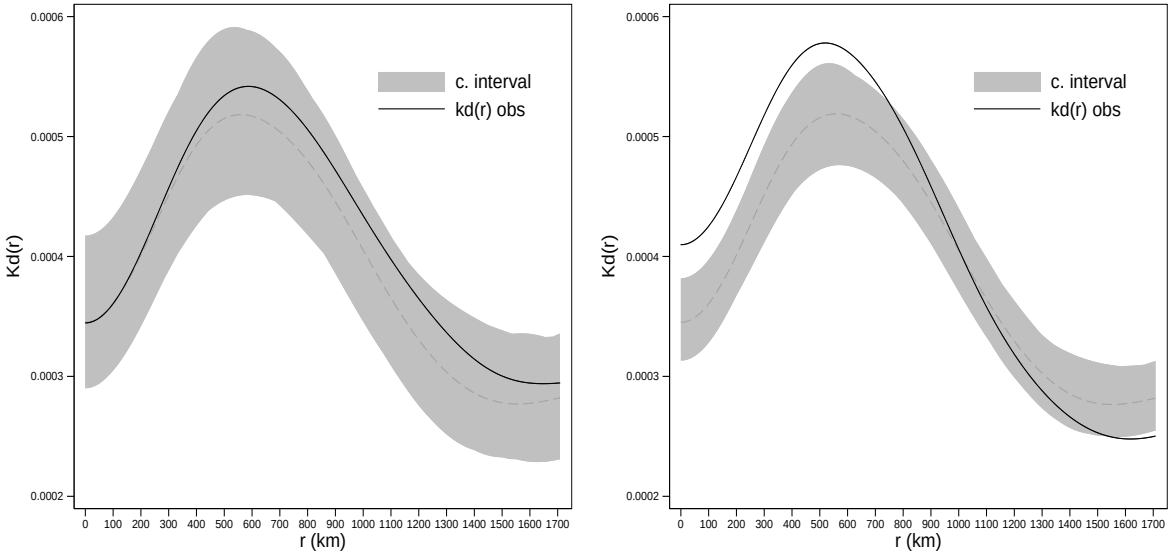
[9] Two opposite examples can be seen in Figures B.3 (b) and (d) in Appendix B.2, which show that the entrants in *printing activity* - CNAE 181 and *forging and metal treatment* - CNAE 253 are codispersed with existing establishments.

[10] These percentages are similar those found by Duranton and Overman (2008) for the UK (13%).



(a) Localization (b) Colocalization

M. of pharmaceutical products - CNAE 212



(c) Localization (d) Colocalization

M. of other food products - CNAE 109

Figure 3.2 K-density estimates for manufacture of pharmaceutical and food products

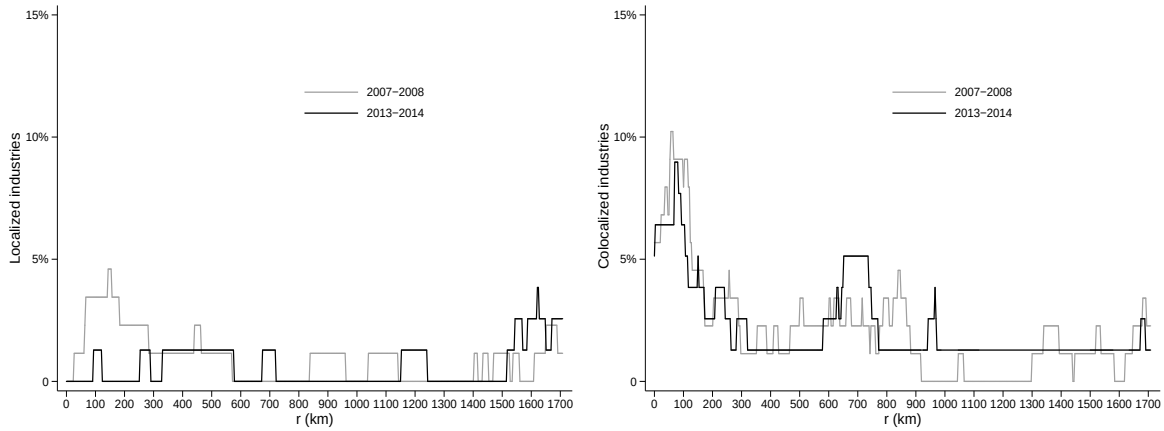
from those observed for existing establishments. We provide the K-density estimates by industry in Table B.5 in Appendix B.2. Interestingly, when we look at these percentages by distance, there is no specific pattern of localization at short distances, although there

are differences in the patterns between the two periods (see Figure 3.3 (a)).[11] In general, this evidence for Brazil is similar to that found for the UK and indicates that there is no clear tendency for manufacturing activities to become systematically more or less clustered over time because of new establishments.

Table 3.1 Localization and colocalization of employment-weighted new establishments

	Localization				Colocalization			
	2007-2008		2013-2014		2007-2008		2013-2014	
	# of ind.	%	# of ind.	%	# of ind.	%	# of ind.	%
Localized	13	14.94	10	12.82	38	42.53	21	26.92
Dispersed	7	8.05	8	10.26	19	21.84	21	26.92
Random	67	77.01	60	76.92	31	35.63	36	46.15
	87[a]	100	78[b]	100	87	100	78	100

Notes: See the Appendix B.2 for details on how to compute the distance-based measures. After the restriction imposed (minimum of 10 plants in each sector): [a] 16 industries were dropped and [b] 15 industries were dropped. Source: Prepared by the author based on estimates.



(a) Significant localization by distance (b) Significant colocalization by distance

Figure 3.3 Shares of industries in which entrants are localized and colocalized

Similarly, but looking at the general colocation patterns, 42.53% and 26.92% of employment-weighted entrants are colocalized with existing establishments in the same periods, as can be seen on the left panel of Table 3.1. These results indicate that the new manufacturing establishments in Brazil tend to be more colocalized than in the UK (9% as shown by Duranton and Overman, 2008), which may indicate that agglomeration

[11] The shares of industries for which entrants are dispersed by distance are reported in Figure B.4 (a) in Appendix B.2.

forces, especially those associated with specialization, are more important in Brazil due to the different urban structures.[12] In contrast, 21.84% and 26.92% are codispersed. Furthermore, unlike the localization results, our evidence for colocalization indicates that entrants are colocalized mainly at short distances, as can be seen in Figure 3.3 (b). For instance, among the colocalized industries for which entrants are closer to existing establishments are *manufacture of pharmaceutical products* - CNAE 212 and *manufacture of other food products* - CNAE 109 (as shown in Figures 3.2 (b) and (d); other food industries such as *fruit & vegetable canning* - CNAE 103, *manufacture of starch products* - CNAE 106, and *manufacture of furniture* - CNAE 310 (less than 10 km); *prepress and graphic finishing services* - CNAE 182 (around 40 km); *manufacture of wood products* - CNAE 162 and *finishing of textile articles* - CNAE 134 (around 70 km).[13]

Our evidence so far provides a details about location and colocation patterns of entrants in Brazilian manufacturing activities, but it does provide any information about the associated agglomeration forces. Note, for example, that there a tendency exists both in 2007-2008 and 2013-2014 for colocalization to take place at short distances (less than 70 km). This is consistent with the idea that the local economy, in particular the presence of existing establishments in the same industry, can determine the location choice of new establishments. In a more detailed analysis, similar to that conducted by Klier and McMillen (2008) for the automotive industry in the US, but looking at all manufacturing activities and focusing on the spatial scope of agglomeration economies, in the next sections we explore the importance of proximity to existing establishments to the location choice of entrants.

3.3 Conceptual framework

If there are agglomeration externalities, new establishments will tend to be located near existing ones. On the other hand, in the absence of positive local externalities generated from the concentration, it is difficult to rationally justify the proximity between establishments. In fact, in this context it is expected that the establishments will be distant from each other because of the various costs associated with spatial concentration,

[12] In line with the calculation of Duranton and Overman (2008) of unweighted K-densities, we also report the results for Brazil with in the unweighted version of the K-densities. As can be seen in Table B.4 in Appendix B.2, the percentage of both localization and colocalization are higher in both periods in the unweighted version, reinforcing the argument that proximity is more important in Brazil.

[13] On the other hand, the maximum of the distribution of the codispersed industries occurs around 1,080 km, as can be seen in Figure B.4 (b) in Appendix B.2.

such as congestion and competition effects. This issue is much broader and involves the existence of the cities themselves. There is a broad set of empirical evidence that shows this in various ways (see Combes and Gobillon, 2015; and Thisse, 2018 for a recent survey).

To evaluate the effects of agglomeration economies on the location choice of new plants, as in Rosenthal and Strange (2003), we can look at the profit maximization problem. Following the authors, we also consider that the establishments can be heterogeneous in their potential profitability. Thus, after normalizing the output price to 1, the profit of an establishment can be written as: $\pi(\mathbf{y},\epsilon) = \mathsf{A}(\mathbf{y})f(\mathbf{x})(1+\epsilon) - \mathrm{c}(\mathbf{x})$, where $\mathsf{A}(\mathbf{y})$ shifts the production function $f(\mathbf{x})$, $\mathbf{y}$ is a vector of any local characteristics that can affect productivity, $\mathbf{x}$ is a vector of inputs, $\mathrm{c}(\mathbf{x})$ is the cost function, and ϵ is the establishment's idiosyncratic productivity shock, independent and identically distributed across establishments according to a cumulative distribution function $\Phi(\epsilon)$. For any $\mathbf{y}$, there is a critical level $\epsilon^*(\mathbf{y})$ such that $\pi(\mathbf{y},\epsilon) \geq 0$ if and only if $\epsilon \geq \epsilon^*(\mathbf{y})$. An establishment will be born when it is possible to earn nonnegative profits and the probability that an establishment is created is $\Phi(\epsilon^*(\mathbf{y}))$.

Any local characteristic that increases productivity also increases both the number of new establishments and employment in these new establishments. As in Rosenthal and Strange (2003), we assume that this characteristic can be partitioned into: (i) $\mathbf{y}_z$, which varies by micro-geographic area, for example, cells of the one square kilometer; and (ii) $\mathbf{y}_d$, which varies by slightly larger geographical areas, for example, at the district level.[14] This distinction allows us to assess how the agglomeration economies are attenuated with distance within the city, for example at the neighborhood level.

The agglomeration economies do not necessarily act homogeneously within a city or region, so some effects can be highly localized (Rosenthal and Strange, 2003; 2008). The effects of the three sources of local externalities highlighted by Marshall (1890) (knowledge spillovers, labor market pooling and input sharing) can be attenuated differently with distance. For example, the knowledge spillovers occur within a few kilometers of the current establishment and are, therefore, highly localized (Fu, 2007; Andersson *et al.*, 2009). On the other hand, labor market pooling linked to local labor supply may occur at the metropolitan level. And the third Marshallian force, the sharing of physical inputs, which is often associated with the transportation structure, may occur at larger spatial

[14]Districts are intra-municipal areas defined by local governments with their own legislation and can reflect socioeconomic characteristics such as population and type of local economic activity. According to Brazilian Instutute of Geography and Statistics (IBGE), in 2007 there were 10,009 districts. The urban areas are also the most disaggregated. For example, the municipality of São Paulo (core city of the São Paulo Metropolitan Region) is divided into 96 districts (see Figure B.2 in Appendix B.1).

scales, such as the regional level (Rosenthal and Strange, 2020). Thus, the term $\mathbf{y}_z$ will include a set of factors (detailed in the next section) that describe how economic activity is organized around each microgeographic area while the term $\mathbf{y}_d$ represents the factors that act in larger spatial extensions.

3.4 Empirical strategy

3.4.1 Model specification

Our objective is to evaluate the spatial extent of agglomeration economies on the quantity of birth of new establishments and the employment levels that they choose in a given geographical area. To do so, it is necessary to solve two important secondary issues: (i) choosing the appropriate empirical model for our problem; and (ii) determining the geographical areas of analysis. The first has an immediate solution given the characteristic of our problem, count data. Thus, we model our problem using count models, more specifically the Poisson model, widely used in the literature (see Arauzo-Carod *et al.*, 2010 for a survey) and directly related to the problem described in the previous section from the establishment's location choice problem using the random profit maximization approach (see Carlton, 1983). For the second point, our finely geocoded employer-employee data allow us to freely define spatial units of measurement. All Brazilian territory is divided into around 8.5 million cells exogenously determined with one square kilometer each (1 km $\times$ 1 km).[15] The definition of such small microgeographic areas helps us to deal with a common problem in studies about agglomeration economies, sorting. Technically, if we evaluate the birth of firms in period t and the local characteristics in $t-1$, there is no simultaneity (Jofre-Monseny *et al.*, 2014), but firms can rank the eligible locations one year earlier, i.e., spatial sorting. However, our exogenously defined set of cells is outside the firms' choice. For example, it is difficult to think that the choice of the city of a new establishment is random. Furthermore, within the city itself, the districts can still be chosen. On the other hand, the new firm does not choose the specific cell because this depends on the availability of land and land use.

Once these initial conditions are established, we assume that new establishments are

[15]Recently, with the availability of microgeographic data, other studies have used similar strategies (see, e.g., Larsson, 2014; Andersson *et al.*, 2014; 2019; Li *et al.*, 2020a). Obviously, most of these cells do not have any kind of economic activity. Technically, an obvious criterion to select our study's geographic area is that most cells are uninhabited areas, such as forests, lakes and rivers.

opened at locations chosen from among of the square kilometer (cell $z = 1, \cdots, Z$) of Brazilian territory. Since our spatial unit of measurement is homogeneous, additional concerns regarding differences in the sizes of the geographical units are not necessary.[16] To capture the spatial extent of the agglomeration economies, we construct five concentric rings: 0-1, 1-5, 5-10, 10-20, and 20-40 km from the centroid of each cell, to measure our agglomeration variables. These variables are measured as usually done in location choice studies (see, e.g., Figueiredo *et al.*, 2002; Jofre-Monseny *et al.*, 2014; Li *et al.*, 2020a, just to cite a few), i.e., the own-industry j employment, emp_{jrt}, and the employment in other industries, emp_{-jrt}, in ring r and period t. The first measure captures the local intra-sectoral externalities (localization economies) and is associated with proximity to existing employment in establishments belonging to the same industry as the new establishment. The spatial concentration of plants of a particular manufacturing sector operates as a pool of favorable conditions, providing local specialized labor, sharing of intermediate input markets and generating knowledge spillovers. The second measure captures more general inter-sectoral local externalities associated with concentration of general economic activity in a particular area.[17] This type of externality can be internalized by all plants in the same area. Thus, for each 3-digit industry, the following Poisson model is estimated:

$$\mathbf{E}(\mathrm{Y}_{jzt+2}) = \exp\left(\sum_r \beta_{jr}^{loc} \text{emp}_{jrt} + \sum_r \beta_{jr}^{urb} \text{emp}_{-jrt} + \mathbf{X}_{jzt}\tau + \gamma_d\right) \tag{3.1}$$

where Y_{jzt+2} is the number of new plants or the number of jobs in these new plants in industry j, cell z and period $t+2$,[18] $\mathbf{X}_{jzt}$ is a vector of control variables containing location determinants other than agglomeration economies, and γ_d is the district fixed effect. We use the Poisson pseudo-maximum likelihood (PPML) estimator with multiple high-dimensional fixed effects recently developed by Correia *et al.* (2020) to deal with the large number of district fixed effects.

The main concern in the estimation of equation 3.1 is to obtain unbiased and consistent estimates for the set of parameters of interest, β_{jr}^{loc}, where $r = 1, \cdots, 5$. By including the fixed district effect, we control for any observed and unobserved characteristics fixed in time and specific to the district, which minimizes possible biases of omitted variables.

[16] Rosenthal and Strange (2003), for example, used Zip code areas of the US to deflate both births and new-establishment employment.

[17] Since we use only employment in manufacturing, these variables measure part of the urbanization economies.

[18] We calculate our outcomes variables in the periods 2007-2008, 2009-2010, 2011-2012 and 2013-2014 and the explanatory variables in 2006, 2008, 2010 and 2012. This is a common strategy in studies like this (see, e.g., Rosenthal and Strange; 2003; Jofre-Monseny, 2009).

Furthermore, we also test the robustness of the estimates by including a comprehensive set at cell and municipality level control variables that may affect the new establishments' location choice. The next subsection presents more details of these control variables.

3.4.2 Control variables

To address the omitted variable bias concerns, we test the robustness of our estimates by including control variables for economic environment, previously existing transportation infrastructure, geographical characteristics, and local development policies around the place chosen by the new establishment.

We begin with the economic environment around a specific cell. There is abundant evidence that incumbent local industrial structures can influence the level of local entrepreneurship. In particular, the presence of many small establishments is associated with employment growth in start-ups (see, e.g., Chinitz, 1961; Glaeser and Kerr, 2009; Rosenthal and Strange, 2010; Ghani *et al.*, 2014). Thus, as in Rosenthal and Strange (2003) and Li *et al.* (2020a), we include proxies for local industry organization and industry diversity. To be more precise, we include two Herfindahl indices within 40 km of the cell's centroid z. The first index captures, for example, the competition effects around the z and is measured for each 3-digit industry by $\sum_j(\text{emp}_{ijzt}/\text{emp}_{jzt})^2$, where emp_{ijzt} is the employment level of plant i in industry j in the region within 40 km of z in period t, and emp_{jzt} is the employment level of industry j in the region within 40 km of z in period t. This variable controls for local industrial organization around z. The second index captures the local diversity of economic activities and is measured by $\sum_j(\text{emp}_{jzt}/\text{emp}_{zt})^2$, where $\text{emp}_{jzt}/\text{emp}_{zt}$ is the industry j's share of total employment within 40 km of the centroid of z in period t.

Also at the cell level, we include a set of time-invariant transport and geographic controls. There is a broad set of evidence that transportation infrastructure can influence the location choice and productivity of plants (see, e.g., Holl, 2004a; 2004b; 2016; Mayer and Trevien, 2017; Gibbons *et al.*, 2019), thus including the set of transportation controls eliminates the effects of the previously installed transportation infrastructure on our estimates. At the same time, the proximity of rivers or lakes can affect the choice of location of resource-intensive industries by reducing input transportation costs (Ellison and Glaeser, 1999; Ellison *et al.*, 2010; Rosenthal and Strange, 2001). Furthermore, by including these variables, we also control for zoning and planning restrictions, something relevant in an intra-urban context. The transportation controls include the distance

between each cell's centroid and the nearest airports, public ports, railways, federal highways, and state highways. The geographic control is the distance between each cell's centroid and the nearest river. We interact these time-invariant controls with time effects to capture differential trends across cells.

Local development policies can also play an important role in attracting new investments (e.g., Glaeser *et al.*, 2010b; Chatterji *et al.*, 2014), and thus affect the location choice of new enterprises. So, at municipal level, we include values of capital expenditures (investments) and housing and town planning expenses per hundred thousand inhabitants, as a proxy for the quality of previously existing urban infrastructure; municipal taxes per hundred thousand inhabitants to control for differences in tax costs between municipalities; tax incentive policies implemented previously by local governments; and exports and imports per hundred thousand inhabitants, as a proxy for the access of firms previously located in the municipality to the international market. We also include homicides per hundred thousand inhabitants, which provides an indicator of the efficiency of public security policies implemented by local governments; and traffic fatalities per hundred thousand inhabitants, which acts as an indicator of the quality of the municipal public transportation system. We present a complete description of these control variables as well as the source of these data in Appendix B.1.

3.4.3 Remaining heterogeneities and control function approach

As discussed earlier, we use district fixed effects and a comprehensive set of control variables to minimize the bias in the estimation of localization economy effects. Even so, we cannot guarantee that the problem of omitted variables is completely solved. Various factors can make the estimates of parameters associated with these variables biased upward or downward. For example, one potential source of endogeneity is the factors in $\mathsf{A}(\mathbf{y})$, in particular, any local unobserved characteristics that can affect productivity can cause higher existing business concentrations, while at the same time, also attract a higher number of new establishments (Combes *et al.*, 2008; Li *et al.*, 2020a).

To address these potential concerns, we complement the analysis with a shift-share instrumental variable that exploits the changes in national employment growth specific to the industry (a "shift") to generate exogenous variation at the concentric ring level, in a control function approach.[19] The measure consists of the growth in employment

[19]This is an approach to estimate nonlinear models with endogenous explanatory variables (Terza *et al.*, 2008; Wooldridge, 2014). To illustrate the application of this approach to this study, we follow Navarro (2008) and Cameron and Trivedi (2013).

that would have occurred had each industry in a concentric ring grown at its national rate of growth (Bartik, 1991). More specifically, we use 33 industries at the 4-digit level to calculate the instrument for employment growth of eight 3-digt industries colocalized at short distances (mentioned in subsection 3.2.2) in each concentric ring r at time t. Formally:

$$\text{IV}_{jrt} = \sum_{k} \sum_{c} \omega_{rc} \text{emp}_{kc1995} \ln \left(\frac{\text{emp}_{kt} - \text{emp}_{kRt}}{\text{emp}_{k1995} - \text{emp}_{kR1995}} \right), \text{ with } \omega_{rc} = \frac{\mathsf{A}_{r \cap c}}{\mathsf{A}_{c}} \tag{3.2}$$

where $\mathsf{A}_{r \cap c}$ is the intersection area between concentric ring r and municipality c; A_{c} is total area of the municipality; emp_{kc1995} is the employment in industry k at the 4-digit level belonging to industry j at the 3-digit level in municipality c in reference year;[20] emp_{kt} is the national employment in industry k and year $t = (1999, \cdots, 2003)$; emp_{kRt} is the employment in industry k in area $R = \sum_{r} \mathsf{A}_{r}$ and year t. In other words, like Moretti and Thulin (2013), we are discounting from national employment in industry k the sum of employment in the five concentric rings in industry k.

This instrument isolates the variation that comes from nationwide changes at the 4-digit industry level k and uses the sum of the industrial mix components to calculate the variation of 3-digit industry j. To understand the logic, consider as example two concentric rings with the same size and the same share of manufacturing jobs in 1995, but a different industry mix within 3-digit manufacturing. If employment in a given industry increases nationally (where we remove from nationwide changes the local changes within a radius of 40 km), the concentric ring where industry employs a larger share of the labor force experiences a positive shock to the labor demand in the manufacturing sector. On the other hand, if employment in a given industry decreases, the concentric ring experiences a negative shock to the labor demand in the manufacturing sector (Moretti, 2010).

Using this instrument to construct a control function, the coefficients can be estimated in two steps. In the first step, our proxies for localization economies in each concentric ring, emp_{jrt}, is regressed using ordinary least squares (OLS) estimation on observed characteristics, presented in the previous subsection, and the instruments. In the second step, the count model is estimated with the residuals of the first step entering as a control for the unobserved confounder bias. This procedure is also known in the literature as

[20] We defined the reference year as the first year (1995) for which the National Classification of Economic Activities - CNAE is available. Thus, we used the first version of the CNAE (or CNAE 1.0), which contains 564 four-digit groups of industries, of which 268 are manufacturing industries. We merged the old CNAE version with the new version CNAE 2.0 (available from 2006) from the correspondence tables provided by IBGE, available at <https://cnae.ibge.gov.br/>.

two-stage residual inclusion (2SRI) and is more appropriate to deal with endogenous explanatory variables in nonlinear models than the extension of the popular linear two-stage least squares estimator for nonlinear models (Terza *et al.*, 2008; Terza, 2017). While this approach is widely employed in health econometric research (see, e.g., Stuart *et al.*, 2009; Lazuka, 2018; Ghanbariamin and Chung, 2020), its application in econometric studies of regional and urban economics is rare, despite the widespread use of nonlinear models with potentially endogenous explanatory variables, especially in studies of locational choice.

To illustrate this procedure, consider the following version of the count model presented in equation 3.1:

$$\mathbf{E}(\mathrm{Y}_{jzt+2}) = \exp\left(\sum_r \beta_{jr}^{loc}\mathrm{emp}_{jrt} + \sum_r \beta_{jr}^{urb}\mathrm{emp}_{-jrt} + \mathbf{X}_{jzt}\tau + \mathbf{u}_{jzt}\right) \tag{3.3}$$

where $\mathbf{u}_{jzt}$ are the effects not captured by the control variables included. Thus, the first step equations are given by:

$$\mathrm{emp}_{jrt} = \sum_r \delta_{jr}\mathrm{IV}_{jrt} + \sum_r \theta_{jr}^{urb}\mathrm{emp}_{-jrt} + \mathbf{X}_{jzt}\lambda + \mathbf{w}_{jrt},\ r = 1, \cdots, 5. \tag{3.4}$$

where IV_{jrt} is the shift-share instrumental variable, and $\mathbf{w}_{jrt}$ are omitted factors that can influence local industry concentration. If the terms $\mathbf{u}_{jzt}$ and $\mathbf{w}_{jrt}$ are correlated for any of the reasons mentioned above, then emp_{jrt} and $\mathbf{u}_{jzt}$ are correlated, so the Poisson regression of $\mathbf{E}(\mathrm{Y}_{jzt+2})$ on emp_{jrt} and the other covariables yields inconsistent parameter estimates (Cameron and Trivedi, 2013). However, we can obtain consistent estimates if $\mathbf{u}_{jzt} = \rho_j \mathbf{w}_{jrt} + \varepsilon_{jzt}$ with ε_{jzt} independent of $\mathbf{w}_{jrt}$, and estimating Poisson regression by substituting the term $\mathbf{u}_{jzt}$ in equation 3.3 by $\mathbf{w}_{jrt}$ estimated from the first step by OLS (Wooldridge, 1997; 2010). That is, the second step equation is:

$$\mathbf{E}(\mathrm{Y}_{jzt+2}) = \exp\left(\sum_r \beta_{jr}^{loc}\mathrm{emp}_{jrt} + \sum_r \beta_{jr}^{urb}\mathrm{emp}_{-jrt} + \mathbf{X}_{jzt}\tau + \rho_j \hat{\mathbf{w}}_{jrt}\right) \tag{3.5}$$

Additional concerns are related to the estimates obtained in the second step. We use $\hat{\mathbf{w}}_{jrt}$ as opposed to $\mathbf{w}_{jrt}$, i.e., a generated regressor, so we need to adjust the standard error estimates in the second step to take this extra source of variation into account (Petrin and Train, 2010; Cameron and Trivedi, 2013). We implemented bootstrapping to adjust the standard errors of the second step.[21] The estimated coefficient for $\hat{\mathbf{w}}_{jrt}$ provides the direction of unobserved confounder bias, and its statistical significance will

[21] We performed 400 bootstrap replications following the examples in Cameron and Trivedi (2013).

indicate whether the variable emp_{jrt} is indeed endogenous (Wooldridge, 1997).

3.5 Results

3.5.1 Baseline results

We showed in subsection 3.2.2 that the new establishments in some industries are colocalized at short distances. To be more precise, we showed that the entrants engaged in *manufacture of pharmaceutical products* - CNAE 212, *manufacture of other food products* - CNAE 109, *fruit & vegetable canning* - CNAE 103, *manufacture of starch products* - CNAE 106, *manufacture of furniture* - CNAE 310, *prepress and graphic finishing services* - CNAE 182, *manufacture of wood products* - CNAE 162, and *finishing of textile articles* - CNAE 134 are colocalized less than 70 km. Now, we evaluate if entrants' location choice in these industries can be affected by agglomeration economies. In particular, we evaluate whether the probability of an entrant being located in a specific cell depends on proximity to existing establishments in the same industry. Tables 3.2 and 3.3 present estimates of equation 3.1 when our outcome variables are the number of new establishments per cell and new-establishment employment per cell, respectively.

For example, consider initially the first estimated coefficient for the *prepress and graphic finishing services* - CNAE 182 (column 6 of Table 3.2), for which the localization effects are among the most pronounced. Adding 100 prepress workers up to 1 km would generate, on average, an increase of 37.6% in the expected number of births and 35.7% in the expected number of employees (as can be seen in the top panel of Tables 3.2 and 3.3). Adding 100 additional employees to the 1-5 km ring would result, on average, an increase of 5% in the expected number of births. On the other hand, when the outcome is new-establishment employment, the coefficient is not significant. For distances larger than 10 km, the coefficients are not significant, except for the 20-40 km ring in Table 3.3, which is negative, suggesting some kind of competition at greater distances. Note that these estimates are free of any bias caused by omitted variables that are time-invariant at the district level and any specific tendency related to previously existing transportation infrastructure, geographic characteristics, and local development policies.

We also provide the results of the models without district fixed effects and/or without control variables for the eight industries and note the dispersion in our data. As can be seen in Table B.3 in Appendix B.1, which provides the descriptive statistics by industry, in general there is overdispersion in the data for new-establishment employment, so the

negative binomial model is more appropriate. When the district fixed effects and/or control variables or urbanization variables are omitted from the Poisson or negative binomial models, there is, of course, a reduction in the magnitude of the estimated coefficients, and in most cases only the coefficients associated with the smaller rings remain positive and strongly significant for both firm birth and new-establishment employment. However, the pattern of attenuation with distance remains, as can be seen in Appendix B.3.

Looking in particular at entrants engaged in *manufacture of pharmaceutical products* - CNAE 212 (column 1 of Tables 3.2 and 3.3), the results also indicate that the probability of an entrant choosing a cell, but not the new-establishment employment, is higher when there are already other establishments in the same industry up to 1 km from location of the new establishment. In contrast, existing establishments located at larger distances, for example, 5, 10, 20 and 40 km, do not affect the probability of an entrant's choice. This pattern confirms our initial impressions (as shown in Figures 3.1 (a) and 3.2 (b)) that localization economies can determine the location choice of new establishments in this industry. More than that, there is a pattern of spatial decay of the effects generated by localization externalities consistent with the idea of gains generated by spillovers at short distances. In addition, estimates for the more general effects, associated with urbanization externalities, are also highly concentrated, much stronger up to 1 km, five times smaller up to 5 km and not significant thereafter. This indicates that the effects generated by proximity to existing establishments in other industries are also attenuated with distance.

Similarly, column 2 of Tables 3.2 and 3.3 report the results for the entrants engaged in *manufacture of other food products* - CNAE 109. In general, the results point in the same direction, i.e., there is a pattern of spatial decay in both localization and urbanization effects, but unlike what is observed for pharmaceutical products, in the food industry the localization externalities extend to large distances, precisely up to 10 km for firm birth outcome and 20 km for new-establishment employment, consistent with the pattern observed previously in the Figures 3.1 (b) and 3.2 (d). We know (from the chapter 2 of this dissertation) that this is a low-tech industry localized (at large distances) relative to manufacturing activity as a whole, a pattern consistent with the evidence found here indicating that this large spatial extension of localization effects is associated with sharing the local labor market. Therefore, although our model identifies agglomeration effects based on within-district variation of the data, our results are broadly consistent with previous works that was based on between-city variation in the data, i.e., also providing evidence of Marshallian agglomeration forces that act at larger spatial scales as labor market pooling (Rosenthal and Strange, 2003; 2020). In subsection 3.5.3 we provide a

Table 3.2 Spatial scope of localization and urbanization externalities - plant birth. Poisson regression

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pharma. Products CNAE 212	Food Products CNAE 109	Fruit & vegetable Canning CNAE 103	Starch Products CNAE 106	Furniture CNAE 310	Prepress Services CNAE 182	Wood Products CNAE 162	Fin. of textile Articles CNAE 134
Localization Effects								
0 to 1 km	9.83e-04**	4.70e-04***	1.55e-03**	1.94e-03***	6.54e-04***	3.76e-03***	2.20e-03***	1.91e-03***
	(3.90e-04)	(5.22e-05)	(6.53e-04)	(4.13e-04)	(7.29e-05)	(4.34e-04)	(2.83e-04)	(5.25e-04)
1 to 5 km	7.62e-05	1.79e-04***	5.64e-04	-2.32e-05	1.10e-04**	5.05e-04**	9.64e-05	3.05e-04
	(3.29e-04)	(3.80e-05)	(5.18e-04)	(3.75e-04)	(5.00e-05)	(2.00e-04)	(2.11e-04)	(2.11e-04)
5 to 10 km	1.18e-04	8.64e-05***	4.46e-05	-5.99e-05	-6.20e-06	-2.32e-04	1.13e-04	-4.25e-04**
	(2.09e-04)	(3.32e-05)	(4.84e-04)	(3.16e-04)	(5.03e-05)	(1.78e-04)	(1.78e-04)	(1.72e-04)
10 to 20 km	1.32e-05	4.47e-05	1.69e-04	-8.02e-05	-5.95e-05	8.49e-05	-8.96e-06	-1.40e-06
	(1.78e-04)	(2.96e-05)	(4.09e-04)	(2.64e-04)	(3.73e-05)	(1.51e-04)	(1.52e-04)	(1.11e-04)
20 to 40 km	1.04e-04	-1.10e-05	4.25e-04	1.82e-04	3.02e-06	-1.99e-05	-1.18e-04	-1.89e-04**
	(1.87e-04)	(2.64e-05)	(3.31e-04)	(2.34e-04)	(2.56e-05)	(1.70e-04)	(1.00e-04)	(9.47e-05)
Urbanization Effects								
0 to 1 km	1.95e-04***	1.32e-04***	1.52e-04***	2.02e-04***	1.52e-04***	7.04e-05***	1.77e-04***	1.84e-04***
	(4.21e-05)	(8.47e-06)	(3.38e-05)	(2.33e-05)	(8.56e-06)	(1.37e-05)	(1.41e-05)	(1.64e-05)
1 to 5 km	3.72e-05**	1.89e-05***	1.12e-05	1.89e-05**	2.12e-05***	2.33e-05***	1.01e-05*	2.44e-05***
	(1.83e-05)	(3.13e-06)	(9.87e-06)	(8.84e-06)	(2.89e-06)	(5.06e-06)	(5.35e-06)	(5.87e-06)
5 to 10 km	-2.73e-06	-6.44e-06***	2.55e-06	-3.20e-06	4.12e-06*	-3.53e-06	-2.35e-06	-7.39e-09
	(1.19e-05)	(2.19e-06)	(6.57e-06)	(6.88e-06)	(2.18e-06)	(3.78e-06)	(4.24e-06)	(4.10e-06)
10 to 20 km	-3.36e-06	-4.27e-06**	-4.13e-06	8.23e-06*	7.38e-07	-2.44e-06	1.24e-07	-3.15e-06
	(9.24e-06)	(1.72e-06)	(5.56e-06)	(4.32e-06)	(1.44e-06)	(3.22e-06)	(2.57e-06)	(2.49e-06)
20 to 40 km	1.13e-06	-2.06e-06	-4.46e-06	3.55e-06	2.67e-06**	-5.06e-06*	-1.23e-06	2.62e-06
	(4.90e-06)	(1.39e-06)	(4.34e-06)	(3.94e-06)	(1.16e-06)	(2.93e-06)	(2.29e-06)	(2.14e-06)
Average Change in Localization Effect per km								
0.5 to 3 km	-3.63e-04	-1.16e-04	-3.93e-04	-7.87e-04	-2.18e-04	-1.30e-03	-8.40e-04	-6.43e-04
3 to 7.5 km	9.32e-06	-2.06e-05	-1.15e-04	-8.14e-06	-2.58e-05	-1.64e-04	3.67e-06	-1.62e-04
7.5 to 15 km	-1.40e-05	-5.56e-06	1.65e-05	-2.72e-06	-7.11e-06	4.23e-05	-1.62e-05	5.65e-05
15 to 30 km	6.06e-06	-3.71e-06	1.71e-05	1.75e-05	4.17e-06	-6.98e-06	-7.30e-06	-1.25e-05
# of district FE	81	1874	381	663	1579	395	957	426
Pseudo R^2	0.1482	0.0743	0.1178	0.1245	0.1011	0.1257	0.1151	0.1606
Pseudo-LL	-523.8067	-26,834.97	-2,694.916	-4,507.093	-24,581.67	-4,507.879	-9,027.774	-5,030.079
Observations	19,151	118,376	47,495	61,350	111,668	56,812	82,976	58,389

Notes: This table reports the localization and urbanization effects when the dependent variable is the number of new establishments in each cell. Heteroscedasticity-robust standard errors are reported in parentheses. All columns include the diversification and competition control variables, transport and geographic controls, municipality level controls, and district fixed effects. The transportation controls include the distance to the nearest airport, public port, railway, federal highway, and state highway interacted with time effects. The geographic control is the distance to the nearest river interacted with time effects. The municipality level controls include proxies for insertion in international trade (exports and imports), municipal taxes, capital investments, housing and town planning expenses, homicides and traffic fatalities. Change per kilometer is computed by differencing the adjacent localization coefficients and dividing by the number of kilometers between the midpoints. Significance level: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: Prepared by the author based on estimates.

Table 3.3 Spatial scope of localization and urbanization externalities - new employment. Poisson regression

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pharma. Products CNAE 212	Food Products CNAE 109	Fruit & vegetable Canning CNAE 103	Starch Products CNAE 106	Furniture CNAE 310	Prepress Services CNAE 182	Wood Products CNAE 162	Fin. of textile Articles CNAE 134
				Localization Effects				
0 to 1 km	4.79e-04	1.54e-03***	4.19e-03***	3.12e-03***	7.93e-04***	3.58e-03**	5.78e-03***	2.08e-03*
	(9.98e-04)	(2.92e-04)	(1.23e-03)	(1.06e-03)	(2.02e-04)	(1.49e-03)	(8.63e-04)	(1.22e-03)
1 to 5 km	-9.69e-04	4.55e-04**	1.58e-03	-1.42e-03	-8.70e-05	8.32e-04	1.57e-03**	-4.25e-05
	(8.58e-04)	(2.17e-04)	(1.39e-03)	(9.60e-04)	(1.70e-04)	(5.77e-04)	(6.75e-04)	(4.21e-04)
5 to 10 km	-5.45e-05	3.49e-04*	1.78e-03	-1.62e-03	-3.94e-05	-8.41e-04	5.17e-04	-7.10e-04*
	(6.41e-04)	(1.84e-04)	(1.15e-03)	(1.06e-03)	(1.25e-04)	(6.00e-04)	(7.18e-04)	(3.97e-04)
10 to 20 km	-2.38e-04	2.44e-04*	1.38e-03	-1.54e-03*	-1.75e-05	-1.56e-04	-7.17e-04	-1.53e-04
	(8.07e-04)	(1.45e-04)	(1.08e-03)	(8.69e-04)	(9.06e-05)	(3.73e-04)	(6.04e-04)	(2.31e-04)
20 to 40 km	-2.49e-04	5.47e-05	-2.16e-04	-3.22e-04	6.20e-05	-1.13e-03*	-6.68e-04*	-5.48e-05
	(1.03e-03)	(1.22e-04)	(1.23e-03)	(6.82e-04)	(1.14e-04)	(6.04e-04)	(3.97e-04)	(1.66e-04)
				Urbanization Effects				
0 to 1 km	1.97e-04**	1.93e-04***	2.22e-04***	3.22e-04***	1.99e-04***	4.70e-05**	1.91e-04***	1.94e-04***
	(9.28e-05)	(3.19e-05)	(6.46e-05)	(6.62e-05)	(1.85e-05)	(2.18e-05)	(2.81e-05)	(5.84e-05)
1 to 5 km	4.48e-05	9.68e-06	2.28e-05	7.64e-05***	2.12e-05**	5.94e-06	-4.33e-06	4.24e-05***
	(5.08e-05)	(1.10e-05)	(3.70e-05)	(2.10e-05)	(9.79e-06)	(1.35e-05)	(1.62e-05)	(1.50e-05)
5 to 10 km	-2.38e-05	-1.84e-05*	-4.85e-06	2.46e-05**	-5.39e-07	-8.57e-07	2.25e-05	1.23e-05
	(4.06e-05)	(1.07e-05)	(1.59e-05)	(1.23e-05)	(6.73e-06)	(1.04e-05)	(2.03e-05)	(9.65e-06)
10 to 20 km	2.21e-05	-1.38e-05	-3.67e-05***	-7.51e-07	-1.33e-06	-3.91e-06	1.84e-05**	-1.99e-05***
	(2.20e-05)	(9.76e-06)	(1.20e-05)	(1.52e-05)	(3.86e-06)	(9.67e-06)	(7.83e-06)	(4.89e-06)
20 to 40 km	9.71e-06	-3.90e-06	-3.76e-05***	1.82e-05	3.70e-07	-2.08e-06	2.34e-05**	1.58e-06
	(1.84e-05)	(7.69e-06)	(1.14e-05)	(1.17e-05)	(4.27e-06)	(6.98e-06)	(9.14e-06)	(5.64e-06)
				Average Change in Localization Effect per km				
0.5 to 3 km	-5.79e-04	-4.34e-04	-1.04e-03	-1.82e-03	-3.52e-04	-1.10e-03	-1.68e-03	-8.50e-04
3 to 7.5 km	2.03e-04	-2.35e-05	4.46e-05	-4.34e-05	1.06e-05	-3.72e-04	-2.33e-04	-1.48e-04
7.5 to 15 km	-2.45e-05	-1.40e-05	-5.36e-05	1.02e-05	2.93e-06	9.14e-05	-1.65e-04	7.43e-05
15 to 30 km	-7.47e-07	-1.26e-05	-1.06e-04	8.12e-05	5.30e-06	-6.52e-05	3.25e-06	6.54e-06
# of district FE	41	1374	214	385	1,175	245	682	308
Pseudo R^2	0.5028	0.2535	0.5626	0.3411	0.2558	0.2908	0.4039	0.3095
Pseudo-LL	-24,369.76	-113,055.7	-14,575.2	-32,034.53	-71,729.64	-12,520.69	-37,538.29	-16,137.08
Observations	11,089	105,114	33,052	42,262	98,583	42,243	66,745	46,925

Notes: This table reports the localization and urbanization effects when the dependent variable is the number of new employments in each cell. Heteroscedasticity-robust standard errors are reported in parentheses. All columns include the diversification and competition control variables, transport and geographic controls, municipality level controls, and district fixed effects. The transportation controls include the distance to the nearest airport, public port, railway, federal highway, and state highway interacted with time effects. The geographic control is the distance to the nearest river interacted with time effects. The municipality level controls include proxies for insertion in international trade (exports and imports), municipal taxes, capital investments, housing and town planning expenses, homicides and traffic fatalities. Change per kilometer is computed by differencing the adjacent localization coefficients and dividing by the number of kilometers between the midpoints. Significance level: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: Prepared by the author based on estimates.

more detailed comparison of our results with those obtained for other countries. Note also that the effects of urbanization are attenuated up to 5 km, which suggests that the presence of existing establishments in other industries at short distances around the cell can act as an attraction force, increasing the probability of choosing the specific cell.

Estimates in columns 3-8 of both Tables 3.2 and 3.3 for all other industries previously classified as colocalized at short distances also indicate that the effects of both proximity to establishments in the same industry and to establishments in other industries attenuate rapidly with distance. Observe also that in all cases the localization effects are more important than urbanization effects. In particular, for the first concentric ring of employment, the coefficient of the localization employment variable is, on average, 13.6 times larger than the coefficient of the corresponding urbanization employment variable when the outcome is the birth of firms and 20 times larger when the outcome is the new-establishment employment. For the second concentric ring, the difference is, on average, 12.1 times larger when the outcome is the birth of firms and not significant when the outcome is the new-establishment employment. As explained by Rosenthal and Strange (2003), this provides a clear distinction between urbanization and localization economies, since it is expected that the gains from information spillovers and the ability to share both intermediate inputs and specialized labor diminish monotonically with increasing distance, while the urbanization effects can be of any sign because of tradeoff between the benefits and congestion costs of locating near densely developed areas. This evidence also points in the same direction as the results found by Henderson (1986) for Brazil and US, Nakamura (1985) for Japan with aggregate data, and more recently, based on within-city variation, by Li *et al.* (2020a) for China, but is more robust since we control for several heterogeneities not included in the previous studies.

In summary, our key geographical results are that for most industries, the localization economies attenuate rapidly in the first few kilometers. Another way to observe this pattern, as in Rosenthal and Strange (2003), is presented at the bottom of the tables, i.e., change per kilometer (CPkm) in the localization effects for each industry, measured by the difference in coefficients between each of the adjacent pairs of concentric rings divided by the number of kilometers between the midpoints of the two rings. For births (Table 3.2), the averaging across all eight industries of the ratio of CPkm values in the 0.5 km to 3 km range relative to the 3 km to 7.5 km range, $\text{CPkm}_{(0.5 \text{ to } 3)}/\text{CPkm}_{(3 \text{ to } 7.5)}$, is 9.5, while the ratio $\text{CPkm}_{(3 \text{ to } 7.5)}/\text{CPkm}_{(7.5 \text{ to } 15)}$ is -6.93, and the ratio $\text{CPkm}_{(7.5 \text{ to } 15)}/\text{CPkm}_{(15 \text{ to } 30)}$ is 4.88. When looking at the new-establishment employment (Table 3.3), the analogous values are 13.97, 7.22, and 0.88 respectively.

3.5.2 Control function results

In this subsection we report the results of the estimates in two steps as a robustness test, i.e., when we include in addition to the comprehensive set of control variables presented above, the fitted residual from the first-stage regression. In Panel A (top) of Table 3.4, we present the estimated coefficients in the second step when our outcome is the count of new establishments in each cell. In Panel B (bottom), we present the results when our outcome is the level of new-establishment employment by cell.

The endogeneity of emp_{jrt} $(r = 1, \cdots, 5)$ can be tested based on the coefficient of the first-stage residual reported in Table B.15 of Appendix B.3. For most industries we reject the null hypothesis of exogeneity of emp_{j1t}, i.e., the exogeneity of own-industry employment in the first concentric ring. For the concentric rings farthest from the cell's centroid, in most cases we fail to reject the null hypothesis of exogeneity of emp_{jrt}.

The coefficient of own-industry employment in the first concentric ring (0-1 km) is now, on average, 4.7 and 11.7 times as large as under the exogeneity assumption when the outcome is the birth of firms and new-establishment employment, respectively. For the second concentric ring (1-5 km), the statistically significant coefficients in both cases (Tables 3.2 and 3.4) are now, on average, 4.8 times large for the outcome firm birth, while there is no major change when the outcome is employment in new establishments. For comparison, consider again the *prepress and graphic finishing services* - CNAE 182. Adding 100 prepress workers up to 1 km would generate, on average, an increase of 60% in the expected number of births as can be seen in the column 6 of Panel A of Table 3.4, while the equivalent coefficient when the outcome is the new-establishment employment is not significant (Panel B). The coefficients of own-industry employment in the second concentric ring are not significant.

Another interesting example, for which the difference is more pronounced, is the *manufacture of pharmaceutical products* - CNAE 212. Under the exogeneity assumption, adding 100 pharmaceutical products workers up to 1 km would generate, on average, an increase of 9.8% in the expected number of births, but is not significant for employment at new establishments (as reported in the column 1 of Tables 3.2 and 3.3), while in the two-stage estimates, adding the same number of workers up to 1 km would generate, on average, an increase of 90% in the expected number of births and 280% in the expected number of employees (column 1 of Panels A and B of Table 3.4).

While these comparisons between the coefficients estimated under the exogeneity assumption and the two-stage estimated coefficients provide some indications about the

Table 3.4 Second stage of two-step estimates of localization effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pharma. Products CNAE 212	Food Products CNAE 109	Fruit & vegetable Canning CNAE 103	Starch Products CNAE 106	Furniture CNAE 310	Prepress Services CNAE 182	Wood Products CNAE 162	Fin. of textile Artifacts CNAE 134
Panel A: The dependent variable is births of new establishments								
0 to 1 km	9.02e-03**	1.14e-03***	6.40e-03**	5.33e-03***	3.86e-03***	6.02e-03**	7.60e-03***	1.63e-02***
	(3.50e-03)	(2.96e-04)	(2.57e-03)	(1.18e-03)	(5.14e-04)	(3.01e-03)	(8.31e-04)	(5.16e-03)
1 to 5 km	-4.53e-04	2.90e-04***	2.31e-04	4.88e-04	-5.99e-05	8.65e-05	5.45e-04***	2.16e-03***
	(3.91e-04)	(5.13e-05)	(6.59e-04)	(2.99e-04)	(6.49e-05)	(3.92e-04)	(1.66e-04)	(5.23e-04)
5 to 10 km	1.76e-04	-1.31e-05	-1.48e-04	-6.02e-05	2.40e-04***	3.55e-05	5.14e-04**	-6.02e-04*
	(1.77e-04)	(3.07e-05)	(4.53e-04)	(2.69e-04)	(6.07e-05)	(2.40e-04)	(2.07e-04)	(3.42e-04)
10 to 20 km	7.10e-05	1.34e-05	5.05e-04***	1.64e-04	7.59e-06	1.92e-04*	-1.72e-04	8.59e-04***
	(8.98e-05)	(1.48e-05)	(1.95e-04)	(1.20e-04)	(2.54e-05)	(1.12e-04)	(1.06e-04)	(9.71e-05)
20 to 40 km	-3.94e-05	-4.48e-05***	2.18e-04	1.31e-04	6.05e-05***	-4.16e-04***	1.89e-04***	-2.57e-05
	(9.56e-05)	(1.16e-05)	(1.38e-04)	(1.01e-04)	(7.57e-06)	(9.75e-05)	(4.65e-05)	(6.00e-05)
Panel B: The dependent variable is new-establishment employment								
0 to 1 km	2.80e-02*	2.81e-03***	2.20e-02***	4.55e-03*	6.19e-03***	6.40e-03	1.01e-02***	3.27e-02***
	(1.59e-02)	(9.23e-04)	(7.48e-03)	(2.51e-03)	(9.74e-04)	(1.38e-02)	(2.56e-03)	(9.07e-03)
1 to 5 km	-1.90e-03	4.78e-04**	-3.84e-03	2.32e-03***	-3.05e-04**	8.67e-04	-6.92e-04	3.26e-03***
	(1.77e-03)	(2.24e-04)	(2.79e-03)	(7.92e-04)	(1.53e-04)	(1.44e-03)	(6.84e-04)	(1.03e-03)
5 to 10 km	7.55e-04	-1.75e-04	3.70e-04	-2.51e-04	5.44e-04***	8.64e-04	1.84e-03*	-1.77e-03**
	(7.02e-04)	(2.30e-04)	(2.04e-03)	(5.15e-04)	(2.03e-04)	(1.27e-03)	(9.74e-04)	(7.23e-04)
10 to 20 km	4.74e-04	2.15e-04	1.47e-03**	7.71e-04**	-7.68e-06	3.22e-04	-5.15e-04	1.12e-03***
	(3.57e-04)	(1.34e-04)	(6.83e-04)	(3.06e-04)	(8.36e-05)	(4.13e-04)	(3.65e-04)	(2.87e-04)
20 to 40 km	3.51e-04*	-1.47e-04	-1.88e-04	2.85e-04	7.02e-05***	-9.20e-04**	4.52e-04***	-9.03e-05
	(2.11e-04)	(1.04e-04)	(6.25e-04)	(1.99e-04)	(2.70e-05)	(4.49e-04)	(1.30e-04)	(1.17e-04)

Notes: This table reports the localization effects when we include a control function to address endogeneity concerns (equation 3.5). All models are estimated with 139,527 observations. All columns report the results from Poisson regressions where the dependent variable is the births of new establishments (Panel A) and the new-establishment employment (Panel B), and the variable of interest is the number of workers in the same industry in each concentric ring. All columns include the urbanization, diversification and competition variables, along with the transport and geographic controls, as presented in the previous section. Standard errors based on 400 bootstrap replications are reported in parentheses. Significance level: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: Prepared by the author based on estimates.

sign and magnitude of the bias, the main result of this subsection is that the pattern of attenuation with distance can also be clearly observed in the two-stage results. Although in both cases the second stage results show more specific patterns. One is in column 8 for *finishing of textile articles* - CNAE 134, where a negative effect appears at 10 km and then becomes positive at 20 km (Panels A and B); another is in column 5 for *manufacture of furniture* - CNAE 310, where a negative effect appears at 5 km and then becomes positive at 10 km (Panel B). This may represent some kind of competition generated by the emergence of new clusters. But note that in both cases the residuals of the first-stage equation are not significant and we do not reject the null hypothesis of exogeneity (see Table B.15). In summary, our main findings remain valid, indicating that our results are robust to possible biases caused by the potential endogeneity of our main explanatory variables.

Our results for Brazil are consistent with theoretical models of urban areas and previous empirical evidence for other countries. For example, Rosenthal and Strange (2003) found similar results for US (localization effects attenuated around 10 km). For developing countries, the evidence on the subject is scarcer. Indeed, there is only the recent study of Li *et al.* (2020a), who found that the localization effects in some industries are attenuated more rapidly with distance in China than in developed countries. As in that study, when we compare our general results with Rosenthal and Strange (2003)'s results, we find that localization effects are also attenuated more rapidly in Brazil. Note that in most cases for both births and new-establishment employment, the positive effects disappear after 5 km, which may suggest that localization effects are attenuated more rapidly because of the local urban infrastructure. We present a more detailed analysis of this phenomenon in the next subsection.

Also exploring the spatial scope of agglomeration economies, but using wages as an outcome, Rosenthal and Strange (2008) found that agglomeration economies attenuate rapidly with distance in the US. In particular, the human capital spillovers were found to be stronger up to 8 km from the individual's workplace. For other developed countries, also using wages as an outcome variable, Addario and Patacchini (2008) for the Italy found that urbanization externalities are attenuated up to 12 km, and more recently Håkansson and Isacsson (2019) found that urbanization externalities are attenuated around 25 km in Sweden. These studies, although addressing a different question than ours, also present evidence in the same direction, i.e., the agglomeration forces are heterogeneous within cities, so the use of geographically aggregated data does not allow exploring in detail the spatial scope of these externalities.

3.5.3 Other key industries and comparison with US and Chinese results

We have thus far presented the spatial scope of agglomeration economies in Brazil. In this subsection, we contrast our estimates with estimates from Rosenthal and Strange (2003) for the US and Li *et al.* (2020a) for China to reveal possible differences in attenuation patterns with distance between countries. Rosenthal and Strange (2003) estimated the determinants of firm birth for six industries (software - SIC 7371, 7372, 7373, and 7375, food products - SIC 20, apparel - SIC 23, printing and publishing - SIC 27, fabricated metals - SIC 34, and industrial and commercial machinery - SIC 35). Li *et al.* (2020a) estimated the determinants of firm birth for all Chinese manufacturing industries. Except for the software industry, we contrast the US and Chinese results for the remaining five industries with the results obtained for similar industries in Brazil. In the previous sections we have already presented the results for *manufacture of other food products* - CNAE 109 and *prepress and graphic finishing services* - CNAE 182, so to make comparison possible, we complement the analysis here by including *manufacture of wearing apparel* - CNAE 141, *manufacture of metal structures* - CNAE 251 and *manufacture of machinery* - CNAE 282.[22]

Before we begin a comparison by industry, we highlight some important general differences between the previous two papers and the present study. Unlike Rosenthal and Strange (2003) and Li *et al.* (2020a), we have panel data, which allows us to control for any observed and unobserved heterogeneities fixed in time in different areas within cities (districts). We use an exogenous microgeographic spatial partitioning structure that, as we discussed earlier, minimizes sorting bias. We also use a comprehensive set of control variables for previously existing transportation infrastructure and geographic features around the cell and local development policies.

Now, to get a detailed view of localization effects on the firm birth by industry, each panel in Figure 3.4 reports the results of the comparison for a specific industry (results from Tables 3.2 and B.7). As in Li *et al.* (2020a), to allow for easy comparison and interpretation, here we also define the vertical axis in each figure so that the magnitude of the spillover effects in the machinery industry within the first ring in the corresponding study is equal to one and all other spillover effects are measured relative to this value. The horizontal axis measures the spatial distance between firms in the same industry.

[22]The results of the estimation of equation 3.1 for these industries can be seen in Table B.7 in Appendix B.3. In Panel A (top panel) we present the results when the dependent variable is births of establishments, while in Panel B (bottom panel) we present the results when the dependent variable is new-establishment employment.

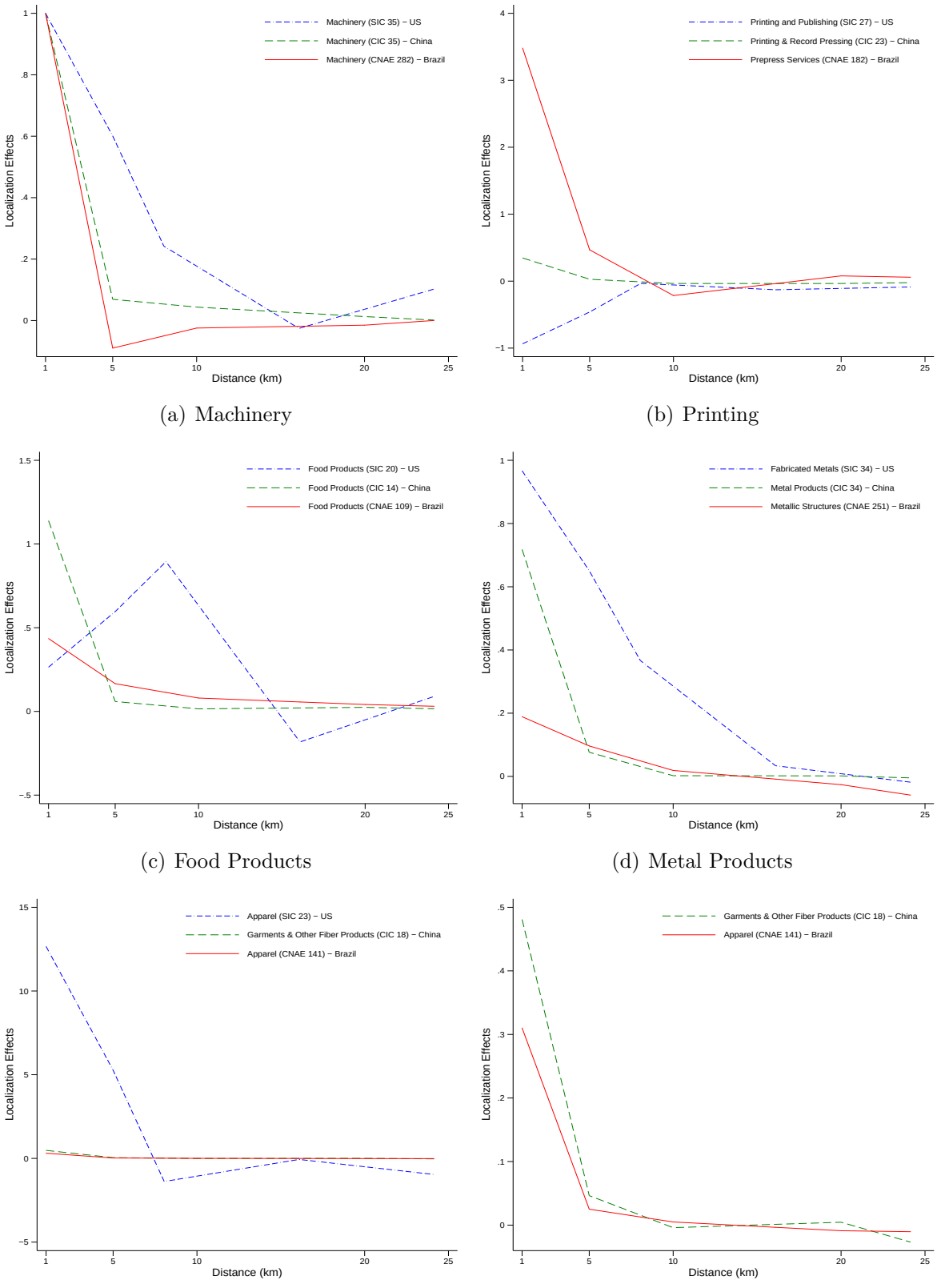
Since the scale and the measurement unit vary among the three studies,[23] we first convert the unit of measurement used by Rosenthal and Strange (2003) (miles) to kilometers, and then we obtain the intermediate values of the estimated coefficients by linear interpolation.

Some interesting patterns emerge from these comparisons. We start with the most general. After adjusting the different scales and distance units used, for 3 of the 5 industries analyzed (machinery, printing and food products), the attenuation of localization economies is faster in Brazil than in the US (as can be seen in Figure 3.4 (a-c)). For 2 industries (machinery and apparel), the attenuation patterns in Brazil are similar to those in China. The contrast can imply that the knowledge spillovers and/or labor pooling mechanism in the US, and to a lesser degree in China, acts over larger distances, so that firms farther from each other can still share knowledge and the specialized labor market. This evidence conforms very well with the pattern of high concentration of manufacturing industries in Brazil.

In the machinery industry, the attenuation in Brazil is similar to China, while in printing, the attenuation is faster in Brazil. For the food industry in Figure 3.4 (c), the attenuation is faster in Brazil relative to the US but not to China. This contrast with the US may be related to the different extension of the transportation infrastructure in the two countries. Similar arguments were also used by Li *et al.* (2020a) to explain the differences in attenuation patterns that are slower in the US than in China, and also apply very well to the characteristics of the costly transportation system in Brazil. A costly and underdeveloped transportation system that makes collaboration between more distant firms hard may restrict the effects of location economies to short distances. Relative to the results for China, the slower attenuation of localization effects in the food industry in Brazil may be associated with the supply of inputs spread throughout all regions of the country. For the metal industry, depicted in Figure 3.4 (d), the attenuation is slower in Brazil, both in relation to the US and China.

Another interesting pattern that contrasts with what has been discussed so far can be seen in Figure 3.4 (e) for the apparel industry. The attenuation of localization economies in this industry in the US is very fast when compared to the pattern observed in the corresponding industry in Brazil. For example, for this industry in the US, the localization effects disappear (are not significant) at distances greater than 1.6 km, while in Brazil

[23]For example, Rosenthal and Strange (2003) used four concentric rings set between 0-1 miles (0-1.6 km), 1-5 miles (1.6-8 km), 5-10 miles (8-16 km), and 10-15 miles (16-24 km). In turn, Li *et al.* (2020a) used five concentric rings set between 0-1 km, 1-5 km, 5-10 km, 10-20 km, and 20-30 km. For comparison purposes, we set the maximum distance on the horizontal axis equal to 24 km (distance of the largest ring used by Rosenthal and Strange (2003)).



(a) Machinery

(b) Printing

(c) Food Products

(d) Metal Products

(e) Apparel

(f) Apparel without US

Figure 3.4 Attenuation of localization economies for five selected manufacturing industries in comparison with Rosenthal and Strange (2003) and Li et al. (2020)

for both births and new-establishment employment these positive effects disappear at distances greater than 5 km (as can be seen in column 1 of Table B.7). This pattern is consistent with the argument that the apparel industry in the US could be more directly engaged in designing and advertising, which benefits more from knowledge spillovers because it requires more idea sharing and networking (Li *et al.*; 2020a). The apparel industry in Brazil, as in China, is more associated with manufacturing processing, which depend less on knowledge spillovers. In Figure 3.4 (f), we plot the estimated coefficients excluding results for the US, because they have different scales, to make it easier to compare the results for Brazil with those for China. The spatial decay of the effects is similar to that observed in China.

3.6 Concluding remarks

The objective of the article has been to examine the spatial scope of agglomeration economies in Brazil. In order to do so, we use a unique and rich microgeographic database for all Brazilian manufacturing industries and estimate the local determinants of the number of births per square kilometer and their associated employment levels as functions of the own-industry employment in different distance bands, controlling for the economic environmental characteristics around the site chosen by the new establishment. To better understand the geographical distribution of the new establishments, initially we address location and colocation patterns of new manufacturing plants using the nonparametric approach of Duranton and Overman (2005; 2008). After obtaining the pattern of geographic distribution of the new plants, we analyze the spatial scope of the agglomeration economies considering mainly the industries classified as colocalized at short distances.

Unlike previous studies on the same subject, such as Rosenthal and Strange (2003) and Li *et al.* (2020a), we use panel data, which allows us to control for time-fixed observed and unobserved heterogeneities at the district level and a comprehensive set of control variables for the economic environment, previously installed transportation infrastructure, geographical characteristics, and local development policies applied to the place chosen by the new establishment. We also use instrumental variables to address any remaining sources of heterogeneities in a control function approach.

From our initial nonparametric investigation, two main features emerge revealing details not available in the literature for developing countries, although they were found by Duranton and Overman (2008) for the UK. Among all manufacturing activities,

14.94% and 12.82% of entrants in 2007-2008 and 2013-2104 were localized while 77% in both periods did not have statistically different location patterns from those observed for existing establishments. The other feature is that, in contrast to what is observed for localization, there is a colocation pattern at short distances, which suggests that agglomeration economies are important. Furthermore, 42.53% and 26.92% of entrants were colocalized with existing establishments in the same periods.

In our parametric investigation, we find that in nearly all cases for both births and new-establishment employment, localization effects (own-industry employment at 3-digit level) are important, mainly up to 5 km from the birthplace. These results show that the infrastructure of cities in Brazil and the way manufacturing activities are geographically distributed benefit the emergence of new establishments in areas where there is existing geographical concentration. Moreover, both the births of new establishments and the level of employment they choose are higher in places that present high local specialization, which consequently affects the local choice of the new plant. This pattern conforms very well with the high geographic concentration of manufacturing observed in Brazil and with the high interregional mobility of workers. This evidence is also consistent with previous results for other countries indicating that localization effects attenuate rapidly over the first few kilometers. Our results are robust to the inclusion of a comprehensive set of controls, district fixed effects and also the inclusion of instrumental variables in a control function approach to deal with the potential endogeneity of our key explanatory variables.

As mentioned earlier, this paper provides evidence about a topic relevant to the formulation of economically efficient public policies focused on manufacturing entrepreneurship that considers the intrinsic forces of attraction in the market. In summary, although it is not in the scope of this study to determine the sources of agglomeration that cause the observed pattern, in general our evidence is in accordance with the spatial scope of action of the three Marshallian agglomeration forces, namely knowledge spillovers at short distances, labor market pooling and shared inputs at greater distances. Exploring these sources in detail is on the agenda for future studies.

Chapter 4

The spatial extent of human capital spillovers in a transition country: Evidence from Brazil

4.1 Introduction

It is well established in the literature that agglomeration economies are important to understand the distribution of economic activities in geographic space (see, e.g., Duranton and Puga, 2004; Rosenthal and Strange, 2004). The central question in the agglomeration literature is idea that spatial concentration enhances productivity. The action of these forces helps to determine the optimal size of cities (Chauvin *et al.*, 2017). Most of the existing evidence on the subject is based on assumptions that the effects of agglomeration economies are homogeneous in space, implicitly assumed to be "club goods" (see Rosenthal and Strange, 2003), which operate homogeneously within cities and intermediate regions or even at more aggregated levels (see Combes and Gobillon, 2015, for a recent survey). However, some externalities, especially those associated with the interaction between workers (face-to-face), such as knowledge spillovers, may not occur homogeneously in the city as a whole and are stronger at short distances (see, e.g., Rosenthal and Strange, 2008; Fu, 2007; Andersson *et al.*, 2009).

While understanding spatial decay of agglomeration economies in a micro-geographic context is politically relevant and essential to understand their nature,[1] there is little empirical evidence about this. Furthermore, this evidence is almost exclusively for developed countries. Using individual wages, Fu (2007) evaluated the spatial scope of different types of local externalities in the Boston Metropolitan Area and found that the effects of human capital attenuate sharply with distance. The contribution of Rosenthal and Strange (2008) revealed that the effects of agglomeration economies are localized and attenuate rapidly in the US. In particular, that the external returns to education are

[1] See, e.g., Rosenthal and Strange (2020) for a discussion of the different geographic scales of operation of micro-foundations of the agglomeration economies.

greater at short distances (around 8 km). Addario and Patacchini (2008) provided similar evidence for Italy from urbanization externalities that occur mainly up to 4 km and are attenuated up to 12 km. More recently, Håkansson and Isacsson (2019) indicated that the effects of urbanization are localized (attenuated up to 25 km) and asymmetric across percentile ranks in the wage distribution in Sweden.

In developing countries, aggregated geographic data provide evidence that the effects of agglomeration economies are greater compared with estimates for developed countries (Duranton, 2016a; Barufi *et al.*, 2016; Chauvin *et al.*, 2017; Combes *et al.*, 2013; 2020), despite the numerous costs and disadvantages associated with city size. The structure of cities is different in developing countries, where problems associated with the provision of public services such as transport, and consequently commuting time, in general are greater (Glaeser and Henderson, 2017; Thisse, 2018), which can substantially affect the geographic spread of agglomeration externalities. Little is known about the spatial scope of agglomeration economies in these environments. To be more precise, using micro-geographic data, only Li *et al.* (2020a) provides some evidence for China. The authors evaluated the effects of localization externalities after controlling for urbanization in different distance bands on the share of new firms. The main findings suggested that attenuation is very different among industries, but in general is faster in China than in the US (compared to Rosenthal and Strange (2003)'s results), indicating that agglomeration economies can be spatially attenuated more quickly because of the local urban structure.

This paper seeks to reduce part of this gap in the literature by analyzing the spatial extent of external returns to education in Brazil. For this, we employ finely geocoded employer-employee panel data in the period 2006-2014 from Brazil to assess if the effects of human capital spillovers on workers' productivity (proxied by their hourly wage) change with distance. More specifically, we evaluate the spatial extent of the externalities generated by the concentration of college-educated workers in different distance bands in relation to the individual's workplace.

Beyond these general differences in the economic environment between developed and developing countries, other characteristics of the Brazilian economy make the investigation of this phenomenon in the country particularly interesting. For example, relative to China, one immediate difference is the free interregional mobility of workers, which can strongly affect the spatial scope of agglomeration economies, in particular human capital spillovers, which are more spatially localized.[2] The spatial concentration of high-tech manufacturing

[2]An example of this type of restriction is China's *hukou* system, which controls population movement by restricting workers' social rights mostly to their birthplace (Combes and Gobillon, 2015).

is positively correlated with the share of college-educated workers (as we showed in the chapter 2), which may suggest that spillovers occur over short distances. Brazil is historically among the countries with the highest levels of income inequality in the world (Fishlow, 1972; Mendonça and Barros, 1995; Narita *et al.*, 2003), but at the beginning of the 21st century, there was a reduction of per capita household income inequality associated with the reduction of educational differentials (Barros *et al.*, 2007a; 2007b; Oliveira and Silveira Neto, 2013; 2016). By analyzing the period 2006-2014 we investigate if there was any change in spatial scope of external return to education consistent with this inequality reduction.

In addition to making a contribution to the scarce empirical literature, unlike previous studies, such as Rosenthal and Strange (2008), we implement a set of tools taking advantage of characteristics of our rich database and exploit exogenous education policy shocks at the national level to identify the causal effect of spillovers in different distance bands. We use panel data, which allows us to control besides the observed characteristics, the unobserved individual and plant heterogeneities, industry-year specific trends, and region fixed effects in wage regressions. Since our data are point data, our geographic units of analysis are squares measuring 1 km^2 and are defined exogenously from Brazil's geographical boundaries. This allows us to explore very small geographic contexts, for example, smaller than neighborhoods, in line with the literature suggesting that human capital spillovers are very local (Fu, 2007; Rosenthal and Strange, 2008). Our choice of units also allows us to minimize potential problems associated with the a priori definition of official administrative areas, such as heterogeneous sizes (Briant *et al.*, 2010). To deal with the potential endogeneity of the human capital variables, we combine exogenous shocks in Brazilian educational policies with the lagged demographic structure of each distance band in a shift-share instrumental variable approach. We also explore how our main results vary when our estimates are conditional to a diverse set of additional controls, such as worker-plant match fixed effects (or job-spell fixed effects), worker-city match fixed effects, mass of low-schooling workers, and the transportation infrastructure surrounding the individual's workplace.

The main results indicate that both the externalities generated by the concentration of general workers and college-educated workers are highly localized and much stronger in the first distance band (0-1 km). The positive effects of external return to education are generally attenuated up to 10 km. For example, in our main specification, adding 1,000 college-educated workers up to 1 km would increase the wages of workers on average by 6.78 percent. On the other hand, if the same number of workers are added to the 1-5

km or 5-10 km range, the wages of workers would increase, on average, by 1.95 and 1.27 percent, respectively. This evidence is robust to different specifications and shows that the speed of decay of agglomeration economies is higher in Brazil than observed in developed countries.

The remainder of the paper is structured as follows. In the next section we present the theoretical framework that provides the basis for our empirical approach. Section 4.3 describes the empirical strategy. Sections 4.4 and 4.5 present the results and final comments.

4.2 Theoretical framework

In this section we present a theoretical framework to address the spatial scope of external returns to education. We define external returns to education as the effect of an increase in the number of educated workers in a specific location and in neighboring locations on total wages minus the effect due to private returns to education (Acemoglu and Angrist, 2000; Moretti, 2004a). We adapt the theoretical structure proposed by Moretti (2004a) to identify the externalities generated by concentration of educated workers, assuming there are two types of workers, with low and high education, who are imperfect substitutes. But different from that author, we expand the model so that we can capture the spatial scope of human capital externalities. As mentioned by Rosenthal and Strange (2020), the human capital spillovers as envisioned by Marshall (1890) are likely to be highly local. The empirical evidence confirms this hypothesis (e.g., Fu, 2007; Rosenthal and Strange, 2008), and as highlighted by Charlot and Duranton (2004; 2006), although the improvement in information technology allows effective communication with distant partners, there is no evidence that this type of communication replaces face-to-face meetings, but instead is complementary.

In the model presented by Moretti (2004a), the human capital spillovers are treated as club goods that act homogeneously within cities. Our structure generalizes this hypothesis and allows these effects to be attenuated with geographic distance and therefore allows them to be different within the same city. The production function is given by:

$$\mathbf{Y}_{jz} = (\theta_{1jz}\mathbf{N}_{1jz})^{\alpha_1}(\theta_{2jz}\mathbf{N}_{2jz})^{\alpha_2}\mathbf{K}_{jz}^{1-\alpha_1-\alpha_2} \tag{4.1}$$

where $\mathbf{Y}_{jz}$ is the output of firm j located at z; $\mathbf{N}_{1jz}$ is the number of workers with high education in firm j located at z; $\mathbf{N}_{2jz}$ is the number of low-schooling workers; $\mathbf{K}_{jz}$ is

capital; and θ's are productivity shifters.

We allow for human capital spillovers by letting workers' productivity depend on the number of educated workers in neighboring locales in a continuous space,[3] as well as on their own human capital:

$$\log(\theta_{\ell jz}) = \phi_{\ell jz} + \gamma \sum_{\bar{z}} f(d(z,\bar{z}))\overline{\mathbf{N}}_{1\bar{z}} \qquad \ell = 1,2 \tag{4.2}$$

where $\phi_{\ell jz}$ is a group-specific effect that captures the direct effect of own human capital on productivity in a specific firm and place ($\phi_{1jz} > \phi_{2jz}$); $\overline{\mathbf{N}}_{1\bar{z}}$ is the number of workers with high education in all firms $k \neq j$ located at $\bar{z}$, so the term $\sum_{\bar{z}} f(d(z,\bar{z}))\overline{\mathbf{N}}_{1\bar{z}}$ captures the effects of Marshallian externalities resulting from the concentration of educated workers in other firms in the same and neighboring localities, weighted by a spatial decay function $f(d(z,\bar{z}))$ with $f(0) = 1$, $f'(d(z,\bar{z})) < 0$; and $d(z,\bar{z})$ is the distance between z and $\bar{z}$. As in Moretti (2004a), if there are positive spillovers, $\gamma > 0$.

Define R as the set of all locations. Now we can define the total of educated workers in R by $\mathbf{N}_{1\mathrm{R}}$ and the total of low-schooling workers by $\mathbf{N}_{2\mathrm{R}}$. Thus, the total number of highly educated workers can be decomposed into $\mathbf{N}_{1jz} + \sum_{\bar{z}} \overline{\mathbf{N}}_{1\bar{z}}$, where the second term is the number of highly educated workers in firms $k \neq j$ in the same (if $z = \bar{z}$) and neighbouring localities (if $z \neq \bar{z}$). If wages are equal to the marginal product of each type of worker and the spillover is external to individual firms in z but internal to the R (take $\overline{\mathbf{N}}_{1\bar{z}}$ as given), the logarithm of wages for highly and low-schooling workers is given by:

$$\begin{aligned} \log(w_{1jz}) = {} & \log(\alpha_1) + \alpha_1\phi_{1jz} + \alpha_2\phi_{2jz} + (\alpha_1+\alpha_2)\gamma \sum_{\bar{z}} f(d(z,\bar{z}))\overline{\mathbf{N}}_{1\bar{z}} \\ & + (\alpha_1 - 1)\log(\mathbf{N}_{1jz}) + \alpha_2 \log(\mathbf{N}_{2jz}) + (1-\alpha_1-\alpha_2)\log(\mathbf{K}_{jz}) \end{aligned} \tag{4.3}$$

$$\begin{aligned} \log(w_{2jz}) = {} & \log(\alpha_2) + \alpha_1\phi_{1jz} + \alpha_2\phi_{2jz} + (\alpha_1+\alpha_2)\gamma \sum_{\bar{z}} f(d(z,\bar{z}))\overline{\mathbf{N}}_{1\bar{z}} \\ & + \alpha_1 \log(\mathbf{N}_{1jz}) + (\alpha_2 - 1)\log(\mathbf{N}_{2jz}) + (1-\alpha_1-\alpha_2)\log(\mathbf{K}_{jz}) \end{aligned} \tag{4.4}$$

Therefore, the effect of $\mathbf{N}_{1\mathrm{R}}$ on workers' productivity is the sum of three effects: (i) that generated by the number of highly educated workers within firm j (neoclassical effect); (ii) that generated by the number of highly educated workers in other firms k in the same location z; and (iii) that generated by the number of highly educated workers in

[3] To approximate our theoretical framework of our empirical model in the next section, we consider a summation in the second term of equation 4.2, but the results remain valid when we consider infinitesimal variations (integral).

other firms k in neighboring locations $\bar{z}$. Formally:

$$\frac{\partial \log(w_{\ell jz})}{\partial \mathbf{N}_{1\mathrm{R}}} = \frac{\partial \log(w_{\ell jz})}{\partial \mathbf{N}_{1jz}} + \frac{\partial \log(w_{\ell jz})}{\partial \overline{\mathbf{N}}_{1\bar{z}}}\bigg|_{\bar{z}=z} + \sum_{\bar{z}\neq z} \frac{\partial \log(w_{\ell jz})}{\partial \overline{\mathbf{N}}_{1\bar{z}}} \qquad \ell = 1,2 \tag{4.5}$$

Which results in:

$$\frac{\partial \log(w_{1jz})}{\partial \mathbf{N}_{1\mathrm{R}}} = \frac{\alpha_1 - 1}{\mathbf{N}_{1jz}} + (\alpha_1+\alpha_2)\gamma + (\alpha_1+\alpha_2)\gamma \sum_{\bar{z}\neq z} f(d(z,\bar{z})) \quad \forall\, z,\bar{z} \tag{4.6}$$

$$\frac{\partial \log(w_{2jz})}{\partial \mathbf{N}_{1\mathrm{R}}} = \frac{\alpha_1}{\mathbf{N}_{1jz}} + (\alpha_1+\alpha_2)\gamma + (\alpha_1+\alpha_2)\gamma \sum_{\bar{z}\neq z} f(d(z,\bar{z})) \quad \forall\, z,\bar{z} \tag{4.7}$$

In words, both types of workers are affected by spillovers generated by proximity to highly educated workers. But note that as in Moretti (2004a), low-schooling workers, w_{2jz}, benefit for two reasons. First, an increase in the number of workers with high education raises low-schooling workers' productivity because of imperfect substitution ($\alpha_1/\mathbf{N}_{1kz} > 0$). Second, the spillover raises their productivity ($(\alpha_1+\alpha_2)\gamma+(\alpha_1+\alpha_2)\gamma\sum_{\bar{z}\neq z} f(d(z,\bar{z})) > 0$). For highly educated workers, w_{1jz}, the impact of an increase in the number of highly educated workers depends on two opposite effects. The first is the conventional supply effect, which makes the economy move along a demand curve ($(\alpha_1-1)/\mathbf{N}_{1jz} < 0$), and the second is the spillover effect, which raises productivity.

Unlike Moretti (2004a), in adapting the model proposed in this section, the spillover effects can be attenuated with the distance between z and $\bar{z}$. This can be observed by considering the behavior of $f(d(z,\bar{z}))$ and obtaining the derivative of the return to education with respect to distance between two localities ($z \neq \bar{z}$).

$$\frac{\partial^2 \log(w_{\ell jz})}{\partial \mathbf{N}_{1\mathrm{R}} \partial d(z,\bar{z})} = (\alpha_1+\alpha_2)\gamma \frac{\partial f(d(z,\bar{z}))}{\partial d(z,\bar{z})} < 0 \qquad \forall z \neq \bar{z}, \ell = 1,2 \tag{4.8}$$

It is interesting to note that when z index microgeographic areas of a continuous space, the adaptation of the theoretical structure satisfies to the objective for which it was proposed, because it allows spillover effects to be captured in very small areas, for example, smaller than neighborhoods. In the next section we describe our empirical strategy based on these conclusions about the attenuation of human capital spillovers.

4.3 Data and empirical strategy

4.3.1 Data and variables

Our main source of data is the Annual Report of Social Information (*Relação Anual de Informações Sociais*, or RAIS) available each two years in the period 2006-2014, i.e., a total of 5 years, which encompasses all formal workers in Brazil and is available from the Ministry of Labor. This dataset allows us to monitor workers and plants across years and provides detailed information at worker-level such as wages, gender, education, age, tenure, hiring data, number of hours worked, kind of contract, occupation, and identifier of the plant in which the worker is employed (National Register of Legal Entities - CNPJ number). At plant-level, detailed information is available about address, plant size (classification based on the number of workers employed), National Classification of Economic Activities (CNAE), which is compatible with the International Standard Industrial Classification of all Economic Activities (ISIC) revision 4, and opening and closing dates (if applicable). We use the address information available in RAIS data and the Google Maps base to obtain the geographic coordinates of each plant to construct a unique set of point data.[4] We also use other complementary data sources, such as the Census, National Household Survey (*Pesquisa Nacional por Amostra de Domicílios*, or PNAD),[5] data from geographic information systems (GIS), all provided by the Brazilian Institute of Geography and Statistics (IBGE), the Brazilian Soil Survey provided by Embrapa and the Brazilian Geological Survey provided by CPRM to calculate our instrumental variables (described in subsections 4.3.3 and 4.3.4).

Our sample consists of plants engaged in manufacturing. We recognize that human capital externalities can occur in other sectors, but there are reasons for restricting our sample. The most immediate one is that we do not have geocoded data for the other sectors of the economy. Moreover, in the case of the Brazilian economy, industry is the sector with the least informality,[6] which contributes to the representativeness of our study since we use data from the formal labor market. Another interesting characteristic is that, as confirmed

[4]Details of our database are presented in the Appendix of Chapter 2 of this dissertation.

[5]This survey provides annually general characteristics of the population, education, labor, income and housing, with household as the unit of survey. It is conducted in nine metropolitan regions (Belém, Fortaleza, Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo, Curitiba and Porto Alegre) chosen to be representative of the country at large.

[6]According to the PNAD data for the year 2014 provided by IBGE, the share of people employed in informal jobs in general industry (including extractive, transformation, electricity and gas, and water and sewage) is 23.9%, while the same measure for the construction sector is 57.9%, agriculture 73%, commerce 36.9%, and other services (excluding domestic) 33.4%.

in previous studies, manufacturing can benefit more from agglomeration economies (see, e.g., Barufi *et al.*, 2016) and human capital spillovers, which are greater when the sectors are economically close, since they presumably interact more in manufacturing (Moretti, 2004c).

The primary focus of the paper is the spatial scope of human capital spillovers observed from the individual's workplace. In order to achieve this focus, we first define the geographical context of our study. Our geocoded data allow us to freely define spatial units of measurement, so we exogenously divide the Brazilian territory into a uniform set of grid cells (1 km $\times$ 1 km) and associate each plant (point data) with its respective cell.[7] By using such small microgeographic units, we do not face the common problems of this type of approach when previously defined areas are used, and we minimize endogeneity problems associated with sorting. For example, Rosenthal and Strange (2008) used the Place of Work Public Use Micro Areas (PWPUMAs) for the US and found that the presence of large PWPUMAs can generate measurement errors in the agglomeration variables, biasing the estimates of the influence of agglomeration towards zero. So we don not need to impose any sample restrictions on the territory because our spatial measurement is homogeneous. In addition, official Brazilian geographic divisions for areas smaller than municipalities, such as census sectors[8] (equivalent to census tracts), are defined based on local factors such as the number of households. The choice of these areas as units of analysis in our approach may generate biased estimates due to simultaneity.

Initially, we had around 8.5 million cells, but not all cells, of course, have plants, and this varies every year as plants are created and/or existing plants are closed/moved.[9] To eliminate cells that are irrelevant for our purpose, we selected only those that have at least one plant and are within metropolitan areas existing in 2006 that encompass all five macro-regions of the country (see Figure C.2 in Appendix C.1 for more details). There are economic and technical factors that justify this restriction. Economically, the externalities generated by the agglomeration economies occur mainly in urban areas, from the concentration of workers and firms. Furthermore, in the Brazilian context the

[7] A similar strategy was used by Larsson (2014) and Andersson *et al.* (2014; 2019) for Swedish cities and by Li *et al.* (2020a) for China.

[8] Census data provide official spatial divisions called census sectors. By definition of IBGE, census sectors are the territorial units established for registration control purposes, formed by a continuous area, located in a single urban or rural setting, with the size and number of households that allow the survey by a census taker.

[9] An example of this division can be seen in Figure C.1 in Appendix C.1, which presents the division of the four largest Brazilian metropolitan regions. We only present cells that contained at least one manufacturing plant in 2014, the others are omitted.

population and economic activity are mainly concentrated in urban areas. Based on data from the 2010 Census, 84.4% of the population lived in urban areas, occupying 1.07% of national territory. Technically, by choosing only the metropolitan areas we are working with spatially smaller municipalities compared to the less urbanized municipalities. This provides more variation in our instrument for the number of college-educated workers, as explained in the next sections.

We define the centroids of these exogenous cells and from them, to capture the agglomeration economies at different distances, we follow Rosenthal and Strange (2003; 2008) by specifying five concentric ring variables, each of which measures the number of workers present at a given distance from the individual's workplace: between 0-1, 1-5, 5-10, 10-20, and 20-40 km. The motivation for choosing the size of the concentric rings is related to the spatial extension of Brazilian cities. The smallest ring can be considered to cover effects at a geographic level smaller than neighborhood. The next two distance ranges, 1-5 and 5-10 km, cover the distances of most common commuting distances within core cities. The two distance bands further away from the centroid, 10-20 and 20-40 km, cover commuting from neighboring cities to the core city and interactions at the level of the metropolitan region as a whole, respectively (see Figure 4.1).

In this context, another important difference from the strategy employed by Rosenthal and Strange (2008) is their assumption that employment in each PWPUMA is uniformly distributed throughout the given PWPUMA. Then, for each concentric ring, the authors had an approximation weighted by the areas of the PWPUMA forming the concentric ring of the true number of workers. Our microgeographic database provides the location of each plant, so we can get more precise measures of the number of workers in each concentric ring, which consequently minimizes potential endogeneity problems associated with measurement errors. Still in this respect, we also have information for different years, which allows us to observe the trajectory of a plant and its workers over the years, therefore controlling for any observed and unobserved heterogeneity that is fixed in time and minimizing potential endogeneity problems associated with omitted variables. We will return to this issue in the next subsection.

To calculate the individual wages in all models, we impose a few more restrictions on our sample. Hourly wage rates are calculated by dividing monthly wage[10] earnings by the usual number of hours worked per week and the number of weeks worked in the month by male workers between the ages of 18 and 56 who work than 20 hours or more per week.

[10]Nominal wages in December deflated by the National Wide Consumer Price Index (*Índice Nacional de Preços ao Consumidor Amplo*, or IPCA) (2017=100).

This sample restriction implies that the remaining workers exhibit a lower unobserved variation in possible endogenous decisions to work full time and increases the possibility that any remaining variations are absorbed by the control variables (Rosenthal and Strange, 2008; Håkansson and Isacsson, 2019). To calculate the agglomeration variables in each concentric ring, we used both restricted and unrestricted databases.

4.3.2 Empirical model specification

The objective of the article is to estimate the spatial scope of human capital spillovers. For this purpose, based on the theoretical structure presented above, which suggests a positive relationship between individual workers' wages and the concentration of college-educated workers, with a decreasing effect with geographic distance, we propose an empirical specification to assess the spatial extent of these effects. Formally our worker-level specification is given by:

$$\mathsf{w}_{izt} = \mathbf{X}_{it}\lambda + \mathbf{H}_{jt}\gamma + \sum_{r} \beta_r \mathsf{S}_{rt} + \alpha_i + \mu_c + \psi_{pt} + \epsilon_{izt} \tag{4.9}$$

where w_{izt} is the natural log of the real hourly wage of worker i in cell z (in plant j, industry p, metropolitan region c), and year t; $\mathbf{X}_{it}$ is a matrix of worker-level control variables such as age, age squared, tenure, and education; $\mathbf{H}_{jt}$ is a matrix of plant-level control variables, such as size; S_{rt} is our explanatory variable of interest and represents the number of workers with college degree or higher in each concentric ring r;[11] α_i, μ_c and ψ_{pt} are worker, metropolitan region, and industry-year fixed effects, respectively; and ϵ_{izt} is the error term.

Our parameters of interest are β_r, $r = 1, \cdots, 5$. The challenge of the exercise is to identify variation in the individual wages that is driven by concentration of college-educated workers around the individual's workplace and hence exogenous to other factors that affect local wages. If the location of the college-educated workers were random, this parameter would capture the effect of the concentration of college-educated workers in each distance band on the wages. There are, however, different mechanisms that make the hypothesis of S_{zrt}'s exogeneity doubtful.

One source of endogeneity common in these approaches, as explained by Rosenthal and Strange (2008), is measurement error. To deal with potential problem, as we discussed

[11] Note that we can generalize this structure and incorporate the urbanization (economic mass) effects measured by the total sum of employment in each ring as in Rosenthal and Strange (2008).

earlier, our geocoded data allow us to set each point (plant) in the exact concentric ring to which it belongs. Therefore, S_{zrt}, $r = 1, \cdots, 5$, is measured more precisely and we considerably reduce the part of the measurement error included in the residual of equation 4.9, ϵ_{izt}.

Another source of endogeneity is associated with omitted variables correlated with the concentration of college-educated workers in the concentric rings that surround the individual's workplace. Since individuals choose where to live and work, it is obvious that spatial sorting of observable and unobservable characteristics may bias our estimates of the human capital spillover effects. We address the potential endogeneity caused by spatial sorting in three ways. First, our micro-geographic units of one square kilometer are generally outside the set of workers' locational choice and thus the choice of the surrounding concentration of college-educated workers. Second, we introduce a set of control variables for characteristics related to the workers and establishment that may be relevant in this context, such as age, age squared, degree of education, tenure, tenure squared, and occupation for workers (all included in $\mathbf{X}_{it}$); and plant size ($\mathbf{H}_{jt}$). Third, we also include the worker fixed effects that control for unobserved individual characteristics, such as "ambition" or "ability".

In addition to these controls, we also include metropolitan region fixed effects, μ_c, which absorb time-invariant metropolitan region characteristics and conditions, such as geographical location, industrial structure, weather and amenities; and to control for industry common time trends, like sector-specific growth path at 2-digit level, we include industry-by-year fixed effects, ψ_{pt}.

Even using exogenously defined cells and a series of controls in equation 4.9, β_r can be biased by the influence of unobservable confounding trends. Any unobserved time-varying factors that affect simultaneously both wages and concentration of college-educated workers can make our estimates of human capital spillovers biased, e.g., transitory productivity shocks that attract highly educated workers and raise wages, $\text{cov}(\epsilon_{izt}, \mathsf{S}_{zrt}) \neq 0$. Thus, we cannot guarantee that our human capital variables calculated for each ring are exogenous.

We address these potential concerns by proposing instruments for our potentially endogenous human capital variables. Our identification strategy consists of using shift-share instrumental variables (SSIV). This approach, which uses weighted averages of a common set of shocks, with weights reflecting heterogeneous shock exposure, is increasingly common in many contexts (e.g., Bartik, 1991; Blanchard *et al.*, 1992; Autor *et al.*, 2013) and have had their properties formally discussed in the recent literature (see Borusyak *et al.*, 2018; Adao *et al.*, 2019; Goldsmith-Pinkham *et al.*, 2020). First, we measure

the population with college degree or higher based on differences in the demographic structure in each concentric ring in 1991. Next, we predict the number of people with that educational level in each concentric ring by the national change in the share of the population with college degree or higher as a result of changes in the federal government's educational policy in the period 1991-2004, weighted by population with college degree or higher in 1991 by age groups. We present a more detailed discussion of our shift-share regression designs in the next section.

4.3.3 Educational policy changes and identification

Our analysis exploits large shifts in national education policy between 1991 and 2004 as an exogenous source of variation in the number of college-educated people across concentric rings within Brazilian metropolitan regions to identify the effect of the concentration of college-educated workers in the concentric rings that surround the individual's workplace on the individual wages (our proxy to labor productivity). As the 1991 Census data and the 2004 PNAD data show, in this period there was 39.6% growth in the share of the population with college degree or higher. The changes in the national higher education policy are also clearly seen when we evaluate the growth in the number of higher education institutions and the number of students enrolled in undergraduate programs. For example, the number of higher education institutions and the number of students enrolled in undergraduate programs in the 1995-2003 period grew by 108% and 120%, respectively.[12]

In addition to the large national variation in the number of people with college degree or higher caused by the change in higher education policy ("shift"), the spatial distribution of these people by age group is also heterogeneous across concentric rings. Part of these cross-ring contemporary differences was certainly due to varying economic environment, industrial structures, and labor demand influencing wage and employment growth. To reduce the chances that age structure is itself endogenous, we use lagged weights. That is, as a generalization of the variable used by Moretti (2004a), we use the demographic structure in each concentric ring in 1991 to define our "exposure shares". To the extent that the relative number of people of different cohorts varies across concentric rings, this will lead to differential trends in the number of college-educated people across concentric rings. That is, we construct an educational-policy-driven instrument that retains only

[12]Data from Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), a federal research institution linked to the Ministry of Education, available at <http://inep.gov.br>.

the portion of growth in the number college-educated people at concentric ring level attributable to national policy fluctuations.

As Moretti (2004a) pointed out for the US, each new generation has a larger share of more educated individuals – and therefore of the number of educated people when the country's population is growing or remains constant – so young people today are more educated, on average, than young people from previous decades. Particularly in the case of Brazil, as we have mentioned, this trend has been exogenously shifted upward due to educational policies. The identification comes from differences in the relative magnitude of the cohorts that entered and left the labor force between 1991 and 2004. For example, consider two identical concentric rings, except in the age structure. If in one of the rings the number of young adults is higher, then the number of college-educated people is expected to be higher in this ring. This increase in the number of workers depends, thus, on the specific demographic structure in each ring. This is Moretti's argument, since the demographic structure varies between cities (in our case rings), each has its own tendency to increase the share (in our case the number of college-educated workers).

To make our exogenous spatial division compatible with official census data, we used a similar strategy to Rosenthal and Strange (2008) and Verstraten *et al.* (2018), which is based on the area of municipalities contained in each ring to create geographic weights to be associated with the college-educated population in each cohort.[13] The measure is calculated based on the data from the 1991 Census at the municipal level about the number of workers with college degree or higher in each ring. To instrument the population with college in the ring of 0-1 km, we use data from the distance range 0-5 km; for 1-5 km, we use 20-40 km; for 5-10 km, we use 40-80 km; and for 10-20 km, we use 80-120 km. We highlight two main points that motivated this choice. The first is the concern that the exposure weights are themselves endogenous even though they are lagged in time. To reduce the chances of this occurring, we use both time and space lagged weights. The second is that increasing the width of rings farther from the centroid minimizes multicollinearity problems (Verstraten *et al.*, 2018).

Formally, the instrument for number of workers with college degree or higher is given by:

$$\text{College}_{rt} = \sum_{m}\sum_{c} \omega_{rc} \mathsf{P}_{cm1991} \ln\left(\frac{\mathsf{P}_{mt}}{\mathsf{P}_{m1991}}\right), \text{ with } \omega_{rc} = \frac{\mathsf{A}_{r\cap c}}{\mathsf{A}_{c}} \tag{4.10}$$

[13]For example, if a concentric ring includes all municipality 1 and 20 percent of the area of municipality 2, then population in the ring is set equal to the population in cohort m in municipality 1 plus 20 percent of the population in cohort m in municipality 2. For Brazilian data, the merger between the rings and the old census data is only feasible at the municipal level.

where $\mathsf{A}_{r\cap c}$ is intersection area between concentric ring r and municipality c; A_c is total area of the municipality; P_{cm1991} is population[14] with college degree or higher in municipality c and cohort m (we defined three age groups: young 16-25, middle-aged 26-50, and old 51-70) in reference year; and P_{mt} is national population with college degree or higher in cohort m and year $t = (2000, \cdots, 2004)$.

4.3.4 Geological instruments

Although it is not our main focus, as we discussed earlier, we can generalize our empirical model to capture part of the economic mass effects, since we are considering only the manufacturing sector in each concentric ring. To do this, we simply substitute S_{zrt} for the total number of workers in each distance band. We can interpret this model specification as a more general test for agglomeration economies, which includes a mix of effects, some of which even have opposite signs, such as congestion effects.

All endogeneity concerns mentioned in the previous subsections remain in this context. One way to deal with reverse causality in productivity-agglomeration economies relationships is the use of historical and/or geological instruments (see, e.g., Ciccone and Hall, 1996; Combes *et al.*, 2008; 2010). Here research we use local soil characteristics as instrument for the total number of workers in each of the distance bands. More specifically, we use data from the Brazilian Geological Survey[15] to compute two measures: the fraction of the ring underlain by sedimentary rock and the fraction classified as mainly having ultisols (red clay soils).

These soil characteristics are associated with the supply of buildings affecting the structure and consequently the spatial distribution of economic activities within cities (Combes *et al.*, 2010). The presence of sedimentary rock has a direct bearing on the foundation of building construction. For example, Rosenthal and Strange (2008) used the example of Manhattan as motivation and explained that the type of soil, in particular where bedrock is relatively accessible, is associated with the height of buildings. On the other hand, the physical properties of ultisols are commonly favorable for most agricultural and non-agricultural uses (West *et al.*, 1997). In this context, soil properties certainly drove the location of populations when agriculture was the main sector of the economy,

[14] As in Moretti (2004a), the weights are estimated using data from the entire population, since the age structure of the labor force may be endogenous.

[15] We use data about the types of soils made available by Brazilian Agricultural Research Corporation (Embrapa) and data about geologic features made available by Geological Survey of Brazil (Companhia de Pesquisa de Recursos Minerais - CPRM).

but it is hard to imagine an effect on the current wages in the manufacturing sector (Combes *et al.*, 2011). The idea is that these variables affect wages only indirectly through the number of workers within each concentric ring.

4.3.5 Summary statistics

In this subsection we present some descriptive statistics that show how our data vary across distance bands. Table 4.1 summarizes descriptive statistics of concentric ring employment variables.[16] Column 2 shows the average number of workers, college-educated workers and number of plants, and column 3 shows the deviations in each concentric ring in 2006. Similarly, columns 4-7 show the same measurements for 2014 and for 2006-2014. It is interesting to note that the average number of college-educated workers increased, on average within five rings by 58.57% while general workers increased only 1.68%. As mentioned, this expansion can be associated with the exogenous increase of the share of population with college degrees in the period 1991-2004. Note also the high standard deviation, indicating the heterogeneity of the spatial distribution of employment within metropolitan areas. The bottom table presents similar measures for the number of plants in each ring, the number of cells in each year (2006 and 2014) and the average number of cells in the period 2006-2014. One possible problem associated with defining ring size is lack of variation of data within each ring, particularly in the smallest rings (0-1 km). Our data show that there is significant variation within the smaller rings, as can be observed both for general workers, such as college-educated workers, and for plants.

As an illustration of the heterogeneity, Figure 4.1 presents the spatial distribution of college-educated workers in the four largest metropolitan regions of the country (São Paulo, Rio de Janeiro, Belo Horizonte and Porto Alegre). Note that except the Rio de Janeiro Metropolitan Region (RJMR) (Figure 4.1 (b)), the areas with the highest concentration of college-educated workers in manufacturing are outside the core city. For example, in the São Paulo and Belo Horizonte Metropolitan Regions (SPMR and BHMR), the areas with a high concentration of college-educated workers are in the 10-20 km range (Figures 4.1 (a) and (c)), while in the Porto Alegre Metropolitan Region the range is 20-40 km range (Figure 4.1 (d)). The neighbors of the core municipality seem attractive to the large industries as they are still close to the central business district (CBD), so the industrial establishments can still benefit from the positive externalities while avoiding

[16]Descriptive statistics for worker characteristics by metropolitan region and for the sample as a whole are provided in Table C.1 in Appendix C.1.

Table 4.1 Descriptive statistics of concentric ring employment variables

	2006		2014		All sample	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
# of workers						
Within 0 to 1 km	568.33	1,241.33	573.00	1,238.43	595.26	1,332.80
Within 1 to 5 km	10,077.43	15,398.52	10,109.23	14,596.50	10,540.13	15,800.01
Within 5 to 10 km	24,964.54	38,857.74	25,077.06	36,588.83	26,157.23	39,677.05
Within 10 to 20 km	69,756.78	103,448.40	70,687.17	98,222.49	73,288.39	105,977.80
Within 20 to 40 km	130,705.50	178,786.20	137,873.50	179,009.90	140,154.60	188,226.70
College-or-more, 0 to 1 km	44.78	181.39	70.75	323.24	58.82	282.45
College-or-more, 1 to 5 km	846.75	2,018.22	1,329.71	2,965.81	1,107.00	2,608.39
College-or-more, 5 to 10 km	2,227.10	4,918.54	3,497.00	7,027.02	2,919.03	6,256.79
College-or-more, 10 to 20 km	6,630.83	12,332.47	10,321.55	17,513.41	8,651.46	15,616.67
College-or-more, 20 to 40 km	12,118.67	20,452.63	19,990.25	30,525.30	16,324.91	26,688.37
# of plants						
Within 0 to 1 km	22.77	47.57	24.21	44.68	23.76	47.76
Within 1 to 5 km	417.16	674.58	440.14	645.45	433.58	673.00
Within 5 to 10 km	1,021.34	1,624.74	1,074.58	1,574.11	1,060.68	1,624.94
Within 10 to 20 km	2,741.58	4,037.25	2,897.68	3,975.69	2,845.82	4,058.87
Within 20 to 40 km	4,702.92	6,571.15	5,144.90	6,750.07	4,972.65	6,750.83
# of cells	14,030		16,094		15,099[a]	

Notes: The number of cells in Brazilian metropolitan areas represents the cells with at least one plant. In 2006, for example, our analysis encompasses 14,030 km^2 of urban areas. [a]: average number of cells. Source: Author' computations using information from RAIS.

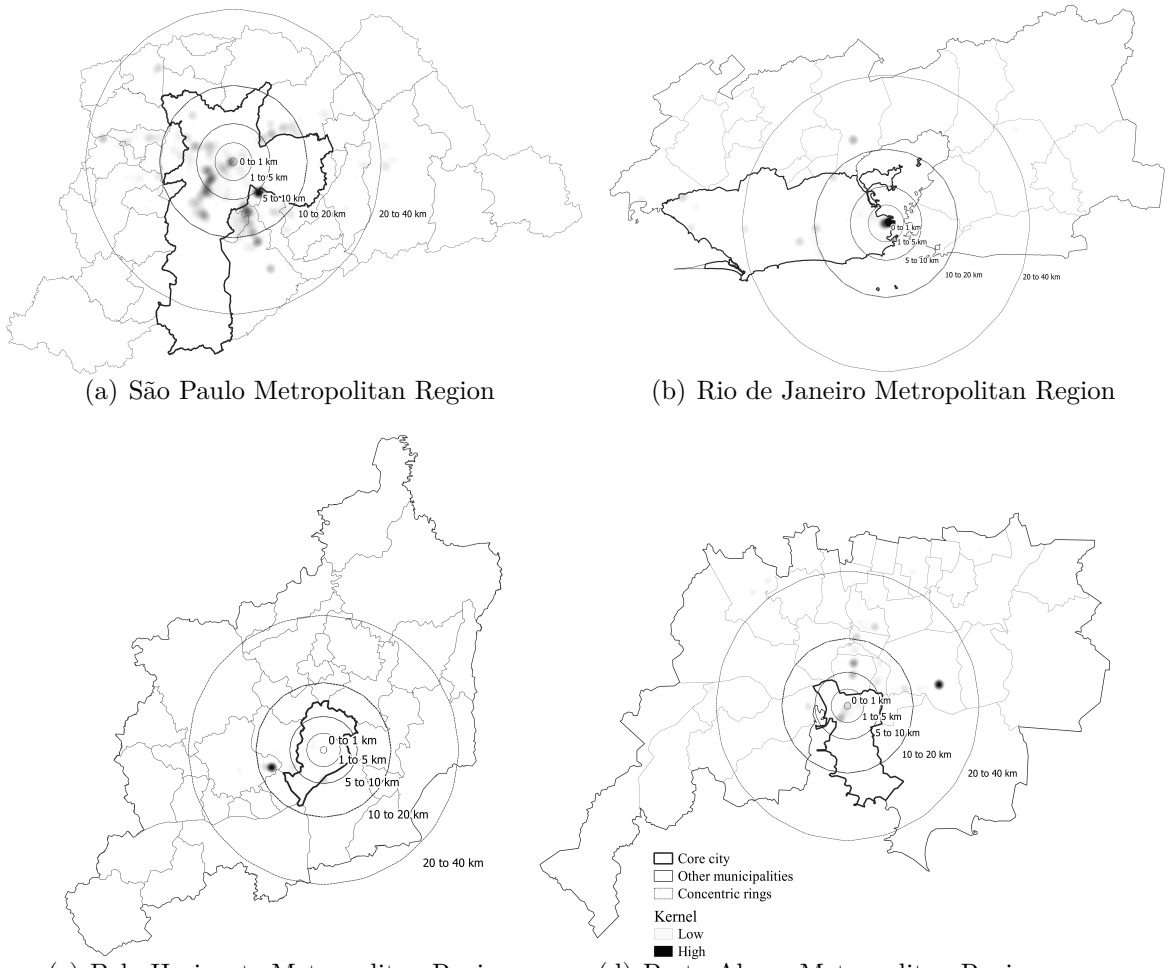
most congestion effects. This location pattern is in conformity with the pattern commonly found in literature on city structure (see, e.g., Anas *et al.*, 1998; Anderson and Bogart, 2001; Coffey and Shearmur, 2002; Billings and Johnson, 2012).

4.4 Results

In this section we present and discuss the results. While not our main focus, we begin by presenting the results of a more general test in a narrower cross section data context, on the spatial extent of effects of the surrounding economic mass of individuals' workplace on wages. The remaining subsections present our main results about the attenuation of human capital spillovers.

4.4.1 Spatial scope of agglomeration gains

As a preliminary exercise, we seek to understand how the local externalities associated with geographic proximity are attenuated with distance, or in other words, how individuals' productivity can be affected by the concentration of other workers at different distances from their work place. As mentioned by Rosenthal and Strange (2008), this question can



(a) São Paulo Metropolitan Region (b) Rio de Janeiro Metropolitan Region

(c) Belo Horizonte Metropolitan Region (d) Porto Alegre Metropolitan Region

Figure 4.1 Distribution of workers with college-or-more within select metropolitan regions

Notes: Kernel density is estimated using workplace data for workers with college-or-more in 2014. We selected a cell within the core city from which we defined five concentric rings around its centroid.

be answered by considering linear changes in the spatial distribution of employment. So, in all the estimations presented below we use log-linear models. For all models starting now, the variables associated with individual characteristics such as age, age squared, dummies for different levels of education[17] and tenure were consistent with estimates in the labor literature and are not reported, but only mentioned as worker-level controls. In this set of controls we also include tenure squared to capture nonlinearities and dummies

[17]Illiterate, incomplete primary school, complete primary school to incomplete high school, complete high school to incomplete college and college degree or more.

for the different 2-digit occupations.[18] Like for plants, dummies associated with plant size are mentioned as plant-level controls.

We begin by exploring the general effects of economic mass on the productivity of workers at different distances. To do this, we use the total number of manufacturing workers within 0-1, 1-5, 5-10, 10-20 and 20 to 40 km from the individual's workplace. As previously discussed, we do not have geocoded data for all sectors of the economy; our database only covers manufacturing.[19] So, the estimates capture a mix of forces that can be associated with both localization and diversification economies and congestion effects. Because of this, the expected effect on each of the rings is unclear. For example, the microfoundations of local externalities, namely labor market pooling, input sharing, and knowledge spillovers, operate at different geographic scales (Rosenthal and Strange, 2020). Empirically, as the results of Rosenthal and Strange (2003) and Li *et al.* (2020a) show for the US and China, respectively, the localization and diversification effects can vary considerably between industries.

As we mentioned above, to deal with endogeneity of economic mass in the wage earnings equation, we use instrumental variables based on local soil characteristics. The hypothesis behind this strategy is that soil characteristics affect the supply of buildings and consequently commercial and residential development, which in turn affects economic mass in each distance band. Some studies about the spatial extent of agglomeration economies use local soil characteristic as an instrument for local employment (see, e.g., Rosenthal and Strange, 2008; Håkansson and Isacsson, 2019; Li *et al.*, 2020a). Nevertheless, a limitation imposed by the use of these instruments is that they do not vary across years, which limits the models to the use of cross-section data. Because of this, in this initial more general analysis, all models are estimated with cross-section data, more specifically for 2010 (last available census).

The results can be seen in Table 4.2. Column 1 reports the results of a simple ordinary least squares (OLS) association conditional on observable worker and plant characteristics but without any control for local characteristics. Although it is a simple exercise that hides, of course, several other observable and unobservable effects besides simultaneity, there is already a suggestive pattern in the results, that is, the distance from the current

[18]We use the Brazilian Occupation Classification version from 2002 that is compatible with the International Statistical Classification of Occupations (ISCO-88).

[19]We recognize that the effects of urban density can extend beyond manufacturing. Generally, the large urban centers and, to be more precise, the core city, concentrate other sectors such as FIRE, health, education, and other kind of services (see, e.g., Almeida *et al.* (2021) for the case of SPMR). Because of this our explanatory variables in each ring should be interpreted as a proxy that captures part of the effects of urbanization.

establishment matters. In column 2 we include a set of dummies for each industry at the 2-digit level and metropolitan region. The spatial decay becomes clearer, larger in the first ring (0-1 km) and decreases until the last ring (20-40 km). In column 3, we include different controls for existing transportation infrastructure (distance in kilometers from the cell's centroid to the nearest railroad, federal highway, state highway, airport, and to the nearest port) and geographic characteristics (distance in kilometers from the cell's centroid to the nearest river/lake) around the individual's workplace. There is evidence that local transportation infrastructure (see, e.g., Holl, 2016; Mayer and Trevien, 2017; Gibbons *et al.*, 2019) and proximity to rivers or lakes (see, e.g., Ellison and Glaeser, 1999; Ellison *et al.*, 2010; Rosenthal and Strange, 2001) can influence the location of plants and their productivity, and therefore influence wages at these plants. Even after including these controls, the general pattern of attenuation remains.

Column 4 reports the results obtained using instruments for the agglomeration variables in each ring, estimated by the generalized method of moments (GMM). The coefficients of the first ring remain positive and strongly significant while the second ring is not significant, the third ring is negative and significant and the last two are not significant. Two patterns emerge when we use the IV-GMM estimator. The first already appeared in OLS models and is associated with the externalities captured by the first coefficient, indicating that the wages are higher, on average, when there are more workers up to 1 km from the current establishment. In particular, the first coefficient is higher (20 times) than the first coefficient in the most complete OLS specification (column 3), indicating the existence of a negative bias in the estimation when we assume exogeneity of our agglomeration variables. Second, in the IV model, in the third ring the coefficient is negative, an effect not captured by OLS models. This possibly indicates that the positive effects associated with concentration are not strong enough from the third ring outward to compensate for the dispersion forces, making the net result negative. The negative sign may also indicate some kind of competition effect between the rings.

At the bottom of Table 4.2 (column 3) we report the instrument diagnostic test results. To test the over-identification condition, we used the Hansen J-test. As can be seen, the statistics fail to reject the over-identifying restrictions, and thus are consistent with the idea that instruments are exogenous. To detect weak instruments, we can observe the first-stage partial F-statistics of the excluded instruments (Stock *et al.*, 2002). These measure whether the coefficients of the excluded instruments are significantly different than zero. As a "rule of thumb", they are expected to be large (Staiger and Stock, 1997). As can be seen, the F-statistics of the first stage are large and meet the established

Table 4.2 Spatial scope of agglomeration externalities

Dependent variable: individual hourly wage (in log)				
# of workers	OLS	OLS	OLS	GMM
	(1)	(2)	(3)	(4)
0 to 1 km	1.04e-05***	5.89e-06***	6.27e-06***	1.28e-04***
	(2.57e-07)	(2.65e-07)	(2.67e-07)	(3.43e-05)
1 to 5 km	7.07e-07***	8.81e-07***	5.25e-07***	2.45e-07
	(5.66e-08)	(5.59e-08)	(5.83e-08)	(3.62e-06)
5 to 10 km	9.31e-07***	6.80e-07***	7.41e-07***	-5.00e-06*
	(3.74e-08)	(3.76e-08)	(3.91e-08)	(2.99e-06)
10 to 20 km	-7.27e-08***	2.12e-08	2.32e-07***	6.20e-07
	(1.69e-08)	(1.88e-08)	(2.05e-08)	(8.10e-07)
20 to 40 km	6.01e-07***	1.65e-07***	2.21e-07***	7.87e-08
	(6.80e-09)	(1.74e-08)	(1.88e-08)	(6.84e-07)
Worker-level controls	Yes	Yes	Yes	Yes
Plant-level controls	Yes	Yes	Yes	Yes
Industry FE	No	Yes	Yes	Yes
Metropolitan region FE	No	Yes	Yes	Yes
Transport infrastructure	No	No	Yes	Yes
Geographical characteristics	No	No	Yes	Yes
Hansen-J over id test[a]				3.89
Kleibergen-Paap rk F[b]				1.34
Kleibergen-Paap rk LM[c]				4.25
1 st stage F-stat. 0 to 1 km				27.93
1 st stage F-stat. 1 to 5 km				59.66
1 st stage F-stat. 5 to 10 km				48.43
1 st stage F-stat. 10 to 20 km				14.83
1 st stage F-stat. 20 to 40 km				68.09
F-stat.	11,417.83	8,317.09	7,709.76	–
R squared	0.5598	0.6256	0.6252	–
Observations	508,457	508,457	497,259	497,259

Notes: This table presents the estimates obtained from equation 4.9 when we consider the total number of workers in the manufacturing industry in each ring. Worker-level controls include all the individual characteristics detailed above. Plant-level controls are dummies for plant size. Industry FE are dummies for industries at the 2-digit level. Metropolitan region FE are dummies at the metropolitan region level. Transport infrastructure includes the distance in kilometers from the cell's centroid to the nearest railroad, nearest federal highway, nearest state highway, nearest airport, and nearest port. Geographic characteristics include the distance in kilometers from the cell's centroid to the nearest river or lake. Robust standard errors are in parentheses. The 1 st stage F-statistic is the F test of excluded instruments. [a]: H_0 - all instruments are valid. [b]: H_0 - weakly identified model. [c]: H_0 - under-identified model. Significance levels: *** $p < 0.01$. Source: Prepared by the author based on estimates.

conditions. We also report a more rigorous test of instrument relevance, calculation of the Kleibergen-Paap Wald F-statistic, which is valid in the presence of non-homoscedastic errors. As reported in Rosenthal and Strange (2008), these test statistics are sensitive to the manner in which the model's standard errors are clustered. In this study, clustering the standard errors at the metropolitan region level greatly lowered the Kleibergen-Paap

F-statistic, increasing the tendency to view the instruments as weak. The Kleibergen-Paap F-statistic was smaller than the "rule thumb" generally adopted in literature, according to which the F statistic should be at least 10 for weak identification not to be a problem (Baum *et al.*, 2007). The same applies to under-identification testing. For this we used the Kleibergen-Paap Lagrange multiplier test, which is valid under heteroscedasticity (Kleibergen and Paap, 2006).

Evidence for other countries also suggests that agglomeration economies measured from economic mass are stronger when located near the individual's workplace.[20] For example, stronger up to 8 km and attenuated up to 80 km for the US (Rosenthal and Strange, 2008). Addario and Patacchini (2008) showed it is strongest up to 4 km and attenuated sharply until 12 km in Italy. More recently, the Håkansson and Isacsson (2019)'s study of Sweden also indicated a negative effect. In some specifications, the authors found negative effects from 25 km of the individual's workplace. In the Brazilian context, this effect appears after 5 km. When observed from this standpoint, the speed of attenuation on wages is much greater. In other words, in the Brazilian case, the local externalities generated by the spatial concentration of workers appear more localized than in developed countries. We must be cautious when evaluating these results. We recognize that in wage regressions there are unobservable effects such as ability or family background that are correlated both with wages and economic mass between and within cities. We address this issue directly in the next subsection, when our focus is only a specific source of local externality, the spatial concentration of college-educated workers.

4.4.2 Spatial scope of human capital spillovers

As pointed out earlier, in the previous subsection we explored a mix of factors that may be correlated with worker productivity. In this subsection we focus on the attenuation of external returns to education. For this, we consider only the number of workers with college degree or more in each concentric ring. The social return to education (private return plus external return) can be associated with different factors, for example, crime reduction, more informed political decisions when voting and enhanced productivity of other workers (see, e.g., Lochner and Moretti, 2004; Milligan *et al.*, 2004; Moretti, 2004b). We focus exclusively on the third example, after discounting the private returns. Proximity to educated workers can enhance productivity of other workers (Moretti, 2004a).

[20] There is a body of evidence on the spatial extent of the agglomeration economies using other techniques, but which generally points in the same direction. See, e.g., Rice *et al.* (2006) for UK, Rosenthal and Strange, (2003; 2005) for US and Andersson *et al.*, (2014; 2019) for Sweden.

Proximity determines the intensity of the effect we are evaluating. So, we now consider only the four closest rings of the individual's workplace, i.e., 0-1, 1-5, 5-10 and 10-20 km. There are two main reasons. First, the human capital externalities occur mainly at short distances (Fu, 2007; Rosenthal and Strange, 2008, 2020; Li *et al.*, 2020b). Second, as presented in Figure 4.1, part of the 20-40 km ring is outside the metropolitan regions.

Here we provide only the results obtained from the restricted sample (Table 4.3).[21] Column 1 reports the pooled OLS estimates. In column 1, we address the spatial classification into observables by controlling for the individual employee characteristics presented above; we also control for observed plant-level heterogeneity, to be precise, controlling for plant size; add time-varying industry-specific effects at the 2-digit level (controlling for industry-specific productivity shocks that vary across years and can affect worker earnings); and we add the metropolitan region fixed effect to control for heterogeneity across geographic areas. The coefficients of the first three rings are positive and strongly significant while the coefficient associated with the 10-20 km ring is negative and strongly significant. This exercise indicates a pattern of decay of the externalities generated by the number of college-educated workers, stronger in the range of 0-1 km and smaller in the other rings.

To understand how the pattern of spatial attenuation changes as we include the heterogeneities highlighted above, we also report the results obtained by estimating simpler models in Table C.4 in Appendix C.2, i.e., without controlling for plant, industry-year, and metropolitan region heterogeneities. There are differences between these simple initial specifications, particularly with regard to the magnitude of the coefficients. But a common feature in all is that the positive association between the number of high-skilled workers and wages are greater at short distances (0-1 km).[22] Note also that the coefficient associated with the 10-20 km ring becomes negative when we include the region fixed effects (column 1 in Table 4.3). One possible interpretation is that two effects can act simultaneously within each ring, a positive one associated with external return to education and a negative one associated with competition between different locations. Nevertheless, when we control for unobserved heterogeneity fixed in time at the metropolitan region level (comparing rings within the same metropolitan region), the positive effect is strong

[21] The results for the unrestricted sample can be seen in Table C.3 in Appendix C.2. Our main results, about attenuation with geographical distance, remain valid.

[22] All coefficients associated with the 0-1 km ring are positive and strongly significant. The intensity of decay between ranges, however, is very different. For example, in models OLS 1 to 3 in Table C.4, the 0-1 km coefficient for proximity to college-educated workers is, on average, 4 times larger than the corresponding 1-5 km effect and 27.7 times larger than the corresponding 5-10 km effect.

enough to offset the negative only at short distances (up to 10 km).

Table 4.3 Spatial scope of human capital spillovers

Dependent variable: individual hourly wage (in log)			
# of workers with college-or-more	OLS (1)	FE (2)	FE + IV (3)
0 to 1 km	2.49e-05***	6.68e-06***	6.78e-05***
	(4.31e-07)	(3.28e-07)	(6.25e-06)
1 to 5 km	9.04e-06***	-2.57e-07	1.95e-05***
	(1.62e-07)	(2.24e-07)	(2.44e-06)
5 to 10 km	2.21e-06***	-6.90e-07***	1.27e-05***
	(9.43e-08)	(1.44e-07)	(1.55e-06)
10 to 20 km	-3.09e-07***	-5.75e-07***	-1.06e-05***
	(5.20e-08)	(9.08e-08)	(6.62e-07)
Worker-level controls	Yes	Yes	Yes
Plant-level controls	Yes	Yes	Yes
Industry × year FE	Yes	Yes	Yes
Metropolitan region FE	Yes	Yes	Yes
Kleibergen-Paap rk F[a]			1,466.29
Kleibergen-Paap rk LM[b]			5,346.92
1 st stage F-stat. 0 to 1 km			2,363.64
1 st stage F-stat. 1 to 5 km			6,652.19
1 st stage F-stat. 5 to 10 km			20,791.46
1 st stage F-stat. 10 to 20 km			60,470.27
F-stat.	21,639.14	–	–
R squared	0.6347	0.3827	–

Notes: This table presents the estimates obtained from equation 4.9 when we consider the number of college-educated workers in each ring. All models are estimated with 2,387,434 observations. Worker-level controls include all the individual characteristics detailed above. Plant-level controls are dummies for plant size. Industry × year effects are dummies for each 2-digit × year combination. Metropolitan region FE are metropolitan region fixed effects. The 1 st stage F-statistic is the F test of excluded instruments. [a]: H_0 - weakly identified model. [b]: H_0 - under-identified model. Standard errors adjusted for clustering are in parentheses. Significance levels: *** $p < 0.01$. Source: Prepared by the author based on estimates.

Since human capital externalities occur mainly through interaction among workers, the spatial decay can be affected (e.g., be stronger) when the frequency of contacts reduces rapidly with distance (Rosenthal and Strange, 2020). So, the heterogeneity of structure and provision of public services in urban areas can help to understand which occurs when the distance from and individual's workplace increases. Consequently, the local public services, including the provision of public transport infrastructure, can affect the interaction of workers further apart. Notice, however, that these factors may help to understand why geographical proximity is important, but do not provide an interpretation for the negative effects. One interpretation of a negative effect, which is also associated

with the structure of cities, is the competition between different locations (Håkansson and Isacsson, 2019). The expansion of the number of high-skilled workers at greater distances from an individual's current establishment can be associated with a reduction in the number of workers around the current establishment.

In columns 2 and 3 we present the results after including the individual-specific fixed effect (FE) and use the shift-share instrument for the number of college-educated workers in the four distance bands presented in equation (4.10) (FE + IV). The estimates with individual-specific fixed effect (here any individual permanent characteristics are controlled) and all other controls, but without instrumental variables, indicate the importance of geographic proximity of high-skilled workers only 1 km from the individual's workplace. For high-skilled workers farther away from an individual's current establishment, there may be a kind of competition, as in the result for the last concentric ring in column 1. Although all these controls for observable and non-observable variables are important for us to obtain cleaner effects, as mentioned, we will use the nationwide exogenous growth of college-educated workers in a shift-share design to instrument the number of workers with college degree or more in each concentric ring. As can be seen at the bottom of column 3, using the instrumental diagnostic tests mentioned above, the hypotheses of weak instruments and under-identification are strongly rejected. Additionally, as can be seen in Table C.6 in Appendix C.2, which presents the first stage estimates, even after including all control variables and fixed effects presented above, the shift-share IV has strongly significant explanatory power.

The qualitative results regarding the spatial extent of human capital externalities in column 3 resemble those in column 1, but the coefficients for the three rings closest to the current establishment are, on average, 3.5 times larger, indicating the existence of a negative bias in the estimation when we assume exogeneity of our human capital variables. The results of the FE + IV model show that if 1,000 college-educated workers are added at distance up to 1 km (approximately equivalent to the 10/90 spread in Table C.2 in Appendix C.2),[23] wages of workers would increase, on average, by 6.78 percent. On the other hand, if the same number of workers were added to the 1-5 km or 5-10 km range, the wages of workers would increase, on average, by 1.95 and 1.27 percent, respectively.

In column 2, a clear exception is the coefficient of the third ring. The coefficient obtained by the fixed effects estimator without instrumental variables is negative, which may be associated with some kind of simultaneous bias farther from the individual's

[23]Table C.2 presents sample percentiles (the 10th, 25th, 50th, 75th, and 90th) of the concentric ring employment variables.

workplace. These details may indicate it is important to consider endogeneity of proximity of college-educated workers in the wage earnings-human capital spatial distribution relationship. So, summing up all the results of Table 4.3, the main impression is that in Brazilian cities the external returns to education are positive up to 10 km from the individual's workplace. For high-skilled workers farther away from an individual's current establishment, there may be a kind of competition, as indicated by the results in columns 1-3. These results remain strongly significant even after controlling for the private return to education and other observable characteristics at worker level and plant level, along with productivity shocks specific to industry × year, metropolitan region and worker fixed effects, and using to instrumental variables.

In general, our main results are in line with previous evidence in the literature. But unlike most of previous studies, as we have discussed so far, we do not assume that the external return to education is homogeneous within a city. Instead, it attenuates rapidly with distance. At more aggregate levels, for example, both for workers (relating local human capital stocks to wages) and for firms[24] (using TFP), geographic proximity of highly skilled workers increases productivity. Moretti (2004a), for example, found that the elasticity of wages to the share of college graduates at the Metropolitan Statistical Area (MSA) level in US was around 1.2, with small variations with different specifications. More recently, Chauvin *et al.* (2017) indicated that the elasticity was 3.0 to 4.7 for Brazil, 5.2 to 7.2 for China, and 1.9 to 3.2 for India. For all three developing countries, the coefficients are higher than in the US. Similar empirical evidence can be found in other studies on the magnitude of human capital spillovers using different strategies to deal with the endogeneity of aggregate human capital (see, e.g., Moretti (2004b), and more recently Carlino and Kerr (2015) for a detailed review).

The contribution of this paper, however, also provides new insights into these effects within the Brazilian cities, i.e., when our lens narrows to the neighborhood level. In the latter aspect, the evidence is almost exclusively for developed countries. For example, Fu (2007) found that knowledge spillovers were very localized, occurring mainly and strongly at short distances in the Boston Metropolitan Area, within round 2.4 km (in models with 5 rings), which the author called "Smart Café Cities". Rosenthal and Strange (2008) found similar results for the whole US. More specifically, when college-educated workers were less than 8 km away from an individual's workplace, they generated a greater external return than those college-educated workers who were more than 8 km away. With a different

[24]See, e.g., Moretti (2004c) and Liu (2013), who related local human capital stocks to TFP in the US and China, respectively.

empirical strategy but still addressing the same theme, Andersson *et al.* (2009) evaluated the effects of spatial decentralization higher education in Sweden and found substantial and highly localized spillovers (about 5 km from the new university) in productivity gains.

In addition to providing evidence about the attenuation of human capital spillovers in a context very different from that observed in developed countries, and differently from other previous studies, such as Rosenthal and Strange (2008), our panel data combined with the exogenous shock driven by the Brazilian government's education policy shift allows us to obtain the causal effects more clearly. Furthermore, our results conform very well to the Brazilian economic environment. They are in agreement with the high level of geographic concentration of manufacturing, in particular with the correlation concentration $\times$ share of college-educated workers (as shown in the chapter 2 of this dissertation); with the higher inter-regional mobility of workers, which favors the formation of more specialized and dynamic industrial clusters; and with the high level of educational disparities observed in the country.

4.4.3 Evidence for different education groups

In the previous subsection, we assumed that the estimates of human capital spillovers are the same for different employees, i.e., when all skill groups are pooled together. The results obtained are, therefore, an average effect across education groups. There are different reasons why this simplification may not be valid. As predicted by a conventional demand and supply models, the effects for less educated workers tend to be greater. Nevertheless, in spillover models, both types of workers can gain from the local increase in the number of college-educated workers. In particular, the less educated workers benefit even in the absence of any spillovers, while on the other hand, the effect on the wages of college-educated workers depends on the existence of the spillovers (see Moretti, 2004a). Thus, to explore these possible differences, we estimated equation 4.9 by separating the sample into two education groups: less than college degree and college degree or higher.

Table 4.4 reports the results. In Panel A we follow the same structure previously presented in Table 4.3, but our outcome variable is only the wages of less educated workers (less than college degree). Similarly, in Panel B we consider college-educated workers. Four important patterns emerge from this table. First, in both cases the proximity to college-educated workers increases an individual's wage. This is indicated by the positive and highly significant coefficients in the first distance band in all columns and for second and third distance bands in most specifications. Second, in both cases it is important to

consider the spatial extent of the spillover effect. This effect is more pronounced up to 1 km from an individual's current establishment. Third, comparing the results in the first ring of Panel A with those of Panel B, we can observe a pattern according to a model that includes both conventional demand and supply factors and spillovers. In particular, a greater effect (2.3 times) exists for less educated workers in the first distance band. An exception is column 2, with fixed effects estimation without instrumental variables, where the estimated coefficient is higher for college-educated workers. As mentioned regarding the results in Table 4.3, in the first distance band the qualitative results in columns 2 and 3 are not different, but in the other concentric rings there are clear differences, highlighting the importance of using instrumental variables. Fourth, according to the results for the second ring, college-educated workers can benefit from greater closeness to other college-educated workers (columns 1 and 3).

Particularly, there are significant differences in the attenuation pattern between the first and second distance bands between panels A and B when using OLS estimators (column 1). For example, while in Panel A the coefficient of the first ring is 10.9 times larger than the coefficient of the second ring, in Panel B the coefficient of the first ring is 2 times larger than the coefficient of the second ring. This difference, however, is smaller when using FE + IV estimators, as can be seen in column 3 of both panels. When moving farther away from the individual's workplace, the decay of the effect from the second to the third ring is greater for the college-educated group in columns 1. Moreover, when we use instrumental variables, there is no effect in the third ring for the college-educated group. That is, for the less educated group, the human capital externalities are positive and strongly significant up to 10 km. On the other hand, for the college-educated group, human capital externalities occur up to 5 km. A general consideration on these observations is that the attenuation between the first and second ring is larger for the less educated group, and for college-educated workers, in the 1-5 km range external returns are stronger than those observed for less educated workers.

To be more specific, from the results in column 3 for the less educated group (Panel A), the addition of 1,000 college-educated workers in the 0 to 1 km ring implies that wages of a less-educated workers would increase, on average, by 10 percent. For the college-educated sample, the corresponding effect is 4.6 percent. When we look at the second distance band, adding the same number of college-educated workers, wages of less educated group would increase, on average, by 2.6 percent while the wages of the highly skilled group would increase by 4 percent. In the third ring, as mentioned, there is significant effect only for the less educated group, of 1 percent. A possible interpretation of these results

Table 4.4 Spatial scope of heterogeneity of human capital externalities by education groups

Dependent variable: individual hourly wage (in log)			
# of workers with college-or-more	OLS (1)	FE (2)	FE + IV (3)
Panel A: Less than college degree			
0 to 1 km	5.08e-05***	4.59e-06***	1.10e-04***
	(5.91e-07)	(4.34e-07)	(1.30e-05)
1 to 5 km	4.64e-06***	-3.19e-07	2.63e-05***
	(1.77e-07)	(2.56e-07)	(2.97e-06)
5 to 10 km	2.11e-06***	-4.21e-07**	1.03e-05***
	(1.00e-07)	(1.65e-07)	(2.44e-06)
10 to 20 km	-1.58e-07***	-7.46e-07***	-1.09e-05***
	(5.52e-08)	(1.03e-07)	(7.95e-07)
Kleibergen-Paap rk F[a]			17.39
Kleibergen-Paap rk LM[b]			69.21
1 st stage F-stat. 0 to 1 km			1,353.64
1 st stage F-stat. 1 to 5 km			5,606.02
1 st stage F-stat. 5 to 10 km			16,936.74
1 st stage F-stat. 10 to 20 km			34,172.36
R squared	0.5241	0.3664	–
Observations	2,003,730	2,003,730	1,994,174
Panel B: College degree or more			
0 to 1 km	3.09e-05***	6.88e-06***	4.58e-05***
	(1.12e-06)	(8.62e-07)	(7.19e-06)
1 to 5 km	1.44e-05***	-1.10e-06	4.10e-05***
	(4.74e-07)	(6.86e-07)	(1.03e-05)
5 to 10 km	2.81e-06***	-1.62e-06***	-4.25e-06
	(3.06e-07)	(4.17e-07)	(5.88e-06)
10 to 20 km	-1.31e-06***	-2.21e-07	-6.99e-06***
	(1.84e-07)	(2.80e-07)	(1.49e-06)
Kleibergen-Paap rk F[a]			41.56
Kleibergen-Paap rk LM[b]			162.73
1 st stage F-stat. 0 to 1 km			640.42
1 st stage F-stat. 1 to 5 km			1,477.87
1 st stage F-stat. 5 to 10 km			2,423.80
1 st stage F-stat. 10 to 20 km			4,161.27
R squared	0.4378	0.3484	–
Controls to Panel A and B			
Worker-level controls	Yes	Yes	Yes
Plant-level controls	Yes	Yes	Yes
Industry × year effect	Yes	Yes	Yes
Metropolitan region FE	Yes	Yes	Yes

Notes: All models in Panel B are estimated with 187,511 observations. The industry × year effect is computed at the 2-digit level. All the controls shown at the bottom of this table are included in both panels A and B. The 1 st stage F-statistic is the F test of excluded instruments. [a]: H_0 - weakly identified model. [b]: H_0 - under-identified model. Robust standard errors in parentheses. Significance level: ** $p < 0.05$, *** $p < 0.01$. Source: Prepared by the author based on estimates.

is that, as established in the supply and demand models, highly skilled workers are not perfect substitutes for less educated workers, so although they are geographically close, they are not close competitors. On the other hand, they can still generate externalities from the sharing of ideas, which can generate new products and productive processes or

improve existing ones. For educated workers, however, geographic proximity (up to 1 km) to other equally skilled workers may generate greater competition since they are perfect substitutes, but when moving further away from the individual's workplace (from 1 km), the competition effect is less.

While these comparisons are interesting, the main findings of this subsection are that the attenuation patterns found in the previous subsection remain largely valid regardless of the subgroup in our sample. This indicates that external returns to education decay with geographic distance within the same city regardless of worker type.

4.4.4 Robustness checks

In this subsection we report estimates for alternative specifications to check the robustness of our main results. These results were previously reported in column 3 of Table 4.3 and refer to the pattern of attenuation of human capital spillovers. That is, when we used instrumental variables in addition to a broad set of controls to deal with the potential endogeneity of our human capital variables. So, all results presented below are obtained from FE + IV estimators. Basically, our robustness check consists of exploring what happens to our results about the attenuation patterns when we change and/or add other controls, as well as when we restrict the sample by metropolitan regions and by industries.

One of the most immediate ways to test the robustness of our main results is, for example, to control for industry-specific trends at the 3-digit rather than 2-digit level. The higher the industrial aggregation (e.g., 2-digit), the greater the grouping of different sectors will be, so that specialized industry-specific trends may not be captured. For example, in the official classification of economic activities in Brazil (CNAE), some 2-digit sectors include 3-digit industries with different technological intensity (see Cavalcante, 2014) and therefore have workers with different skills and education. So, to check whether our results are sensitive to these effects, we include 3-digit industry-year specific effects. The results are provided in column 1 of Table 4.5. Note that although there are small variations in the magnitude of the coefficients, all remain strongly significant and have the same pattern highlighted above. This indicates that our results are not influenced when we control for 3-digit industry-specific productivity shocks that vary over time.

Another interesting issue refers to match-specific productivity. Although evidence has been reported that the worker-plant match can influence estimates in wage equations (see, e.g., Krishna *et al.* (2014) and Araújo and Paz (2014) for Brazil), studies of the external returns to education generally do not control for this effect. As highlighted by Woodcock

Table 4.5 Robustness checks

Dependent variable: individual hourly wage (in log)						
# of workers with college-or-more	FE + IV					
	(1)	(2)	(3)	(4)	(5)	(6)
0 to 1 km	5.59e-05***	3.12e-05***	6.60e-05***	6.78e-05***	4.28e-05***	2.43e-04***
	(6.66e-06)	(6.14e-06)	(6.35e-06)	(6.25e-06)	(5.62e-06)	(3.82e-05)
1 to 5 km	3.10e-05***	2.85e-05***	2.20e-05***	1.95e-05***	2.68e-05***	2.13e-04***
	(2.42e-06)	(3.64e-06)	(2.45e-06)	(2.45e-06)	(4.73e-06)	(2.58e-05)
5 to 10 km	7.76e-06***	1.43e-05***	1.23e-05***	1.26e-05***	6.86e-06***	7.42e-05***
	(1.45e-06)	(1.71e-06)	(1.48e-06)	(1.74e-06)	(1.81e-06)	(7.31e-06)
10 to 20 km	-9.29e-06***	-1.26e-05***	-1.07e-05***	-1.06e-05***	-5.60e-06***	-3.37e-05***
	(5.78e-07)	(6.18e-07)	(6.56e-07)	(6.96e-07)	(9.40e-07)	(3.44e-06)
Worker-level controls	Yes	Yes	Yes	Yes	Yes	Yes
Plant-level controls	Yes	Yes	Yes	Yes	Yes	Yes
2-digit × year FE	No	No	Yes	Yes	Yes	Yes
3-digit × year FE	Yes	Yes	No	No	No	No
Metrop. region FE	Yes	Yes	No	Yes	Yes	Yes
Worker-plant FE	No	Yes	No	No	Yes	Yes
Worker-Metrop. FE	No	No	Yes	No	No	No
City population	No	No	No	Yes	No	No
Low-schooling workers	No	No	No	No	Yes	Yes
Transport × year FE	No	No	No	No	No	Yes
Kleibergen-Paap rk F[a]	1,050.46	635.25	1,449.44	1,385.05	1,192.05	30.37
Kleibergen-Paap rk LM[b]	3,867.09	2,561.33	5,311.26	5,086.24	1,643.85	117.77
Observations	2,387,434	2,223,550	2,378,812	2,387,434	2,223,550	2,191,252

Notes: Due to computational limitations, we do not report the F test of excluded instruments in the first stage. Robust standard errors in parentheses. [a]: H_0 - weakly identified model. [b]: H_0 - under-identified model. Robust standard errors in parentheses. Significance level: *** $p < 0.01$. Source: Prepared by the author based on estimates.

(2015), in the absence of distortions, good workers will match with good firms, and if match-specific productivity is important in wage determination, the absence of this effect on wage regression can generate biased estimates of the returns to observable worker and firm characteristics. For example, in equation 4.9, firms' unobservable time invariant characteristics can influence our estimates. Firms self-selection of larger cities may arise if only the most productive firms survive in large urban centers. This additional concern can be addressed directly when we modify our wage equation to include firms' fixed effects by considering a wage specification similar to Abowd *et al.* (1999).[25] We check if our results change when we control for worker-plant matched fixed effects (or job-spell fixed effects). The results reported in column 2 show that, in terms of attenuation, there is no changes. All coefficients still remain strongly significant and larger at shorter distances, particularly in the first ring, attenuating up to 10 km. Note, however, that the coefficient of the first ring is less than those of the other models (e.g., the 0-1 km ring coefficient in column 1 is 2 times greater).

As discussed earlier, workers choose the city where their skills are most valued. We deal with this problem using the exogenous expansion of higher education in Brazil. Another

[25]Technical details are available in Appendix C.1.

way to test the robustness of our instruments is to include worker-metropolitan-region matching effects. That is, everything that is specific to a worker-metropolitan-region pair is absorbed by the fixed effect. In this case, the variation that comes from movers is absorbed and the identification is based on stayers and comes from changes in the number of college-educated workers in each ring over time (Moretti, 2004a). It is expected that the results will not be highly sensitive to the inclusion of worker-metropolitan-region matching effects. Otherwise, if the results are highly sensitive, doubts would be cast on the validity of our instruments. Conditional estimates of worker-metropolitan-region matching are reported in column 3. There is no evidence that unobserved individual ability and return to unobserved ability across cities affect the attenuation results.

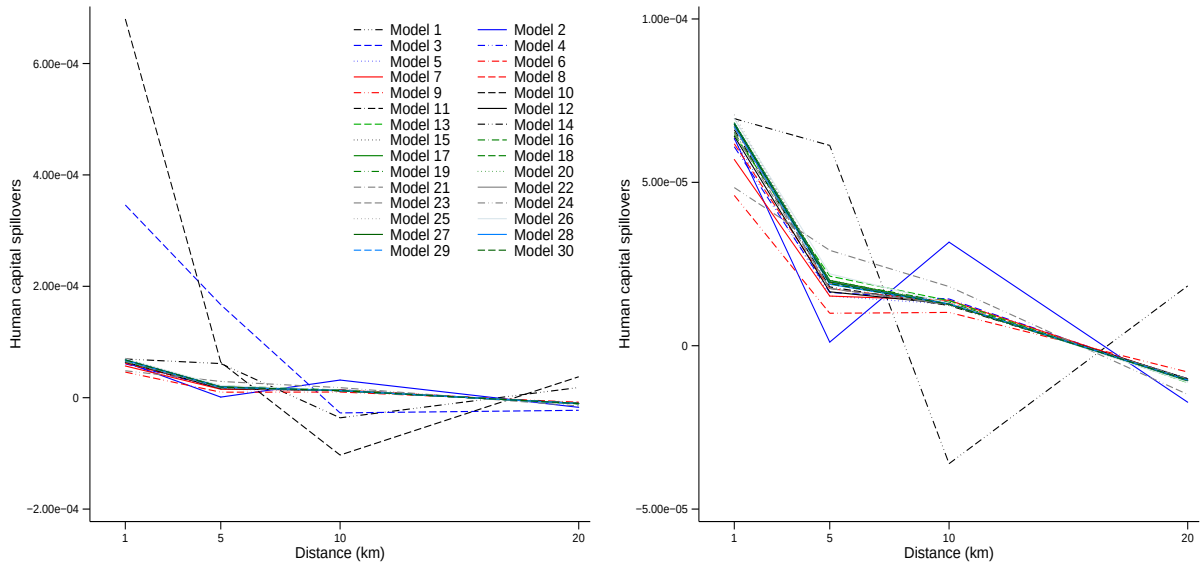
We also can evaluate what happens to the results when we include simultaneously other effects of agglomeration in each ring. The human capital variables are usually correlated with density (Combes and Gobillon, 2015). Therefore, other effects besides human capital spillovers can be captured by human capital variables when not controlling for the presence of other types of workers. To test the robustness of our results, we proceed in two ways: (i) in column 4 we include the city population as a control for these mechanisms, assuming that they act homogeneously within the same city; and (ii) in column 5 we include worker-plant matching fixed effects and the number of low-schooling workers in each ring to control for the presence of possible gains from density that vary with distance. Rosenthal and Strange (2008), for example, showed that the number of low-schooling workers has a negative effect on wages. Here we do not report the estimates (both for population and low-schooling workers) because these variables are potentially endogenous and therefore should be analyzed only as a robustness test. We conclude that other effects associated with the city size and concentration of low-schooling workers at different distances from the individual's workplace are not a major sources of bias. Our main results remain largely robust to these effects.

As mentioned earlier, the transportation infrastructure around establishments can affect worker productivity. Similar to what we did in our preliminary test in subsection 4.4.1, here we also include the same control variables for the transportation infrastructure around the individual's workplace (distance in kilometers from the cell's centroid to the nearest railroad, federal highway, state highway, airport and port). But unlike before, here we interact these time-invariant controls with the time effect to capture trends specific to each cell's transportation infrastructure improvement. In column 6, we include these control variables with worker-plant matching fixed effects and number of low-schooling workers effects. As can be seen, the estimated coefficients remain largely significant, and

more importantly, again show that the pattern of attenuation remains robust, as expected.

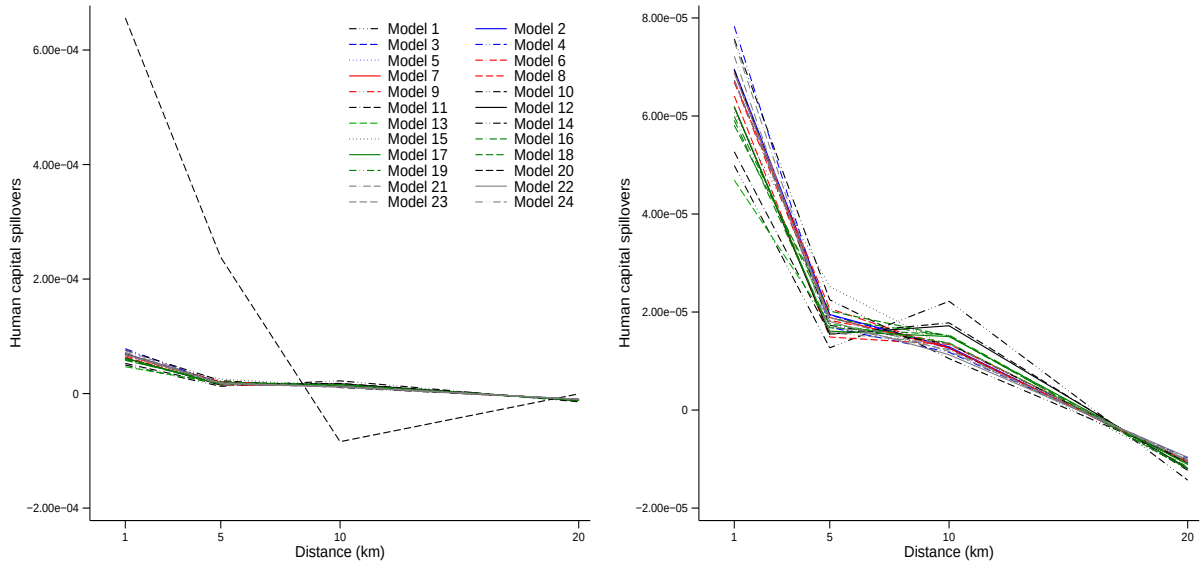
We also test the robustness of our main results using leave-one-out estimates by metropolitan region and by 2-digit CNAE code. In the first case, we exclude one metropolitan region at a time from the sample and estimate the model with the remaining metropolitan regions. This generates a total of 30 different models, reported in Table C.7 in Appendix C.2. To make the comparison of the models clearer, we plot in Figure 4.2 (a-b) these estimated human capital spillovers (vertical axis) as a function of distances between workers (horizontal axis). In Figure 4.2 (b), we exclude models 3 and 10, which have different scales. With few exceptions, the general and most important result is that the attenuation pattern of the coefficients remains largely unchanged. This indicates that our results are not extremely sensitive to the exclusion of any metropolitan region from the sample. Moreover, these results also reveal that we are not capturing a pattern specific to a particular metropolitan region, but a general pattern for all Brazilian metropolitan regions. Two clear exceptions to the general pattern are found in model 1 (in Figure 4.2 (b)), for which the estimates are obtained by excluding the São Paulo Metropolitan Region (SPMR) from the sample; and model 10 (in Figure 4.2 (a)), for which the estimates are obtained by excluding the Campinas Metropolitan Region (CAMR). In these two cases we observe an inflection in the sign of the estimated coefficient for the 5-10 km ring. This different behavior may be associated with (i) considerable reduction of the sample, e.g., the SPMR represents 38% of our sample; and (ii) some kind of complementarity between the two metropolitan regions, e.g., at some points the distance between SPMR and CAMR is 17.1 km.

Similarly to the 2-digit sectors, we exclude one sector at a time from the sample and estimate the model with the remaining sectors, generating the 24 models that are presented in Table C.8 in Appendix C.2. We also plot the estimated coefficients as a function of distances in Figure 4.3 (a-b). In Figure 4.3 (b), we exclude model 20, which has different scales. Again, our results are not sensitive to the exclusion of any specific sector. This indicates that we are not estimating an industry-specific spatial attenuation pattern, but a general pattern observed in all manufacturing activity. An exception, as can be seen in Figure 4.3 (a), is model 20, for which the estimates are obtained excluding the *motor vehicle manufacturing* (CNAE 29). Note, however, that the coefficient for the 5-10 km ring is negative and significant, but the coefficient for the 10-20 km ring is not significant, which is still consistent with a pattern of attenuation at short distances.



(a) All models (b) Without models 3 and 10

Figure 4.2 Leave-one-out estimates by metropolitan region



(a) All models (b) Without model 20

Figure 4.3 Leave-one-out estimates by 2-digit CNAE classification

4.4.5 External versus private returns to education

So far we have presented estimates for the external return to education. In this subsection we compare the external returns to education estimated in the previous subsections with the private returns to education. Following Rosenthal and Strange

(2008), we report the estimates for the private returns to education in Table 4.6, where we omit the agglomeration variables but retain all other controls.[26] This also allows us to compare the relationship between private and external returns to education in Brazil with those obtained by those authors for the US. To do this, in column 1 the estimates are obtained by OLS. Consistent with the literature on the private returns to education, the incremental contribution of a college degree beyond that of a high school diploma (complete high school) on an individual's wage is, on average, 56.26 percent, a larger private return than that obtained by the authors for the US (roughly 30 percent). However, when we control for the worker or worker-plant matched fixed effects (columns 2 and 3), the incremental gain is much smaller, roughly 6 and 4 percent, respectively. This is expected, given that the unobserved heterogeneity of workers and worker-plant matching omitted in the OLS estimates are biasing the private return upward.

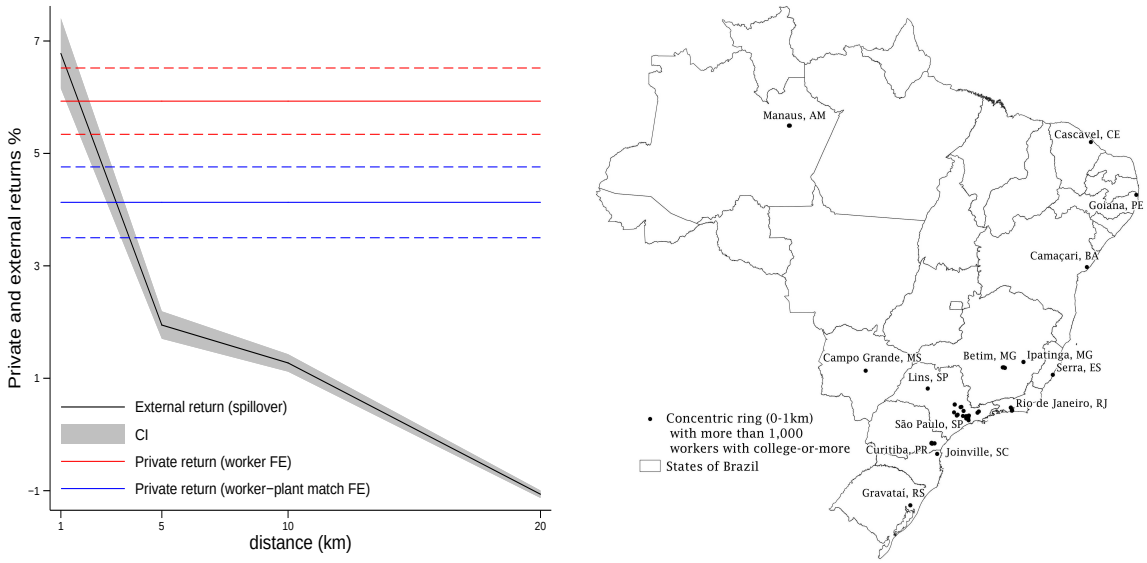
Table 4.6 Private returns to education

Dependent variable: individual hourly wage (in log)			
	OLS	Worker FE	Worker-plant FE
	(1)	(2)	(3)
Illiterate (reference category)			
Incomplete primary school	0.0608***	-0.0127**	-0.0095
	(0.0047)	(0.0057)	(0.0061)
Incomplete high school	0.1823***	-0.0142**	-0.0114*
	(0.0047)	(0.0057)	(0.0061)
Complete high school	0.3142***	-0.0160***	-0.0105*
	(0.0047)	(0.0058)	(0.0061)
Incomplete college	0.6366***	0.0042	0.0036
	(0.0050)	(0.0061)	(0.0065)
College degree or more	0.8772***	0.0433***	0.0308***
	(0.0050)	(0.0059)	(0.0063)
Worker-level controls	Yes	Yes	Yes
Plant-level controls	Yes	Yes	Yes
Industry × year FE	Yes	Yes	Yes
Metropolitan region FE	Yes	Yes	Yes
F-stat.	22,173.68	3,453.35	–
R squared	0.6389	0.3825	0.3622

Notes: This table presents the estimates obtained from equation 4.9 when we omit the agglomeration variables. All models are estimated with 2,387,434 observations. Worker-level controls include all the individual characteristics detailed above. Plant-level controls are dummies for plant size. Industry × year effects are dummies for each 2-digit × year combination. Metropolitan region FE is metropolitan region fixed effects. Standard errors adjusted for clustering are in parentheses. Significance levels: *** $p < 0.01$. Source: Prepared by the author based on estimates.

[26] When we include the agglomeration variables, as shown in Table C.9 in Appendix C.2, the results are very similar.

Two interesting patterns emerge from this evidence. First, as we have shown in the previous sections, adding 1,000 college-educated workers within 1 km would increase an individual's wage by roughly 6.8 to 24 percent, depending on the included controls (considering both the results in column 3 in Table 4.3 and those in Table 4.5). These effects are comparable to 12 to 43 percent of incremental private returns associated with obtaining a college degree in the OLS model. The equivalent percentage for the US is 20 to 50. That is, looking only at the OLS results, the external return measured as a share of the private returns to education is lower in Brazil. On the other hand, when we analyze the results of private returns to education conditional on the worker or worker-plant matching fixed effect, in most cases the external returns to education exceed the private returns, but only at short distances.



(a) Comparison (b) Concentric ring (0 to 1 km)location in 2014

Figure 4.4 External versus private returns

Consider the example in Figure 4.4 (a), where we plot the average percent change (with confidence interval - CI) in workers' wages given an increase of 1,000 college-educated workers in different distance bands (results of column 3 in Table 4.3). To compare external versus private returns, we also plot the average percentage change in wages associated with obtaining a college degree following high school in both worker FE (solid red line with dashed lines representing the CI) and worker-plant matching FE models (solid blue line with dashed lines representing the CI).[27] Note that for distances up to 1 km from the

[27]Figure C.3 in Appendix C.2 plot also the OLS results, which has different scales.

current establishment, human capital spillovers can outperform private returns, provided there is a certain increase (minimum 900) in the number of workers with college degree or higher. These results are particularly interesting in the context of public education policies and the efficiency of investments in education, because they suggest that the external return not only contributes to increasing the social returns to education, but also may be greater than the private return at short distances. We also report where these effects can occur in Figure 4.4 (b), which shows the location of 121 rings (0 to 1 km) in 2014 with 1,000 or more college-educated workers. Most of the rings are in (or near) the SPMR (73) and the remaining ones are located in industrial clusters in the other regions of the country (e.g., Campo Grande, MS; Camaçari, BA; Manaus, AM; and Joinville, SC). Moreover, these results are also in line with Moretti (2004a), who also found similar results with aggregated geographic data.

4.5 Concluding remarks

The objective of this paper has been to analyze the spatial extent of human capital spillovers within Brazilian cities. Some studies have explored this topic, but they are almost exclusively for developed countries, as cited throughout the text. Beyond the lack of evidence, the present examination contributes to the discussion on the subject by evaluating this phenomenon in an economic environment very different from that of developed countries. For this purpose, we have exogenously divided the all Brazilian geographic areas into cells of one square kilometer and used a unique and rich microgeographic panel dataset to calculate the number of college-educated workers in four different distance bands from the geographic centroid of each cell. In addition to using panel data and a broad set of controls for observed and unobserved heterogeneities, our identification strategy is based on a shift-share IV for the federal government's education policy shocks in Brazil in the period 1991-2004.

In the preliminary stage of the analysis, we used soil characteristics as instrument for economic mass. After including detailed controls for observable characteristics of workers and firms, industry and metropolitan region fixed effects, our main considerations are that the externalities generated by the concentration of workers in different distance bands from the individual's workplace are highly localized. The effects are stronger at short distances (up to 1 km). This decay pattern is faster than that observed for developed countries. A possible explanation is the urban structure, e.g., if the local transportation

system is not sufficiently developed to enable the collaboration of workers located in the farthest distance ranges, the agglomeration economies can be restricted to short distances (Rosenthal and Strange, 2020; Li *et al.*, 2020a).

The main set of results provides more detailed evidence when our focus is on the externalities generated by the concentration of college-educated workers. The proposal to isolate this specific effect aims to assess how the external return to education is attenuated with the distance from the current establishment. We used the exogenous expansion of public education in Brazil over the past two decades as an instrument for the number of college-educated workers in each of the rings, besides a broad set of controls including unobservable characteristics of workers, plants, industry, metropolitan region and worker-plant matching to deal with potential endogeneity. The external returns from education are also highly localized and therefore consistent with the idea that interaction between workers (face-to-face) can generate productivity gains from knowledge spillovers. We also found evidence that unskilled workers can obtain higher returns by being spatially close to skilled workers, in line with demand and supply models with spillover (Moretti, 2004a).

The evidence provided here is very consistent with the characteristics of the economic environment in Brazil. We can highlight some of these characteristics with which our results conform very well. The highly localized human capital spillovers we find are consistent with the high geographic concentration of the manufacturing. It is also consistent with the positive correlation of concentration $\times$ share of college-educated workers we found in chapter 2. The larger effect at very short distances also conforms very well with the absence of restriction on worker mobility, which favors the formation of highly specialized and dynamic local labor markets; and with the low quality of urban infrastructure (e.g., public transportation), which can hinder interaction. Our results also fully agree with the evidence presented in chapter 3 on the pattern of attenuation observed for localization economies.

The results presented here are clearly in line with the consensus that urban activities involve increasing returns and hence become more efficient than can be attained in isolation. But it also provides insight on the spatial pattern of the effects within cities in a neighborhood context. The importance is clear of a better understanding of these forces for public policymaking. Urban infrastructure can play an important role, for example, by providing the inputs for collaboration among more distant workers. In developing countries like Brazil, this importance is even greater, since the cities in general have greater structural problems.

References

ABOWD, J. M.; KRAMARZ, F.; MARGOLIS, D. N. High wage workers and high wage firms. *Econometrica*, Wiley Online Library, v. 67, n. 2, p. 251–333, 1999. 4.4.4

ACEMOGLU, D.; ANGRIST, J. How large are human-capital externalities? evidence from compulsory schooling laws. *NBER Macroeconomics Annual*, MIT Press, v. 15, p. 9–59, 2000. <https://doi.org/10.1086/654403>. 4.2

ADAO, R.; KOLESÁR, M.; MORALES, E. Shift-share designs: Theory and inference. *The Quarterly Journal of Economics*, Oxford University Press, v. 134, n. 4, p. 1949–2010, 2019. 4.3.2

ADDARIO, S. D.; PATACCHINI, E. Wages and the city. evidence from Italy. *Labour Economics*, Elsevier, v. 15, n. 5, p. 1040–1061, 2008. 3.5.2, 4.1, 4.4.1

AFONSO, J. R. R.; ARAÚJO, E. A.; JÚNIOR, G. B. Fiscal space and public sector investments in infrastructure: a Brazilian case-study. Instituto de Pesquisa Econômica Aplicada (Ipea), 2005. 2.4

ALEKSANDROVA, E.; BEHRENS, K.; KUZNETSOVA, M. Manufacturing (co) agglomeration in a transition country: Evidence from Russia. *Journal of Regional Science*, Wiley Online Library, 2019. <https://doi.org/10.1111/jors.12436>. 1, 2.1, 2.2.1, 2.3.1, 2.3.2, 2.3.3, 2.4, 2.5, A.2, A.3

ALMEIDA, E. T.; ROCHA, R. M. Labor pooling as an agglomeration factor: Evidence from the Brazilian Northeast in the 2002–2014 period. *EconomiA*, Elsevier, 2018. <https://doi.org/10.1016/j.econ.2018.02.002>. 2.3.3, 2.4

ALMEIDA, E. T. d.; NETO, R. d. M. S.; BASTOS, J. M. B. de; SILVA, R. L. P. da. Location patterns of service activities in large metropolitan areas: the case of São Paulo. *The Annals of Regional Science*, Springer, p. 1–31, 2021. 19

ANAS, A.; ARNOTT, R.; SMALL, K. A. Urban spatial structure. *Journal of Economic Literature*, JSTOR, v. 36, n. 3, p. 1426–1464, 1998. 4.3.5

ANDERSON, N. B.; BOGART, W. T. The structure of sprawl: Identifying and characterizing employment centers in polycentric metropolitan areas. *American Journal of Economics and Sociology*, Wiley Online Library, v. 60, n. 1, p. 147–169, 2001. 4.3.5

ANDERSSON, M.; KLAESSON, J.; LARSSON, J. P. How local are spatial density externalities? Neighbourhood effects in agglomeration economies. *Regional Studies*, Routledge, v. 50, n. 6, p. 1082–1095, 2014. 15, 7, 20

ANDERSSON, M.; LARSSON, J. P.; WERNBERG, J. The economic microgeography of diversity and specialization externalities-firm-level evidence from Swedish cities. *Research Policy*, Elsevier, v. 48, n. 6, p. 1385–1398, 2019. 3.1, 15, 7, 20

ANDERSSON, R.; QUIGLEY, J. M.; WILHELMSSON, M. Urbanization, productivity, and innovation: Evidence from investment in higher education. *Journal of Urban Economics*, Elsevier, v. 66, n. 1, p. 2–15, 2009. 3.3, 4.1, 4.4.2

ANDREWS, M.; SCHANK, T.; UPWARD, R. Practical fixed-effects estimation methods for the three-way error-components model. *The Stata Journal*, SAGE Publications Sage CA: Los Angeles, CA, v. 6, n. 4, p. 461–481, 2006. C.1.2

ARAÚJO, B. C.; PAZ, L. S. The effects of exporting on wages: An evaluation using the 1999 Brazilian exchange rate devaluation. *Journal of Development Economics*, Elsevier, v. 111, p. 1–16, 2014. 4.4.4

ARAUZO-CAROD, J.-M.; LIVIANO-SOLIS, D.; MANJÓN-ANTOLÍN, M. Empirical studies in industrial location: an assessment of their methods and results. *Journal of Regional Science*, Wiley Online Library, v. 50, n. 3, p. 685–711, 2010. 3.1, 3.4.1

AUTOR, D. H.; DORN, D.; HANSON, G. H. The China syndrome: Local labor market effects of import competition in the United States. *American Economic Review*, v. 103, n. 6, p. 2121–68, 2013. 4.3.2

AZZONI, C. R. *Indústria e reversão da polarização no Brasil*. [S.l.]: Instituto de Pesquisas Econômicas, 1986. v. 58. 2.3.2

BAER, W. *Economia Brasileira*. [S.l.]: NBL Editora, 2002. 1, 2.1

BALDWIN, R. E.; KRUGMAN, P. Agglomeration, integration and tax harmonisation. *European Economic Review*, Elsevier, v. 48, n. 1, p. 1–23, 2004. 2.4

BARBOSA, F.; ARAÚJO, H. E.; ARAÚJO, M. Migração interna no Brasil. *Comunicados do IPEA*, Instituto de Pesquisa Econômica Aplicada (IPEA), n. 61, 2010. <http://repositorio.ipea.gov.br/handle/11058/5289>. 2.4

BARLET, M.; BRIANT, A.; CRUSSON, L. Concentration géographique dans lâindustrie manufacturière et dans les services en France: une approche par un indicateur en continu. *Documents de Travail de la DESE-Working Papers of the DESE*, Institut National de la Statistique et des Etudes Economiques, D3E, 2008. 2.5

BARROS, R. P. d.; FRANCO, S.; MENDONÇA, R. Discriminação e segmentação no mercado de trabalho e desigualdade de renda no Brasil. *Texto para Discussão, Instituto de*

Pesquisa Econômica Aplicada (IPEA), Instituto de Pesquisa Econômica Aplicada (Ipea), 2007. <http://repositorio.ipea.gov.br/handle/11058/1842>. 4.1

BARROS, R. P. d.; FRANCO, S.; MENDONÇA, R. A recente queda da desigualdade de renda e o acelerado progresso educacional brasileiro da última década. *Texto para Discussão, Instituto de Pesquisa Econômica Aplicada (IPEA)*, Instituto de Pesquisa Econômica Aplicada (Ipea), 2007. <http://repositorio.ipea.gov.br/handle/11058/1439>. 4.1

BARTHOLOMEU, D. B.; CAIXETA FILHO, J. V. Impactos econômicos e ambientais decorrentes do estado de conservação das rodovias brasileiras: um estudo de caso. *Revista de Economia e Sociología Rural*, SciELO Brasil, v. 46, n. 3, p. 703–738, 2008. 2.4

BARTIK, T. J. Who benefits from state and local economic development policies? WE Upjohn Institute for Employment Research, 1991. 3.4.3, 4.3.2

BARUFI, A. M. B.; HADDAD, E. A.; NIJKAMP, P. Industrial scope of agglomeration economies in Brazil. *The Annals of Regional Science*, Springer, v. 56, n. 3, p. 707–755, 2016. <https://doi.org/10.1007/s00168-016-0768-3>. 1, 3.1, 4.1, 4.3.1

BAUM, C. F.; SCHAFFER, M. E.; STILLMAN, S. Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *The Stata Journal*, SAGE Publications Sage CA: Los Angeles, CA, v. 7, n. 4, p. 465–506, 2007. 4.4.1

BEHRENS, K.; BOUGNA, T. An anatomy of the geographical concentration of Canadian manufacturing industries. *Regional Science and Urban Economics*, Elsevier, v. 51, p. 47–69, 2015. <https://doi.org/10.1016/j.regsciurbeco.2015.01.002>. 1, 2.1, 2.3.1, 2.3.1, 15, 2.3.2, 2.5, A.2, A.3, A.3

BEHRENS, K.; BROWN, W. M. Transport costs, trade, and geographic concentration: Evidence from Canada. In: *Handbook of international trade and transportation*. [S.l.]: Edward Elgar Publishing, 2018. 2.4, 2.4

BEHRENS, K.; BROWN, W. M.; BOUGNA, T. The world is not yet flat: transport costs matter! *Review of Economics and Statistics*, MIT Press, v. 100, n. 4, p. 712–724, 2018. 2.4, 2.4

BILLINGS, S. B.; JOHNSON, E. B. A non-parametric test for industrial specialization. *Journal of Urban Economics*, Elsevier, v. 71, n. 3, p. 312–331, 2012. 4.3.5

BILLINGS, S. B.; JOHNSON, E. B. Agglomeration within an urban area. *Journal of Urban Economics*, Elsevier, v. 91, p. 13–25, 2016. <https://doi.org/10.1016/j.jue.2015.11.002>. 2.4

BLANCHARD, O. J.; KATZ, L. F.; HALL, R. E.; EICHENGREEN, B. Regional evolutions. *Brookings papers on economic activity*, JSTOR, v. 1992, n. 1, p. 1–75, 1992. 4.3.2

BORUSYAK, K.; HULL, P.; JARAVEL, X. *Quasi-experimental shift-share research designs*. [S.l.], 2018. 4.3.2

BRAKMAN, S.; GARRETSEN, H.; ZHAO, Z. Spatial concentration of manufacturing firms in China. *Papers in Regional Science*, Wiley Online Library, v. 96, p. S179–S205, 2016. 1, 2.1, 2.3.1, 2.3.2, 2.3.2, 2.3.3, 2.5, A.3

BRIANT, A.; COMBES, P.-P.; LAFOURCADE, M. Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations? *Journal of Urban Economics*, Elsevier, v. 67, n. 3, p. 287–302, 2010. <https://doi.org/10.1016/j.jue.2009.09.014>. 4.1

CAMERON, A. C.; TRIVEDI, P. K. *Regression analysis of count data*. [S.l.]: Cambridge university press, 2013. v. 53. 19, 3.4.3, 3.4.3, 21

CARLINO, G.; KERR, W. R. Agglomeration and innovation. In: *Handbook of Regional and Urban Economics*. [S.l.]: Elsevier, 2015. v. 5, p. 349–404. 4.4.2

CARLTON, D. W. The location and employment choices of new firms: An econometric model with discrete and continuous endogenous variables. *The Review of Economics and Statistics*, JSTOR, p. 440–449, 1983. 3.1, 3.4.1

CAVALCANTE, L. R. Classificações tecnológicas: uma sistematização. Instituto de Pesquisa Econômica Aplicada (Ipea), 2014. 2.1, 9, 4.4.4, A.6

CHANDRA, A.; THOMPSON, E. Does public infrastructure affect economic activity?: Evidence from the rural interstate highway system. *Regional Science and Urban Economics*, Elsevier, v. 30, n. 4, p. 457–490, 2000. 2.4

CHARLOT, S.; DURANTON, G. Communication externalities in cities. *Journal of Urban Economics*, Elsevier, v. 56, n. 3, p. 581–613, 2004. <https://doi.org/10.1016/j.jue.2004.08.001>. 4.2

CHARLOT, S.; DURANTON, G. Cities and workplace communication: some quantitative French evidence. *Urban Studies*, Sage Publications Sage UK: London, England, v. 43, n. 8, p. 1365–1394, 2006. 4.2

CHATTERJI, A.; GLAESER, E.; KERR, W. Clusters of entrepreneurship and innovation. *Innovation Policy and the Economy*, University of Chicago Press Chicago, IL, v. 14, n. 1, p. 129–166, 2014. 3.4.2

CHAUVIN, J. P.; GLAESER, E.; MA, Y.; TOBIO, K. What is different about urbanization in rich and poor countries? cities in Brazil, China, India and the United States. *Journal of Urban Economics*, Elsevier, v. 98, p. 17–49, 2017. <https://doi.org/10.1016/j.jue.2016.05.003>. 1, 3.1, 4.1, 4.4.2

CHINITZ, B. Contrasts in agglomeration: New york and pittsburgh. *The American Economic Review*, JSTOR, p. 279–289, 1961. 3.4.2

CICCONE, A.; HALL, R. E. Productivity and the density of economic activity. *American Economic Review*, Citeseer, v. 86, n. 1, p. 54–70, 1996. <https://doi.org/10.3386/w4313>. 11, 2.4, 4.3.4

COFFEY, W. J.; SHEARMUR, R. G. Agglomeration and dispersion of high-order service employment in the Montreal metropolitan region, 1981-96. *Urban Studies*, Sage Publications Sage UK: London, England, v. 39, n. 3, p. 359–378, 2002. 4.3.5

COMBES, P.-P. Economic structure and local growth: France, 1984–1993. *Journal of Urban Economics*, Elsevier, v. 47, n. 3, p. 329–355, 2000. <https://doi.org/10.1006/juec.1999.2143>. 2.4, 2.4

COMBES, P.-P.; DÉMURGER, S.; LI, S. Urbanisation and migration externalities in China. CEPR Discussion Paper No. DP9352, 2013. 1, 4.1

COMBES, P.-P.; DÉMURGER, S.; LI, S.; WANG, J. Unequal migration and urbanisation gains in China. *Journal of Development Economics*, Elsevier, v. 142, p. 102328, 2020. 1, 4.1

COMBES, P.-P.; DURANTON, G. Labour pooling, labour poaching, and spatial clustering. *Regional Science and Urban Economics*, Elsevier, v. 36, n. 1, p. 1–28, 2006. 2.4

COMBES, P.-P.; DURANTON, G.; GOBILLON, L. Spatial wage disparities: Sorting matters! *Journal of Urban Economics*, Elsevier, v. 63, n. 2, p. 723–742, 2008. <https://doi.org/10.1016/j.jue.2007.04.004>. 3.4.3, 4.3.4

COMBES, P.-P.; DURANTON, G.; GOBILLON, L. The identification of agglomeration economies. *Journal of Economic Geography*, Oxford University Press, v. 11, n. 2, p. 253–266, 2011. <https://doi.org/10.1093/jeg/lbq038>. 4.3.4

COMBES, P.-P.; DURANTON, G.; GOBILLON, L.; ROUX, S. Estimating agglomeration economies with history, geology, and worker effects. In: *Agglomeration Economics*. [S.l.]: University of Chicago Press, 2010. p. 15–66. <https://ssrn.com/abstract=1141634>. 4.3.4

COMBES, P.-P.; GOBILLON, L. The empirics of agglomeration economies. In: *Handbook of Regional and Urban Economics*. [S.l.]: Elsevier, 2015. v. 5, p. 247–348. <https://doi.org/10.1016/B978-0-444-59517-1.00005-2>. 3.3, 4.1, 2, 4.4.4

CORREIA, S.; GUIMARÃES, P.; ZYLKIN, T. Fast Poisson estimation with high-dimensional fixed effects. *The Stata Journal*, SAGE Publications Sage CA: Los Angeles, CA, v. 20, n. 1, p. 95–115, 2020. 3.4.1

COSTA, A. B.; BIDERMAN, C. A dinâmica da concentração do emprego industrial no Brasil (1991-2011) e o ciclo de vida das empresas. *ANPEC-Associação Nacional dos Centros de Pósgraduação em Economia [Brazilian Association of Graduate Programs in Economics]*, 2016. 2.3.2

DEKLE, R.; EATON, J. Agglomeration and land rents: evidence from the prefectures. *Journal of Urban Economics*, Elsevier, v. 46, n. 2, p. 200–214, 1999. 2.4

DINGEL, J. I.; MISCIO, A.; DAVIS, D. R. *Cities, lights, and skills in developing economies*. [S.l.], 2019. <http://www.nber.org/papers/w25678>. 11, 2.4

DURANTON, G. Growing through cities in developing countries. *World Bank Research Observer*, World Bank Research Observer, v. 39, n. 1, p. 39â73, 2015. 3.1

DURANTON, G. Agglomeration effects in Colombia. *Journal of Regional Science*, Wiley Online Library, v. 56, n. 2, p. 210–238, 2016. <https://doi.org/10.1111/jors.12239>. 1, 4.1

DURANTON, G. Determinants of city growth in Colombia. *Papers in Regional Science*, Wiley Online Library, v. 95, n. 1, p. 101–131, 2016. <https://doi.org/10.1111/pirs.12225>. 11, 2.4

DURANTON, G.; MORROW, P. M.; TURNER, M. A. Roads and trade: Evidence from the US. *Review of Economic Studies*, Oxford University Press, v. 81, n. 2, p. 681–724, 2014. 2.4

DURANTON, G.; OVERMAN, H. G. Testing for localization using micro-geographic data. *The Review of Economic Studies*, Wiley-Blackwell, v. 72, n. 4, p. 1077–1106, 2005. <https://doi.org/10.1111/0034-6527.00362>. 1, 2.1, 2.2.1, 2.3.1, 14, 15, 2.3.2, 2.5, 3.1, 4, 3.2.2, 3.6, A.2, A.2, 2, A.3, B.2, B.2.1, B.2.1

DURANTON, G.; OVERMAN, H. G. Exploring the detailed location patterns of UK manufacturing industries using microgeographic data. *Journal of Regional Science*, Wiley Online Library, v. 48, n. 1, p. 213–243, 2008. <https://doi.org/10.1111/j.1365-2966.2006.0547.x>. 1, 2.1, 3.1, 4, 3.2.2, 7, 3.2.2, 10, 3.2.2, 12, 3.6, B.2, B.2.1, B.2.1, B.2.1

DURANTON, G.; PUGA, D. Micro-foundations of urban agglomeration economies. In: *Handbook of Regional and Urban Economics*. [S.l.]: Elsevier, 2004. v. 4, p. 2063–2117. 2.4, 2.4, 4.1

DURANTON, G.; PUGA, D. The growth of cities. In: *Handbook of Economic Growth*. [S.l.]: Elsevier, 2014. v. 2, p. 781–853. <https://doi.org/10.1016/B978-0-444-53540-5.00005-7>. 1

ELLISON, G.; GLAESER, E. L. Geographic concentration in us manufacturing industries: a dartboard approach. *Journal of Political Economy*, The University of Chicago Press, v. 105, n. 5, p. 889–927, 1997. 3.2.2

ELLISON, G.; GLAESER, E. L. The geographic concentration of industry: does natural advantage explain agglomeration? *American Economic Review*, v. 89, n. 2, p. 311–316, 1999. 2.4, 3.4.2, 4.4.1

ELLISON, G.; GLAESER, E. L.; KERR, W. R. What causes industry agglomeration? evidence from coagglomeration patterns. *American Economic Review*, v. 100, n. 3, p. 1195–1213, 2010. 2.4, 2.4, 2.4, 3.4.2, 4.4.1

FABERMAN, R. J. The relationship between the establishment age distribution and urban growth. *Journal of Regional Science*, Wiley Online Library, v. 51, n. 3, p. 450–470, 2011. 6

FERREIRA, A. L.; DINIZ, M. B. *et al.* Determinantes da concentração geográfica industrial no Brasil. *Revista Econômica do Nordeste*, v. 50, n. 4, p. 163–182, 2019. 2.1, 2.4

FIGUEIREDO, O.; GUIMARAES, P.; WOODWARD, D. Home-field advantage: location decisions of Portuguese entrepreneurs. *Journal of Urban Economics*, Elsevier, v. 52, n. 2, p. 341–361, 2002. 3.4.1

FISHLOW, A. Brazilian size distribution of income. *The American Economic Review*, JSTOR, v. 62, n. 1/2, p. 391–402, 1972. 1, 4.1

FRITSCH, M.; WYRWICH, M. The effect of entrepreneurship on economic development – an empirical analysis using regional entrepreneurship culture. *Journal of Economic Geography*, Oxford University Press, v. 17, n. 1, p. 157–189, 2016. 6

FU, S. Smart café cities: Testing human capital externalities in the Boston metropolitan area. *Journal of Urban Economics*, Elsevier, v. 61, n. 1, p. 86–111, 2007. <https://doi.org/10.1016/j.jue.2006.06.002>. 11, 2.4, 3.1, 3.3, 4.1, 4.2, 4.4.2, 4.4.2

FUJITA, M.; KRUGMAN, P. R.; VENABLES, A. *The spatial economy: Cities, regions, and international trade*. [S.l.]: MIT press, 1999. 2.4

FURTADO, C. *Formação econômica do Brasil*. [S.l.]: Editora Universidade de Brasília, 1963. 1, 2.1

GHANBARIAMIN, R.; CHUNG, B. W. The effect of the national kidney registry on the kidney-exchange market. *Journal of Health Economics*, Elsevier, v. 70, p. 102301, 2020. 3.4.3

GHANI, E.; KERR, W. R.; O'CONNELL, S. Spatial determinants of entrepreneurship in India. *Regional Studies*, Taylor & Francis, v. 48, n. 6, p. 1071–1089, 2014. 3.4.2

GIBBONS, S.; LYYTIKÄINEN, T.; OVERMAN, H. G.; SANCHIS-GUARNER, R. New road infrastructure: the effects on firms. *Journal of Urban Economics*, Elsevier, v. 110, p. 35–50, 2019. 3.4.2, 4.4.1

GLAESER, E.; HENDERSON, J. V. Urban economics for the developing world: An introduction. *Journal of Urban Economics*, Elsevier, v. 98, p. 1–5, 2017. 4.1

GLAESER, E. L. *Entrepreneurship and the City*. [S.l.], 2007. 1

GLAESER, E. L.; KAHN, M. E. *Decentralized employment and the transformation of the American city*. [S.l.], 2001. 2.4

GLAESER, E. L.; KALLAL, H. D.; SCHEINKMAN, J. A.; SHLEIFER, A. Growth in cities. *Journal of Political Economy*, The University of Chicago Press, v. 100, n. 6, p. 1126–1152, 1992. 2.4, 2.4

GLAESER, E. L.; KERR, S. P.; KERR, W. R. Entrepreneurship and urban growth: An empirical assessment with historical mines. *Review of Economics and Statistics*, MIT Press, v. 97, n. 2, p. 498–520, 2015. 6

GLAESER, E. L.; KERR, W. R. Local industrial conditions and entrepreneurship: how much of the spatial distribution can we explain? *Journal of Economics & Management Strategy*, Wiley Online Library, v. 18, n. 3, p. 623–663, 2009. 3.1, 1, 3.4.2

GLAESER, E. L.; KERR, W. R.; PONZETTO, G. A. Clusters of entrepreneurship. *Journal of Urban Economics*, Elsevier, v. 67, n. 1, p. 150–168, 2010. 1

GLAESER, E. L.; ROSENTHAL, S. S.; STRANGE, W. C. Urban economics and entrepreneurship. *Journal of Urban Economics*, Elsevier, v. 67, n. 1, p. 1–14, 2010. 3.4.2

GOLDSMITH-PINKHAM, P.; SORKIN, I.; SWIFT, H. Bartik instruments: What, when, why, and how. *American Economic Review*, v. 110, n. 8, p. 2586–2624, 2020. 4.3.2

GREENSTONE, M.; HORNBECK, R.; MORETTI, E. Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings. *Journal of Political Economy*, The University of Chicago Press, v. 118, n. 3, p. 536–598, 2010. 2.4, 2.4

GUILHOTO, J.; SESSO FILHO, U. A. Estimação da matriz insumo-produto à partir de dados preliminares das contas nacionais, 2004. *Economia Aplicada*, v. 9, n. 2, p. 277–299, 2005. 2.4, A.3

GUILHOTO, J.; SESSO FILHO, U. A. Estimação da matriz insumo-produto utilizando dados preliminares das contas nacionais: Aplicação e análise de indicadores econômicos para o Brasil em 2005 (using data from the system of national accounts to estimate input-output matrices: An application using Brazilian data for 2005). *Available at SSRN 1836495*, 2010. 2.4, A.3

HÅKANSSON, J.; ISACSSON, G. The spatial extent of agglomeration economies across the wage earnings distribution. *Journal of Regional Science*, Wiley Online Library, v. 59, n. 2, p. 281–301, 2019. 3.1, 3.5.2, 4.1, 4.3.1, 4.4.1, 4.4.1, 4.4.2

HANSEN, E. R. Industrial location choice in Sao Paulo, Brazil: a nested logit model. *Regional Science and Urban Economics*, North-Holland, v. 17, n. 1, p. 89–108, 1987. 3.1

HATZICHRONOGLOU, T. Revision of the high-technology sector and product classification. OECD, 1997. 9

HEAD, K.; RIES, J.; SWENSON, D. Agglomeration benefits and location choice: Evidence from Japanese manufacturing investments in the United States. *Journal of International Economics*, Elsevier, v. 38, n. 3-4, p. 223–247, 1995. 3.1

HENDERSON, J. V. Efficiency of resource usage and city size. *Journal of Urban Economics*, Elsevier, v. 19, n. 1, p. 47–70, 1986. 3.5.1

HOFFMANN, R. Queda da desigualdade da distribuição de renda no Brasil, de 1995 a 2005, e delimitação dos relativamente ricos em 2005. *Desigualdade de renda no Brasil: uma análise da queda recente*, v. 1, p. 93–105, 2006. 2.1

HOLL, A. Manufacturing location and impacts of road transport infrastructure: empirical evidence from Spain. *Regional Science and Urban Economics*, Elsevier, v. 34, n. 3, p. 341–363, 2004. 3.4.2

HOLL, A. Transport infrastructure, agglomeration economies, and firm birth: empirical evidence from Portugal. *Journal of Regional Science*, Wiley Online Library, v. 44, n. 4, p. 693–712, 2004. 3.4.2

HOLL, A. Highways and productivity in manufacturing firms. *Journal of Urban Economics*, Elsevier, v. 93, p. 131–151, 2016. 3.4.2, 4.4.1

HOLMES, T. J. Localization of industry and vertical disintegration. *Review of Economics and Statistics*, MIT Press, v. 81, n. 2, p. 314–325, 1999. 2.4

JOFRE-MONSENY, J. The scope of agglomeration economies: Evidence from Catalonia. *Papers in Regional Science*, Wiley Online Library, v. 88, n. 3, p. 575–590, 2009. 18

JOFRE-MONSENY, J.; MARÍN-LÓPEZ, R.; VILADECANS-MARSAL, E. The determinants of localization and urbanization economies: Evidence from the location of new firms in Spain. *Journal of Regional Science*, Wiley Online Library, v. 54, n. 2, p. 313–337, 2014. 3.4.1

KLEIBERGEN, F.; PAAP, R. Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, Elsevier, v. 133, n. 1, p. 97–126, 2006. 4.4.1

KLIER, T.; MCMILLEN, D. P. Evolving agglomeration in the US auto supplier industry. *Journal of Regional Science*, Wiley Online Library, v. 48, n. 1, p. 245–267, 2008. 4, 3.2.2

KOH, H.-J.; RIEDEL, N. Assessing the localization pattern of German manufacturing and service industries: a distance-based approach. *Regional Studies*, Routledge, v. 48, n. 5, p. 823–843, 2014. 2.1, 2.3.2, 2.5, A.3

KOLKO, J. Urbanization, agglomeration, and coagglomeration of service industries. In: *Agglomeration Economics*. [S.l.]: University of Chicago Press, 2010. p. 151–180. 2.4

KRISHNA, P.; POOLE, J. P.; SENSES, M. Z. Wage effects of trade reform with endogenous worker mobility. *Journal of International Economics*, Elsevier, v. 93, n. 2, p. 239–252, 2014. 4.4.4

KRUGMAN, P. R. *Geography and trade*. [S.l.]: MIT press, 1991. 2.4

KRUGMAN, P. R. Increasing returns and economic geography. *Journal of Political Economy*, The University of Chicago Press, v. 99, n. 3, p. 483–499, 1991. 2.4, 2.4

LARSSON, J. P. The neighborhood or the region? reassessing the density–wage relationship using geocoded data. *The Annals of Regional Science*, Springer, v. 52, n. 2, p. 367–384, 2014. 15, 7

LAUTERT, V.; ARAÚJO, N. C. M. d. Concentração industrial no Brasil no período 1996-2001: uma análise por meio do índice de Ellison e Glaeser (1994). *Economia Aplicada*, SciELO Brasil, v. 11, n. 3, p. 347–368, 2007. <http://dx.doi.org/10.1590/S1413-80502007000300002>. 2.1, 2.3.2, 2.3.2, 2.3.3, 3.2.1

LAZUKA, V. The long-term health benefits of receiving treatment from qualified midwives at birth. *Journal of Development Economics*, Elsevier, v. 133, p. 415–433, 2018. 3.4.3

LEFF, N. H. Desenvolvimento econômico e desigualdade regional: origens do caso brasileiro. *Revista Brasileira de Economia*, v. 26, n. 1, p. 3–22, 1972. 1, 2.1, 3.1

LI, J.; LI, L.; LIU, S. Attenuation of agglomeration economies: Evidence from the universe of Chinese manufacturing firms. *Working paper*, 2020. 1, 3.1, 15, 3.4.1, 3.4.2, 3.4.3, 3.5.1, 3.5.2, 3.5.3, 3.5.3, 23, 3.6, 4.1, 7, 4.4.1, 4.5

LI, J.; LIU, S.; WU, Y. Identifying knowledge spillovers from universities: Quasi-experimental evidence from urban China. *Working paper*, 2020. 4.4.2

LIMA, A. C. d. C.; SIMÕES, R. F. Centralidade e emprego na região Nordeste do Brasil no período 1995/2007. *Nova Economia*, SciELO Brasil, v. 20, n. 1, p. 39–83, 2010. 2.3.2

LIU, Z. Human capital externalities in cities: evidence from Chinese manufacturing firms. *Journal of Economic Geography*, Oxford University Press, v. 14, n. 3, p. 621–649, 2013. 24

LOCHNER, L.; MORETTI, E. The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review*, v. 94, n. 1, p. 155–189, 2004. 4.4.2

LYCHAGIN, S.; PINKSE, J.; SLADE, M. E.; REENEN, J. V. Spillovers in space: Does geography matter? *The Journal of Industrial Economics*, Wiley Online Library, v. 64, n. 2, p. 295–335, 2016. 2.4

MARCON, E.; PUECH, F. Evaluating the geographic concentration of industries using distance-based methods. *Journal of Economic Geography*, Oxford University Press, v. 3, n. 4, p. 409–428, 2003. <https://doi.org/10.1093/jeg/lbg016>. 1

MARCON, E.; PUECH, F. Measures of the geographic concentration of industries: improving distance-based methods. *Journal of Economic Geography*, Oxford University Press, v. 10, n. 5, p. 745–762, 2009. <https://doi.org/10.1093/jeg/lbp056>. 1

MARSHALL, A. *Principles of Economics*. [S.l.]: Macmillan London, 1890. 1, 2.4, 2.4, 3.1, 3.3, 4.2

MAYER, T.; TREVIEN, C. The impact of urban public transportation evidence from the Paris region. *Journal of Urban Economics*, Elsevier, v. 102, p. 1–21, 2017. 3.4.2, 4.4.1

MCCANN, P. *Modern urban and regional economics*. [S.l.]: Oxford University Press, 2013. 2.4

MENDONÇA, R. S. P. de; BARROS, R. P. de. A evolução do bem-estar e da desigualdade no Brasil desde 1960. *Revista Brasileira de Economia*, v. 49, n. 2, p. 329–352, 1995. 1, 4.1

MILLIGAN, K.; MORETTI, E.; OREOPOULOS, P. Does education improve citizenship? evidence from the United States and the United Kingdom. *Journal of Public Economics*, Elsevier, v. 88, n. 9-10, p. 1667–1695, 2004. 4.4.2

MORETTI, E. Estimating the external return to higher education: Evidence from cross-sectional and longitudinal data. *Journal of Econometrics*, v. 120, n. 1-2, p. 175–212, 2004. 1, 11, 2.4, 2.4, 4.2, 4.2, 4.2, 4.3.3, 14, 4.4.2, 4.4.2, 4.4.3, 4.4.4, 4.4.5, 4.5

MORETTI, E. Human capital externalities in cities. In: *Handbook of Regional and Urban Economics*. [S.l.]: Elsevier, 2004. v. 4, p. 2243–2291. 4.4.2, 4.4.2

MORETTI, E. Workers' education, spillovers, and productivity: evidence from plant-level production functions. *American Economic Review*, v. 94, n. 3, p. 656–690, 2004. 11, 2.4, 2.4, 4.3.1, 24

MORETTI, E. Local multipliers. *American Economic Review*, v. 100, n. 2, p. 373–77, 2010. 3.4.3

MORETTI, E.; THULIN, P. Local multipliers and human capital in the United States and Sweden. *Industrial and Corporate Change*, Oxford University Press, v. 22, n. 1, p. 339–362, 2013. 3.4.3

NAKAJIMA, K.; SAITO, Y. U.; UESUGI, I. Measuring economic localization: Evidence from Japanese firm-level data. *Journal of the Japanese and International Economies*, Elsevier, v. 26, n. 2, p. 201–220, 2012. <https://doi.org/10.1016/j.jjie.2012.02.002>. 1, 2.1, 2.3.1, 2.3.2, 2.3.2, A.3

NAKAMURA, R. Agglomeration economies in urban manufacturing industries: a case of Japanese cities. *Journal of Urban Economics*, Elsevier, v. 17, n. 1, p. 108–124, 1985. 3.5.1

NARITA, R.; GONZAGA, G.; FIRPO, S. Decomposicão da evolucao da desigualdade de renda no Brasil em efeitos idade, periodo e coorte/explaining income inequality in Brazil: age, period and cohort effects. *Pesquisa e Planejamento Economico*, v. 33, n. 2, p. 211–252, 2003. <http://ppe.ipea.gov.br/index.php/ppe/article/view/91>. 1, 4.1

NASCIMENTO, S. P. d. Guerra fiscal: uma avaliação comparativa entre alguns estados participantes. *Economia Aplicada*, SciELO Brasil, v. 12, n. 4, p. 677–706, 2008. 2.4

NAVARRO, S. Control functions. In: *The New Palgrave Dictionary of Economics*. [S.l.]: Palgrave Macmillan UK, 2008. p. 1–7. 19

NETO, C. A. d. S. C.; SOARES, R. P.; FERREIRA, I. M.; POMPERMAYER, F. M.; ROMMINGER, A. E. Gargalos e demandas da infraestrutura rodoviária e os investimentos do PAC: mapeamento IPEA de obras rodoviárias. *Textos para Discussão - IPEA*, Instituto de Pesquisa Econômica Aplicada (IPEA), n. 1592, 2011. <http://repositorio.ipea.gov.br/handle/11058/1637>. 2.4

OLIVEIRA, R. C.; SILVEIRA NETO, R. d. M. Escolaridade, políticas sociais e a evolução da desigualdade regional de renda no Brasil entre 2003 e 2011: uma análise a partir das fontes de renda. *Revista Econômica do Nordeste*, v. 44, n. 3, p. 651–670, 2013. 4.1

OLIVEIRA, R. C.; SILVEIRA NETO, R. D. M. Expansão da escolaridade e redução da desigualdade regional de renda no Brasil entre 1995 e 2011: progressos recentes e desafios presentes. *Pesquisa e Planejamento Econômico*, v. 46, n. 1, 2016. <http://ppe.ipea.gov.br/index.php/ppe/article/view/1618>. 4.1

OVERMAN, H. G.; PUGA, D. Labor pooling as a source of agglomeration: An empirical investigation. In: *Agglomeration Economics*. [S.l.]: University of Chicago Press, 2010. p. 133–150. 2.4, A.3

PACHECO, C. A. Novos padrões de localização industrial? tendências recentes dos indicadores da produção e do investimento industrial. Instituto de Pesquisa Econômica Aplicada (Ipea), 1999. 2.3.2

PAES, N. L.; SIQUEIRA, M. L. Análise dos efeitos econômicos da implantação do princípio do destino na cobrança do ICMS e suas implicações sobre a pobreza e a desigualdade de renda. *Revista EconomiA*, v. 6, n. 3, p. 91–126, 2005. 3.1

PAES, N. L.; SIQUEIRA, M. L. Desenvolvimento regional e federalismo fiscal no Brasil: em busca da igualdade na distribuição de receitas. *Economia Aplicada*, SciELO Brasil, v. 12, n. 4, p. 707–742, 2008. 3.1

PETRIN, A.; TRAIN, K. A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research*, SAGE Publications Sage CA: Los Angeles, CA, v. 47, n. 1, p. 3–13, 2010. 3.4.3

PUGA, D. The magnitude and causes of agglomeration economies. *Journal of Regional Science*, Wiley Online Library, v. 50, n. 1, p. 203–219, 2010. <https://doi.org/10.1111/j.1467-9787.2009.00657.x>. 1

RESENDE, M.; WYLLIE, R. Aglomeração industrial no Brasil: um estudo empírico. *Estudos Econômicos (São Paulo)*, SciELO Brasil, v. 35, n. 3, p. 433–460, 2005. <http://dx.doi.org/10.1590/S1413-80502007000300002>. 2.1, 2.3.2, 2.3.2, 2.3.3

RICE, P.; VENABLES, A. J.; PATACCHINI, E. Spatial determinants of productivity: analysis for the regions of Great Britain. *Regional Science and Urban Economics*, Elsevier, v. 36, n. 6, p. 727–752, 2006. 20

ROCHA, R. d. M.; ARAÚJO, J. E. S.; ALMEIDA, E. T. d. de. As indústrias da tranformação são concentradas geograficamente? um teste empírico para o Brasil (2002-2014). *Nova Economia*, 2019. 2.1, 2.3.2, 2.3.2, 3.2.1

ROSENTHAL, S. S.; STRANGE, W. C. The determinants of agglomeration. *Journal of Urban Economics*, Elsevier, v. 50, n. 2, p. 191–229, 2001. 2.4, 2.4, 2.4, 3.4.2, 4.4.1

ROSENTHAL, S. S.; STRANGE, W. C. Geography, industrial organization, and agglomeration. *Review of Economics and Statistics*, MIT Press, v. 85, n. 2, p. 377–393, 2003. 1, 3.1, 3.3, 16, 18, 3.4.2, 3.5.1, 3.5.2, 3.5.3, 23, 3.6, 4.1, 4.3.1, 4.4.1, 20

ROSENTHAL, S. S.; STRANGE, W. C. Evidence on the nature and sources of agglomeration economies. In: *Handbook of Regional and Urban Economics*. [S.l.]: Elsevier, 2004. v. 4, p. 2119–2171. 4.1

ROSENTHAL, S. S.; STRANGE, W. C. The geography of entrepreneurship in the New York metropolitan area. *Economic Policy Review*, Citeseer, v. 11, n. 2, p. 29–54, 2005. 20

ROSENTHAL, S. S.; STRANGE, W. C. The attenuation of human capital spillovers. *Journal of Urban Economics*, Elsevier, v. 64, n. 2, p. 373–389, 2008. <https://doi.org/10.1016/j.jue.2008.02.006>. 3.1, 3.3, 3.5.2, 4.1, 4.2, 4.3.1, 4.3.2, 11, 4.3.3, 4.3.4, 4.4.1, 4.4.1, 4.4.2, 4.4.2, 4.4.4, 4.4.5

ROSENTHAL, S. S.; STRANGE, W. C. Small establishments/big effects: Agglomeration, industrial organization and entrepreneurship. In: *Agglomeration Economics*. [S.l.]: University of Chicago Press, 2010. p. 277–302. 3.4.2

ROSENTHAL, S. S.; STRANGE, W. C. How close is close? the spatial reach of agglomeration economies. *Journal of Economic Perspectives*, v. 34, n. 3, p. 27–49, 2020. 3.3, 3.5.1, 1, 4.2, 4.4.1, 4.4.2, 4.4.2, 4.5

SILVA, D. F. C. d.; SILVEIRA NETO, R. d. M. Escolhas de carreiras universitárias e mercado de trabalho: uma análise da influência dos incentivos econômicos. *Nova Economia*, SciELO Brasil, v. 25, n. 3, p. 519–552, 2015. <http://dx.doi.org/10.1590/0103-6351/1941>. 1, 2.3.3

SILVA, M. V. B. d.; SILVEIRA NETO, R. d. M. Dinâmica da concentração da atividade industrial no Brasil entre 1994 e 2004: uma análise a partir de economias de aglomeração e da nova geografia econômica. *Economia Aplicada*, SciELO Brasil, v. 13, n. 2, p. 299–331, 2009. 2.3.2

SILVA, R. L. P. d.; SILVEIRA NETO, R. d. M.; ROCHA, R. Localization patterns within urban areas: evidence from Brazil. *Area Development and Policy*, Routledge, p. 1–20, 2019. <https://doi.org/10.1080/23792949.2019.1571424>. 2.1

SILVEIRA NETO, R. d. M. Concentração industrial regional, especialização geográfica e geografia econômica: evidências para o Brasil no período 1950-2000. *Revista Econômica do Nordeste*, v. 36, n. 2, p. 189–208, 2005. <https://ren.emnuvens.com.br/ren/article/view/732>. 2.1, 2.3.2, 2.3.3, 3.2.1

SILVERMAN, B. W. *Density estimation for statistics and data analysis*. [S.l.]: Routledge, 2018. <https://doi.org/10.1201/9781315140919>. 1

STAIGER, D.; STOCK, J. H. Instrumental variables regression with weak instruments. *Econometrica*, v. 65, n. 3, p. 557–586, 1997. 4.4.1

STOCK, J. H.; WRIGHT, J. H.; YOGO, M. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, Taylor & Francis, v. 20, n. 4, p. 518–529, 2002. 4.4.1

STORPER, M.; VENABLES, A. J. Buzz: face-to-face contact and the urban economy. *Journal of Economic Geography*, Oxford University Press, v. 4, n. 4, p. 351–370, 2004. 2.4

STUART, B. C.; DOSHI, J. A.; TERZA, J. V. Assessing the impact of drug use on hospital costs. *Health Services Research*, Wiley Online Library, v. 44, n. 1, p. 128–144, 2009. 3.4.3

SULIANO, D. C.; SIQUEIRA, M. L. Retornos da educação no Brasil em âmbito regional considerando um ambiente de menor desigualdade. *Economia Aplicada*, SciELO Brasil, v. 16, n. 1, p. 137–165, 2012. <http://dx.doi.org/10.1590/S1413-80502012000100006>. 1, 2.3.3

TATSCH, A. L.; RUFFONI, J.; BATISTI, V. d. S.; ROXO, L. A. T. Análise de políticas para aglomerações no Brasil e em países europeus selecionados. *Planejamento e Políticas Públicas*, Instituto de Pesquisa Econômica Aplicada (Ipea), v. 44, p. 189–228, 2015. 3.1

TERZA, J. V. Two-stage residual inclusion estimation in health services research and health economics. *Health Services Research*, Wiley Online Library, v. 53, n. 3, p. 1890–1899, 2017. 3.4.3

TERZA, J. V.; BASU, A.; RATHOUZ, P. J. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics*, Elsevier, v. 27, n. 3, p. 531–543, 2008. 19, 3.4.3

THISSE, J.-F. Human capital and agglomeration economies in urban development. *The Developing Economies*, Wiley Online Library, v. 56, n. 2, p. 117–139, 2018. <https://doi.org/10.1111/deve.12167>. 1, 3.3, 4.1

VARSANO, R. A guerra fiscal do ICMS: quem ganha e quem perde. *Texto para Discussão 500 - IPEA*, Instituto de Pesquisa Econômica Aplicada (Ipea), 1997. 3.1

VENABLES, A. J. Equilibrium locations of vertically linked industries. *International economic review*, JSTOR, p. 341–359, 1996. 2.4

VERSTRATEN, P.; VERWEIJ, G.; ZWANEVELD, P. J. Complexities in the spatial scope of agglomeration economies. *Journal of Regional Science*, Wiley Online Library, v. 59, n. 1, p. 29–55, 2018. 4.3.3

VITALI, S.; NAPOLETANO, M.; FAGIOLO, G. Spatial localization in manufacturing: a cross-country analysis. *Regional Studies*, Taylor & Francis, v. 47, n. 9, p. 1534–1554, 2013. <https://doi.org/10.1080/00343404.2011.625006>. 1, 2.1, 2.3.1, 2.3.2

WEST, L.; BEINROTH, F.; SUMNER, M.; KANG, B. Ultisols: Characteristics and impacts on society. *Advances in Agronomy*, Elsevier, v. 63, p. 179–236, 1997. 4.3.4

WOODCOCK, S. D. Match effects. *Research in Economics*, Elsevier, v. 69, n. 1, p. 100–121, 2015. 4.4.4

WOOLDRIDGE, J. M. Quasi-likelihood methods for count data. *Handbook of Applied Econometrics*, Blackwell, v. 2, p. 352–406, 1997. 3.4.3, 3.4.3

WOOLDRIDGE, J. M. *Econometric analysis of cross section and panel data*. [S.l.]: MIT press, 2010. 3.4.3

WOOLDRIDGE, J. M. Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. *Journal of Econometrics*, Elsevier, v. 182, n. 1, p. 226–234, 2014. 19

Appendix A

Appendix to Chapter 2

A.1 Data: additional details

In this appendix we provide additional information on the source and treatment of the database used in the research.

Our main source is the RAIS database. The data used for geocoding contains detailed information about each plant's geographical location, such as address and postal code, but not the geographical coordinates. As this information is updated every year, it incorporates any changes to plant addresses. We only consider plants for which location information was available. In the initial phase of geocoding, the addresses informed in RAIS are compared with Google Maps database to capture the geographic coordinates. In this phase, some plants were not located, often because of incomplete information or typographical errors. In a second stage, the plants not previously located were worked on manually. The last phase included cleaning the data through the geocoding precision. Besides the plants for which the address was not available, we excluded from our sample those addresses which for some reason were located outside their states of origin in the initial phase. The final result can be seen in Table A.1.

Table A.1 Plants and employment geocoded by year

Year	All plants	Total employment	Plants geocoded	Employment geocoded	% plants geocoded	% employ. geocoded
2006	248,789	6,253,684	239,679	6,059,652	96.34	96.90
2007	253,529	6,710,807	244,148	6,501,161	96.30	96.88
2008	264,214	6,905,074	254,313	6,692,968	96.25	96.93
2009	269,741	6,932,127	259,469	6,720,477	96.19	96.95
2010	289,189	7,516,523	278,051	7,276,690	96.15	96.81
2011	300,778	7,726,509	289,084	7,489,454	96.11	96.93
2012	309,596	7,754,545	296,719	7,502,956	95.84	96.76
2013	317,661	7,900,136	303,857	7,639,614	95.65	96.70
2014	322,527	7,765,846	308,227	7,510,367	95.57	96.71
2015	316,473	7,185,512	302,252	6,945,722	95.51	96.66
Total	2,892,497	72,650,763	2,775,799	70,339,061	95.97	96.82

Note: Our main source is the official RAIS micro data at plant level from which, through geocoding techniques, we obtain detailed information on the location of manufacturing industries in the country. Source: Prepared by the author using a unique database.

Table A.2 Percent georeferenced of manufacturing by 3-digit CNAE 2.0 codes in 2006 - 2010

		2006			2007			2008			2009			2010		
	Group (three digit CNAE 2.0)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
101	Slaughtering and production of meat products	3,122	3,049	0.98	3,395	3,330	0.98	3,510	3,427	0.98	3,489	3,408	0.98	3,622	3,548	0.98
102	Preservation of fish and M. of fish products	303	293	0.97	305	296	0.97	317	305	0.96	314	306	0.97	326	317	0.97
103	M. of tinned fruit, vegetables and other vegetables	1,339	1,298	0.97	1,382	1,355	0.98	1,447	1,409	0.97	1,514	1,468	0.97	1,582	1,541	0.97
104	M. of vegetable and animal oils and fats	360	353	0.98	414	407	0.98	412	401	0.97	473	462	0.98	409	402	0.98
105	Dairy products	5,511	5,367	0.97	5,604	5,476	0.98	5,745	5,612	0.98	5,839	5,694	0.98	5,800	5,654	0.97
106	Grinding, M. of starch products and animal feed	4,332	4,250	0.98	4,301	4,235	0.98	4,363	4,279	0.98	4,435	4,346	0.98	4,471	4,383	0.98
107	M. and refining of sugar	334	320	0.96	363	357	0.98	374	361	0.97	370	361	0.98	373	361	0.97
108	Coffee roasting and grinding	1,106	1,092	0.99	1,098	1,086	0.99	1,073	1,055	0.98	1,052	1,036	0.98	1,091	1,071	0.98
109	M. of other food products	16,038	15,856	0.99	15,779	15,623	0.99	15,906	15,735	0.99	15,581	15,394	0.99	21,893	21,637	0.99
111	M. of alcoholic beverages	1,276	1,238	0.97	1,207	1,173	0.97	1,179	1,142	0.97	1,196	1,159	0.97	1,190	1,152	0.97
112	M. of non-alcoholic beverages	888	863	0.97	889	866	0.97	922	895	0.97	936	909	0.97	997	960	0.96
121	Industrial tobacco processing	10	0	0.00	36	35	0.97	31	29	0.94	40	39	0.98	39	38	0.97
122	M. of tobacco products	198	191	0.96	175	168	0.96	171	165	0.96	179	173	0.97	184	178	0.97
131	Preparation and spinning of textile fibres	943	906	0.96	930	892	0.96	937	894	0.95	883	842	0.95	870	836	0.96
132	Weaving, not knitted or crocheted	857	830	0.97	839	817	0.97	823	794	0.96	821	793	0.97	847	820	0.97
133	M. of knitted and crocheted fabrics	829	804	0.97	804	784	0.98	786	761	0.97	718	696	0.97	740	716	0.97
134	Finishing of yarns, fabrics and textile articles	1,490	1,441	0.97	1,712	1,673	0.98	1,886	1,830	0.97	2,004	1,943	0.97	2,222	2,153	0.97
135	M. of textile articles, except apparel	4,398	4,271	0.97	4,521	4,373	0.97	4,748	4,597	0.97	4,817	4,659	0.97	5,003	4,820	0.96
141	M. of wearing apparel and accessories	39,633	38,216	0.96	41,214	39,776	0.97	43,419	41,811	0.96	44,777	43,081	0.96	47,587	45,688	0.96
142	M. of knitted and crocheted articles	1,834	1,765	0.96	1,801	1,731	0.96	1,910	1,841	0.96	1,959	1,890	0.96	2,065	1,982	0.96
151	Tanning and other leather preparations	758	716	0.94	752	717	0.95	733	701	0.96	702	671	0.96	686	654	0.95
152	M. of travel goods and miscellaneous leather goods	2,494	2,445	0.98	2,473	2,425	0.98	2,482	2,440	0.98	2,391	2,349	0.98	2,470	2,435	0.99
153	Footwear manufacturing	7,677	7,559	0.98	7,828	7,713	0.99	8,094	7,984	0.99	7,865	7,759	0.99	8,186	8,093	0.99
154	M. of parts for footwear, of any material	554	542	0.98	626	616	0.98	722	713	0.99	822	808	0.98	991	974	0.98
161	Wood unfolding	6,923	6,665	0.96	6,914	6,671	0.96	6,871	6,630	0.96	6,806	6,554	0.96	6,810	6,560	0.96
162	M. of products of wood except furniture	8,028	7,891	0.98	7,892	7,771	0.98	7,971	7,834	0.98	7,880	7,745	0.98	7,958	7,814	0.98
171	M. of pulp and other pulp for papermaking	103	94	0.91	98	90	0.92	92	85	0.92	96	89	0.93	95	87	0.92
172	M. of paper, paperboard and paperboard	303	292	0.96	300	295	0.98	310	297	0.96	313	299	0.96	314	297	0.95
173	M. of paper and corrugated board packaging	1,528	1,477	0.97	1,593	1,544	0.97	1,619	1,569	0.97	1,633	1,585	0.97	1,705	1,651	0.97
174	M. of miscellaneous paper	2,102	2,047	0.97	2,092	2,045	0.98	2,054	2,010	0.98	1,963	1,918	0.98	1,949	1,907	0.98
181	Print activity	5,635	5,516	0.98	5,914	5,801	0.98	6,395	6,279	0.98	6,836	6,695	0.98	7,723	7,564	0.98
182	Pre-press services and graphic finishing	4,481	4,403	0.98	4,500	4,414	0.98	4,690	4,603	0.98	4,829	4,737	0.98	4,398	4,313	0.98
183	Reproduction of recorded materials on any medium	177	172	0.97	165	160	0.97	167	160	0.96	164	157	0.96	184	175	0.95

Continued on next page

Table A.2 – continued from previous page

		2006			2007			2008			2009			2010		
	Group (three digit CNAE 2.0)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
191	Coke ovens	8	7	0.88	12	9	0.75	12	9	0.75	14	12	0.86	13	11	0.85
192	M. of petroleum products	143	137	0.96	149	141	0.95	169	159	0.94	182	171	0.94	254	235	0.93
193	M. of biofuels	290	269	0.93	319	293	0.92	374	343	0.92	382	354	0.93	377	352	0.93
201	M. of inorganic chemicals	988	931	0.94	973	948	0.97	1,042	967	0.93	1,039	967	0.93	992	925	0.93
202	M. of organic chemicals	460	441	0.96	456	431	0.95	440	423	0.96	432	422	0.98	441	430	0.98
203	M. of resins and elastomers	259	251	0.97	239	234	0.98	265	257	0.97	258	249	0.97	262	251	0.96
204	M. of man-made fibres	74	70	0.95	71	66	0.93	79	74	0.94	82	80	0.98	82	80	0.98
205	M. of pesticides and household disinfectants	129	124	0.96	133	129	0.97	145	141	0.97	154	149	0.97	171	168	0.98
206	M. of soaps and personal care products	2,721	2,619	0.96	2,805	2,695	0.96	2,899	2,773	0.96	2,912	2,786	0.96	3,010	2,871	0.95
207	M. of paints and related products	1,009	967	0.96	1,027	984	0.96	1,068	1,026	0.96	1,090	1,038	0.95	1,165	1,119	0.96
209	M. of miscellaneous chemical products	2,270	2,187	0.96	2,251	2,180	0.97	2,273	2,200	0.97	2,294	2,226	0.97	2,283	2,211	0.97
211	M. of pharmochemicals products	175	165	0.94	175	167	0.95	159	154	0.97	164	156	0.95	159	152	0.96
212	M. of pharmaceutical products	955	933	0.98	882	870	0.99	857	835	0.97	821	798	0.97	778	752	0.97
221	M. of rubber products	2,624	2,542	0.97	2,602	2,537	0.98	2,633	2,563	0.97	2,622	2,544	0.97	2,640	2,564	0.97
222	M. of plastic products	10,576	10,270	0.97	10,609	10,315	0.97	10,803	10,471	0.97	10,723	10,394	0.97	10,903	10,556	0.97
231	M. of glass and glass products	606	583	0.96	604	575	0.95	646	616	0.95	680	647	0.95	755	710	0.94
232	M. of cement	175	154	0.88	159	139	0.87	136	114	0.84	135	114	0.84	135	116	0.86
233	M. of articles of concrete	7,087	6,596	0.93	7,263	6,907	0.95	7,849	7,225	0.92	8,386	7,689	0.92	9,030	8,263	0.92
234	M. of ceramic products	6,185	5,370	0.87	6,266	5,403	0.86	6,409	5,501	0.86	6,486	5,543	0.85	6,724	5,708	0.85
239	M. of other non-metallic mineral products	5,267	4,904	0.93	5,444	5,076	0.93	5,627	5,219	0.93	5,851	5,415	0.93	6,081	5,611	0.92
241	Production of pig iron and ferroalloys	281	272	0.97	268	264	0.99	278	269	0.97	254	246	0.97	257	251	0.98
242	Steel	566	554	0.98	577	564	0.98	578	565	0.98	554	539	0.97	586	575	0.98
243	Production of steel tubes other than seamless tubes	248	239	0.96	258	251	0.97	256	249	0.97	241	235	0.98	253	249	0.98
244	Non-ferrous metal metallurgy	1,509	1,474	0.98	1,431	1,398	0.98	1,375	1,337	0.97	1,278	1,241	0.97	1,300	1,261	0.97
245	Foundry	2,076	2,026	0.98	2,040	1,996	0.98	1,990	1,943	0.98	1,934	1,887	0.98	1,925	1,873	0.97
251	M. of metal structures	7,888	7,791	0.99	8,073	8,237	1.02	8,905	8,808	0.99	9,351	9,250	0.99	10,188	10,070	0.99
252	M. of tanks, metal containers and boilers	477	467	0.98	503	498	0.99	526	515	0.98	546	534	0.98	574	566	0.99
253	Forging and metal treatment services	5,160	5,122	0.99	5,011	4,971	0.99	5,614	5,568	0.99	5,991	5,928	0.99	5,918	5,862	0.99
254	M. of cutlery, locksmiths' wares and tools	4,789	4,746	0.99	5,151	5,106	0.99	5,418	5,363	0.99	5,630	5,564	0.99	6,068	5,998	0.99
255	M. of heavy military equipment	30	30	1.00	29	29	1.00	25	25	1.00	26	26	1.00	27	27	1.00
259	M. of metal products not otherwise specified	8,243	8,171	0.99	8,508	8,446	0.99	8,734	8,644	0.99	8,716	8,623	0.99	8,996	8,908	0.99
261	M. of electronic components	777	766	0.99	814	804	0.99	856	846	0.99	875	864	0.99	877	867	0.99
262	M. of computer and peripheral equipment	524	514	0.98	578	562	0.97	596	581	0.97	596	578	0.97	643	626	0.97
263	M. of communication equipment	312	305	0.98	322	315	0.98	331	321	0.97	325	320	0.98	333	329	0.99
264	M. of reception and apparatus for audio and video	239	234	0.98	252	247	0.98	273	264	0.97	272	265	0.97	281	270	0.96

Continued on next page

Table A.2 – continued from previous page

		2006			2007			2008			2009			2010		
	Group (three digit CNAE 2.0)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
265	M. of measuring, testing and control devices	584	577	0.99	638	631	0.99	689	679	0.99	725	714	0.98	799	784	0.98
266	M. of electromedical equipment	267	263	0.99	239	235	0.98	251	246	0.98	252	248	0.98	261	258	0.99
267	M. of optical equipments	138	137	0.99	137	135	0.99	140	138	0.99	130	128	0.98	127	126	0.99
268	M. of magnetic and optical media	6	6	1.00	3	3	1.00	5	5	1.00	7	7	1.00	5	5	1.00
271	M. of electric generators, transformers and motors	440	431	0.98	434	430	0.99	483	470	0.97	481	471	0.98	501	491	0.98
272	M. of batteries and electric accumulators	203	199	0.98	210	205	0.98	197	192	0.97	179	174	0.97	175	170	0.97
273	M. of equip. for distribution of electrical energy	942	927	0.98	987	984	1.00	1,018	1,002	0.98	1,079	1,061	0.98	1,125	1,109	0.99
274	M. of lamps and other lighting equipment	564	553	0.98	600	590	0.98	601	592	0.99	601	592	0.99	622	613	0.99
275	M. of household appliances	370	366	0.99	387	384	0.99	400	396	0.99	409	406	0.99	432	430	1.00
279	M. of apparatus not otherwise specified	1,203	1,186	0.99	1,195	1,186	0.99	1,245	1,227	0.99	1,242	1,222	0.98	1,263	1,241	0.98
281	M. of engines and transmission equipment	898	884	0.98	950	941	0.99	1,008	992	0.98	1,041	1,025	0.98	1,132	1,114	0.98
282	M. of general-purpose machinery and equipment	4,363	4,297	0.98	4,418	4,353	0.99	4,604	4,534	0.98	4,637	4,562	0.98	4,787	4,718	0.99
283	M. of agricultural machinery and equipment	1,264	1,230	0.97	1,335	1,310	0.98	1,452	1,418	0.98	1,456	1,418	0.97	1,584	1,543	0.97
284	M. of machine tools	744	734	0.99	826	816	0.99	893	881	0.99	941	928	0.99	1,020	1,002	0.98
285	M. of machinery for use in mineral extraction	237	232	0.98	257	252	0.98	256	251	0.98	262	257	0.98	301	298	0.99
286	M. of machinery for industrial uses	3,066	3,014	0.98	3,250	3,200	0.98	3,467	3,406	0.98	3,573	3,515	0.98	3,782	3,717	0.98
291	M. of automobiles, vans and utilities	97	94	0.97	92	89	0.97	95	91	0.96	87	85	0.98	90	87	0.97
292	M. of trucks and buses	36	35	0.97	39	38	0.97	37	36	0.97	30	29	0.97	33	31	0.94
293	M. of motor vehicle cabins, bodies and trailers	1,029	997	0.97	1,097	1,070	0.98	1,188	1,149	0.97	1,201	1,161	0.97	1,265	1,222	0.97
294	M. of parts and accessories for motor vehicles	2,385	2,356	0.99	2,525	2,501	0.99	2,624	2,586	0.99	2,702	2,668	0.99	2,779	2,744	0.99
295	Recovery of engines for motor vehicles	815	800	0.98	899	882	0.98	993	964	0.97	1,046	1,015	0.97	1,130	1,094	0.97
301	Shipbuilding	272	261	0.96	295	285	0.97	320	300	0.94	334	312	0.93	362	333	0.92
303	M. of railway vehicles	47	45	0.96	50	47	0.94	44	42	0.95	47	45	0.96	52	50	0.96
304	M. of aircraft	45	42	0.93	54	52	0.96	67	62	0.93	66	63	0.95	77	74	0.96
305	M. of military fighting vehicles	0	0	–	0	0	–	–	–	–	0	0	–	0	0	–
309	M. of transport equipment not otherwise specified	449	419	0.93	457	428	0.94	460	430	0.93	475	449	0.95	488	460	0.94
310	Furniture manufacturing	14,442	13,044	0.90	14,634	13,211	0.90	15,116	13,543	0.90	15,459	13,823	0.89	16,463	14,628	0.89
321	M. of jewellery and related articles	1,245	1,141	0.92	1,320	1,211	0.92	1,386	1,267	0.91	1,473	1,347	0.91	1,600	1,465	0.92
322	M. of musical instruments	116	112	0.97	128	123	0.96	119	116	0.97	132	129	0.98	136	134	0.99
323	M. of fishing and sporting goods	236	225	0.95	254	241	0.95	294	285	0.97	296	285	0.96	333	322	0.97
324	M. of toys and recreational games	495	464	0.94	512	484	0.95	532	505	0.95	557	526	0.94	580	551	0.95
325	M. of instruments for medical use	1,619	1,578	0.97	1,739	1,699	0.98	1,945	1,901	0.98	2,245	2,191	0.98	2,493	2,431	0.98
329	M. of miscellaneous products	2,626	2,479	0.94	2,620	2,483	0.95	2,906	2,762	0.95	3,173	3,043	0.96	3,793	3,624	0.96
331	Repair of machinery and equipment	5,440	4,768	0.88	5,357	4,708	0.88	5,992	5,339	0.89	6,617	6,022	0.91	8,056	7,462	0.93
332	Installation of machinery and equipment	1,422	1,330	0.94	1,423	1,328	0.93	1,588	1,490	0.94	1,793	1,669	0.93	2,282	2,122	0.93

Continued on next page

Table A.2 – continued from previous page

	2006			2007			2008			2009			2010		
Group (three digit CNAE 2.0)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Total	248,789	239,679	0.96	253,531	245,037	0.97	264,214	254,313	0.96	269,741	259,469	0.96	289,189	278,051	0.96

Notes: Columns (1), (2) and (3) represents the total number of plants, total number of georeferenced plants, and the share georeferenced, respectively.

Source: Prepared by the author with data from RAIS.

Table A.3 Percent georeferenced of manufacturing by 3-digit CNAE 2.0 codes in 2011 - 2015

		2011			2012			2013			2014			2015		
	Group (three digit CNAE 2.0)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
101	Slaughtering and production of meat products	3,680	3,615	0.98	3,771	3,702	0.98	3,872	3,800	0.98	3,964	3,886	0.98	4,023	3,948	0.98
102	Preservation of fish and M. of fish products	323	319	0.99	334	328	0.98	364	354	0.97	386	373	0.97	403	390	0.97
103	M. of tinned fruit, vegetables and other vegetables	1,703	1,665	0.98	1,835	1,791	0.98	1,944	1,906	0.98	2,055	2,007	0.98	2,115	2,070	0.98
104	M. of vegetable and animal oils and fats	415	409	0.99	443	435	0.98	428	419	0.98	465	457	0.98	454	447	0.98
105	Dairy products	5,905	5,788	0.98	5,876	5,770	0.98	5,885	5,783	0.98	5,866	5,766	0.98	5,764	5,658	0.98
106	Grinding, M. of starch products and animal feed	4,467	4,383	0.98	4,562	4,468	0.98	4,594	4,507	0.98	4,640	4,558	0.98	4,628	4,548	0.98
107	M. and refining of sugar	382	371	0.97	378	368	0.97	372	364	0.98	368	359	0.98	366	355	0.97
108	Coffee roasting and grinding	1,067	1,047	0.98	1,070	1,054	0.99	1,094	1,074	0.98	1,102	1,086	0.99	1,082	1,062	0.98
109	M. of other food products	21,282	21,072	0.99	20,999	20,793	0.99	21,796	21,608	0.99	23,068	22,866	0.99	24,100	23,891	0.99
111	M. of alcoholic beverages	1,140	1,115	0.98	1,129	1,101	0.98	1,142	1,117	0.98	1,137	1,116	0.98	1,170	1,148	0.98
112	M. of non-alcoholic beverages	1,051	1,024	0.97	1,061	1,037	0.98	1,075	1,053	0.98	1,111	1,089	0.98	1,131	1,112	0.98
121	Industrial tobacco processing	35	34	0.97	31	30	0.97	30	29	0.97	30	29	0.97	31	29	0.94
122	M. of tobacco products	187	178	0.95	184	176	0.96	178	172	0.97	180	174	0.97	187	179	0.96
131	Preparation and spinning of textile fibres	870	832	0.96	847	816	0.96	803	780	0.97	786	761	0.97	739	715	0.97
132	Weaving, not knitted or crocheted	850	819	0.96	825	801	0.97	831	802	0.97	834	806	0.97	800	770	0.96
133	M. of knitted and crocheted fabrics	764	738	0.97	740	712	0.96	695	671	0.97	682	655	0.96	652	627	0.96
134	Finishing of yarns, fabrics and textile articles	2,319	2,250	0.97	2,402	2,319	0.97	2,597	2,505	0.96	2,594	2,504	0.97	2,520	2,427	0.96
135	M. of textile articles, except apparel	5,189	5,006	0.96	5,358	5,171	0.97	5,401	5,207	0.96	5,455	5,254	0.96	5,219	5,016	0.96
141	M. of wearing apparel and accessories	49,552	47,601	0.96	50,148	48,065	0.96	50,816	48,601	0.96	50,464	48,244	0.96	47,264	45,147	0.96
142	M. of knitted and crocheted articles	2,073	1,975	0.95	2,089	1,995	0.96	1,949	1,870	0.96	1,897	1,813	0.96	1,787	1,708	0.96
151	Tanning and other leather preparations	670	646	0.96	654	637	0.97	615	596	0.97	595	576	0.97	580	561	0.97
152	M. of travel goods and miscellaneous leather goods	2,496	2,464	0.99	2,468	2,432	0.99	2,404	2,369	0.99	2,396	2,366	0.99	2,228	2,205	0.99
153	Footwear manufacturing	8,194	8,112	0.99	8,135	8,055	0.99	7,925	7,852	0.99	7,561	7,488	0.99	6,800	6,732	0.99
154	M. of parts for footwear, of any material	1,086	1,071	0.99	1,220	1,203	0.99	1,263	1,248	0.99	1,293	1,280	0.99	1,218	1,204	0.99
161	Wood unfolding	6,788	6,606	0.97	6,739	6,555	0.97	6,622	6,440	0.97	6,458	6,285	0.97	6,243	6,072	0.97
162	M. of products of wood except furniture	8,042	7,908	0.98	8,101	7,976	0.98	7,991	7,864	0.98	7,928	7,808	0.98	7,614	7,501	0.99
171	M. of pulp and other pulp for papermaking	82	75	0.91	85	77	0.91	83	72	0.87	78	67	0.86	72	65	0.90
172	M. of paper, paperboard and paperboard	310	295	0.95	301	286	0.95	291	280	0.96	295	285	0.97	291	280	0.96
173	M. of paper and corrugated board packaging	1,823	1,765	0.97	1,864	1,796	0.96	1,890	1,820	0.96	1,907	1,842	0.97	1,897	1,829	0.96
174	M. of miscellaneous paper	1,971	1,929	0.98	1,938	1,892	0.98	1,926	1,876	0.97	1,946	1,883	0.97	1,892	1,827	0.97
181	Print activity	8,340	8,138	0.98	8,793	8,517	0.97	9,121	8,788	0.96	9,319	9,000	0.97	9,274	8,928	0.96
182	Pre-press services and graphic finishing	4,249	4,152	0.98	4,053	3,958	0.98	3,870	3,762	0.97	3,737	3,621	0.97	3,511	3,398	0.97
183	Reproduction of recorded materials on any medium	177	173	0.98	194	184	0.95	191	181	0.95	175	168	0.96	157	148	0.94

Continued on next page

Table A.3 – continued from previous page

		2011			2012			2013			2014			2015		
	Group (three digit CNAE 2.0)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
191	Coke ovens	13	10	0.77	11	8	0.73	12	10	0.83	12	10	0.83	10	8	0.80
192	M. of petroleum products	247	232	0.94	252	236	0.94	260	247	0.95	241	235	0.98	245	236	0.96
193	M. of biofuels	358	332	0.93	345	322	0.93	334	313	0.94	313	291	0.93	307	287	0.93
201	M. of inorganic chemicals	1,045	981	0.94	1,089	1,024	0.94	1,181	1,107	0.94	1,152	1,079	0.94	1,109	1,047	0.94
202	M. of organic chemicals	442	425	0.96	442	425	0.96	421	406	0.96	398	386	0.97	397	386	0.97
203	M. of resins and elastomers	267	258	0.97	274	264	0.96	276	267	0.97	276	267	0.97	276	268	0.97
204	M. of man-made fibres	80	77	0.96	80	77	0.96	77	75	0.97	79	77	0.97	71	66	0.93
205	M. of pesticides and household disinfectants	192	183	0.95	201	191	0.95	194	181	0.93	196	183	0.93	217	204	0.94
206	M. of soaps and personal care products	3,047	2,906	0.95	3,107	2,963	0.95	3,133	2,967	0.95	3,169	3,012	0.95	3,148	2,996	0.95
207	M. of paints and related products	1,214	1,173	0.97	1,207	1,161	0.96	1,241	1,185	0.95	1,271	1,217	0.96	1,243	1,193	0.96
209	M. of miscellaneous chemical products	2,241	2,181	0.97	2,210	2,147	0.97	2,163	2,091	0.97	2,141	2,079	0.97	2,098	2,036	0.97
211	M. of pharmochemicals products	163	152	0.93	142	137	0.96	146	142	0.97	140	136	0.97	133	128	0.96
212	M. of pharmaceutical products	740	720	0.97	747	721	0.97	727	701	0.96	679	648	0.95	668	638	0.96
221	M. of rubber products	2,689	2,616	0.97	2,672	2,593	0.97	2,646	2,572	0.97	2,604	2,527	0.97	2,543	2,467	0.97
222	M. of plastic products	11,016	10,669	0.97	11,029	10,665	0.97	10,992	10,616	0.97	10,927	10,551	0.97	10,793	10,434	0.97
231	M. of glass and glass products	797	734	0.92	871	794	0.91	939	856	0.91	1,007	906	0.90	1,032	920	0.89
232	M. of cement	144	123	0.85	145	121	0.83	135	111	0.82	133	105	0.79	137	104	0.76
233	M. of articles of concrete	9,722	8,814	0.91	10,483	9,398	0.90	10,927	9,728	0.89	11,055	9,762	0.88	11,055	9,767	0.88
234	M. of ceramic products	6,954	5,874	0.84	7,133	5,997	0.84	7,024	5,865	0.83	6,934	5,747	0.83	6,647	5,522	0.83
239	M. of other non-metallic mineral products	6,509	5,984	0.92	6,798	6,168	0.91	7,163	6,445	0.90	7,539	6,777	0.90	7,590	6,796	0.90
241	Production of pig iron and ferroalloys	241	236	0.98	221	216	0.98	207	203	0.98	203	200	0.99	183	180	0.98
242	Steel	592	581	0.98	584	572	0.98	606	595	0.98	602	589	0.98	595	586	0.98
243	Production of steel tubes other than seamless tubes	245	240	0.98	274	270	0.99	264	260	0.98	259	256	0.99	241	238	0.99
244	Non-ferrous metal metallurgy	1,306	1,270	0.97	1,300	1,265	0.97	1,284	1,250	0.97	1,221	1,189	0.97	1,150	1,115	0.97
245	Foundry	1,934	1,882	0.97	1,858	1,814	0.98	1,807	1,765	0.98	1,715	1,676	0.98	1,603	1,570	0.98
251	M. of metal structures	11,085	10,967	0.99	11,884	11,767	0.99	12,541	12,424	0.99	12,823	12,699	0.99	12,576	12,456	0.99
252	M. of tanks, metal containers and boilers	581	571	0.98	572	562	0.98	600	591	0.99	573	563	0.98	564	556	0.99
253	Forging and metal treatment services	6,239	6,194	0.99	6,506	6,457	0.99	6,921	6,859	0.99	7,177	7,124	0.99	7,271	7,230	0.99
254	M. of cutlery, locksmiths' wares and tools	6,615	6,563	0.99	7,100	7,048	0.99	7,586	7,529	0.99	7,843	7,789	0.99	7,763	7,699	0.99
255	M. of heavy military equipment	24	24	1.00	23	23	1.00	25	25	1.00	28	28	1.00	22	22	1.00
259	M. of metal products not otherwise specified	9,236	9,165	0.99	9,304	9,227	0.99	9,282	9,209	0.99	9,186	9,113	0.99	8,894	8,813	0.99
261	M. of electronic components	896	882	0.98	913	898	0.98	889	876	0.99	880	866	0.98	849	837	0.99
262	M. of computer and peripheral equipment	661	646	0.98	626	614	0.98	622	612	0.98	599	588	0.98	565	556	0.98
263	M. of communication equipment	341	335	0.98	351	346	0.99	332	324	0.98	320	315	0.98	292	288	0.99
264	M. of reception and apparatus for audio and video	290	282	0.97	312	304	0.97	327	319	0.98	316	308	0.97	311	303	0.97

Continued on next page

Table A.3 – continued from previous page

		2011			2012			2013			2014			2015		
	Group (three digit CNAE 2.0)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
265	M. of measuring, testing and control devices	849	834	0.98	847	837	0.99	888	876	0.99	920	908	0.99	899	888	0.99
266	M. of electromedical equipment	254	251	0.99	246	241	0.98	251	247	0.98	239	235	0.98	241	238	0.99
267	M. of optical equipments	127	125	0.98	122	120	0.98	126	125	0.99	122	122	1.00	110	109	0.99
268	M. of magnetic and optical media	11	11	1.00	10	9	0.90	11	10	0.91	11	10	0.91	9	9	1.00
271	M. of electric generators, transformers and motors	518	508	0.98	544	536	0.99	535	526	0.98	538	530	0.99	536	528	0.99
272	M. of batteries and electric accumulators	169	165	0.98	167	164	0.98	153	152	0.99	151	150	0.99	147	146	0.99
273	M. of equip. for distribution of electrical energy	1,182	1,163	0.98	1,241	1,225	0.99	1,329	1,313	0.99	1,402	1,385	0.99	1,425	1,408	0.99
274	M. of lamps and other lighting equipment	630	619	0.98	649	637	0.98	629	617	0.98	642	634	0.99	621	616	0.99
275	M. of household appliances	450	447	0.99	463	459	0.99	461	456	0.99	466	463	0.99	454	452	1.00
279	M. of apparatus not otherwise specified	1,272	1,253	0.99	1,258	1,240	0.99	1,207	1,191	0.99	1,199	1,184	0.99	1,170	1,156	0.99
281	M. of engines and transmission equipment	1,208	1,181	0.98	1,262	1,240	0.98	1,263	1,238	0.98	1,285	1,259	0.98	1,243	1,216	0.98
282	M. of general-purpose machinery and equipment	4,948	4,885	0.99	5,030	4,955	0.99	5,153	5,080	0.99	5,153	5,079	0.99	5,027	4,961	0.99
283	M. of agricultural machinery and equipment	1,623	1,591	0.98	1,678	1,644	0.98	1,778	1,744	0.98	1,790	1,760	0.98	1,805	1,768	0.98
284	M. of machine tools	1,104	1,086	0.98	1,151	1,136	0.99	1,172	1,158	0.99	1,168	1,154	0.99	1,113	1,103	0.99
285	M. of machinery for use in mineral extraction	324	318	0.98	345	342	0.99	347	345	0.99	375	371	0.99	381	378	0.99
286	M. of machinery for industrial uses	3,964	3,901	0.98	4,040	3,972	0.98	4,146	4,086	0.99	4,130	4,064	0.98	3,965	3,907	0.99
291	M. of automobiles, vans and utilities	100	95	0.95	104	100	0.96	97	93	0.96	101	98	0.97	98	96	0.98
292	M. of trucks and buses	37	35	0.95	35	34	0.97	41	40	0.98	43	42	0.98	38	37	0.97
293	M. of motor vehicle cabins, bodies and trailers	1,408	1,376	0.98	1,515	1,470	0.97	1,620	1,571	0.97	1,664	1,620	0.97	1,636	1,596	0.98
294	M. of parts and accessories for motor vehicles	2,877	2,845	0.99	2,932	2,898	0.99	2,960	2,924	0.99	2,943	2,909	0.99	2,863	2,830	0.99
295	Recovery of engines for motor vehicles	1,194	1,162	0.97	1,279	1,249	0.98	1,319	1,295	0.98	1,374	1,349	0.98	1,412	1,392	0.99
301	Shipbuilding	405	361	0.89	444	393	0.89	456	399	0.88	458	406	0.89	455	401	0.88
303	M. of railway vehicles	54	52	0.96	57	55	0.96	66	64	0.97	70	65	0.93	73	69	0.95
304	M. of aircraft	77	72	0.94	77	73	0.95	85	78	0.92	83	75	0.90	75	70	0.93
305	M. of military fighting vehicles	1	1	1.00	0	0	–	0	0	–	0	0	–	0	0	–
309	M. of transport equipment not otherwise specified	494	460	0.93	537	495	0.92	548	500	0.91	555	512	0.92	551	503	0.91
310	Furniture manufacturing	17,530	15,397	0.88	18,672	16,148	0.86	19,753	16,873	0.85	20,666	17,506	0.85	20,451	17,171	0.84
321	M. of jewellery and related articles	1,703	1,539	0.90	1,826	1,620	0.89	1,949	1,707	0.88	1,950	1,689	0.87	1,916	1,641	0.86
322	M. of musical instruments	139	134	0.96	136	132	0.97	135	128	0.95	139	132	0.95	141	136	0.96
323	M. of fishing and sporting goods	359	350	0.97	389	369	0.95	418	388	0.93	430	399	0.93	428	393	0.92
324	M. of toys and recreational games	602	562	0.93	616	567	0.92	629	579	0.92	624	577	0.92	609	557	0.91
325	M. of instruments for medical use	2,802	2,746	0.98	3,139	3,082	0.98	3,497	3,436	0.98	3,733	3,666	0.98	3,892	3,836	0.99
329	M. of miscellaneous products	4,299	4,065	0.95	4,732	4,446	0.94	5,103	4,758	0.93	5,363	5,007	0.93	5,355	4,999	0.93
331	Repair of machinery and equipment	9,539	8,970	0.94	11,002	10,435	0.95	12,363	11,826	0.96	13,688	13,171	0.96	14,174	13,685	0.97
332	Installation of machinery and equipment	2,877	2,667	0.93	3,358	3,101	0.92	3,831	3,538	0.92	4,269	3,926	0.92	4,350	4,025	0.93

Continued on next page

Table A.3 – continued from previous page

	2011			2012			2013			2014			2015		
Group (three digit CNAE 2.0)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Total	300,778	289,084	0.96	309,596	296,719	0.96	317,661	303,857	0.96	322,527	308,227	0.96	316,473	302,269	0.96

Notes: Columns (1), (2) and (3) represents the total number of plants, total number of georeferenced plants, and the share georeferenced, respectively.

Source: Prepared by the author with data from RAIS.

Table A.4 Selected statistics by large regions, state of São Paulo and SPMR

Large regions	2006				2015			
	empl.	# of plants	share # of plants	Avg. empl.	empl.	# of plants	share # of plants	Avg. empl.
Southeast	3,301,753	124,029	49.85	26.62	3,596,277	146,133	46.18	24.61
São Paulo	2,238,987	74,911	30.11	29.89	2,371,621	84,853	26.81	27.95
SPMR	1,001,086	35,827	14.40	27.94	925,665	36,791	11.63	25.16
South	1,624,587	73,749	29.64	22.03	1,919,087	93,492	29.54	20.53
Northeast	798,372	28,794	11.57	27.73	980,464	43,823	13.85	22.37
Midwest	298,999	15,064	6.05	19.85	441,679	23,134	7.31	19.09
North	229,973	7,153	2.88	32.15	248,005	9,891	3.13	25.07
Total	6,253,684	248,789	100	25.14	7,185,512	316,473	100	22.70

Notes: The state of São Paulo is in Southeast region and SPMR is the metropolitan region that contains the capital city (core city of the SPMR). Source: Author's computation using information from RAIS (Annual Report of Social Information).

Table A.5 High- and low-tech manufacturing plants in Brazil by metropolitan regions in 2015

High-tech industries		
	# Plants	%
São Paulo Metropolitan Region	8,200	20.37
Rio de Janeiro Metropolitan Region	1,318	3.27
Belo Horizonte Metropolitan Region	1,373	3.41
Porto Alegre Metropolitan Region	1,900	4.72
Recife Metropolitan Region	508	1.26
# plants in RMs	13,299	33.04
Brazil	40,249	100
Low-tech industries		
	# Plants	%
São Paulo Metropolitan Region	14,101	9.94
Rio de Janeiro Metropolitan Region	3,866	2.73
Belo Horizonte Metropolitan Region	3,086	2.18
Porto Alegre Metropolitan Region	5,212	3.67
Recife Metropolitan Region	1,913	1.35
# plants in RMs	28,178	19.86
Brazil	141,849	100

Note: Percentage of plants by technological level located in the five largest metropolitan regions of the country (according to the 2010 Demographic Census made available by the Brazilian Institute of Geography and Statistics (IBGE)). Source: Elaborated by the author based on a unique database.

A.2 Location Patterns: nonparametric analysis

This appendix provides details on the method used by Duranton and Overman (2005). The procedure for constructing this metric consists of four steps:

First step - Obtain the kernel densities. Using microgeographic data, which allows the location of each firm in space through the geographic coordinates, the Euclidean distance between a pair of plants (i, j) is initially calculated. For n plants, we have $\frac{n(n-1)}{2}$ unique pairs of distances. Using a Kernel density function, the density of the bilateral distances at any target distance r can be calculated according to:

$$\hat{\mathrm{K}}d_{obs}(r) = \frac{1}{n(n-1)h} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f\left(\frac{r - r_{i,j}}{h}\right) \tag{A.1}$$

where $r_{i,j}$ is the distance between plants i and j, h is bandwidth,[1] and $f(.)$ is a Gaussian kernel function. At this stage, we also calculate, for some years, the variation of $\hat{\mathrm{K}}d_{obs}(r)$ weighted by the number of jobs in each plant, given by:

$$\hat{\mathrm{K}}d^{emp}_{obs}(r) = \frac{1}{h \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} e(i)e(j)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} e(i)e(j) f\left(\frac{r - r_{i,j}}{h}\right) \tag{A.2}$$

where $e(i)$ denotes employment of firm i.

Second step - Counterfactual densities. Counterfactuals are generated by sampling (without replacement) of the number of plants by sector considering the total number of sites. In this case, since we work with only one sector of the manufacturing industry, the total population of sites is the sector as a whole (counterfactual) and we subdivide the sample according to the interests of the research (3-digit and 4-digit manufacturing activities). Proceeding in this way, we control for the general agglomeration of the sector analyzed. Given a subdivision m, each sample is a pseudo-m subdivision, for which a density is estimated $\tilde{\mathrm{K}}d(r)$. For each m, thousand[2] $\tilde{\mathrm{K}}d(r)$ are generated so that each simulation is equivalent to a random reshuffling of establishments across sites. These are simulations of null hypotheses that form the confidence interval.

Third step - Confidence bands. Following Duranton and Overman (2005), we consider

[1] Following Silverman (2018), the ideal bandwidth for the Gaussian kernel function is $1.06sn^{-0.2}$, where s is the standard deviation of $n(n\text{-}1)$ bilateral distances.

[2] Following Duranton and Overman (2005), we also repeated our simulations 2,000 and 10,000 times for selected industries. There were no changes in the results.

the distances from 0 to the median, $\bar{r}$ of all bilateral distances in the sample. The analysis is restricted to $r \in [0, \bar{r}]$. The maximum range for Brazil is 1,708.11 km.[3] Because it is a large country, as highlighted by Aleksandrova *et al.* (2019), the window size should be adequate. The authors, in their study for Russia, considered 1,000 km. The study of Behrens and Bougna (2015) for Canada considered 800 km. For each subdivision of the sample in this interval, a $\tilde{\mathrm{K}}d(r)$ is estimated. The lower $\hat{\mathrm{K}}d_{lo}(r)$ and upper $\hat{\mathrm{K}}d_{hi}(r)$ limits are defined so that no less than 95% of the $\tilde{\mathrm{K}}d(r)$ estimated are between $\hat{\mathrm{K}}d_{lo}(r)$ and $\hat{\mathrm{K}}d_{hi}(r)$.

Fourth step - Identification of location patterns. Once we have identified the confidence bands, we can classify the pattern of localization as localized, dispersed or random by following: i) if $\hat{\mathrm{K}}d_{obs}(r) > \hat{\mathrm{K}}d_{hi}(r)$ for at least one r, the industry is localized; ii) if $\hat{\mathrm{K}}d_{obs}(r) < \hat{\mathrm{K}}d_{lo}(r)$ for at least one r and $\hat{\mathrm{K}}d_{obs}(r) < \hat{\mathrm{K}}d_{hi}(r)$ for all r, the industry is dispersed.[4]

For each industry m, the location ($\Gamma_m(r)$) and dispersion ($\Psi_m(r)$) indices are defined, respectively, by:

$$\Gamma_m(r) \equiv \max\left\{\hat{\mathrm{K}}d_{obs,m}(r) - \hat{K}d_{hi,m}(r), 0\right\} \tag{A.3}$$

$$\Psi_m(r) \equiv \begin{cases} \max\left\{\hat{K}d_{lo,m}(r) - \hat{\mathrm{K}}d_{obs,m}(r), 0\right\} & \text{if } \sum_{r=0}^{\bar{r}} \Gamma_m(r) = 0 \\ 0 \qquad \text{otherwise} \end{cases} \tag{A.4}$$

In the graphical analysis, for an industry to be considered localized, it suffices that the estimated distribution of bilateral distances be above the upper confidence range for at least a distance r. On the other hand, for an industry to be considered dispersed, the distance distribution must be below the lower confidence band for at least one distance r and never above the upper limit. The indices of localization and dispersion for all the

[3] In Brazilian case, a cutoff distance of 1,708 kilometers includes interactions within the "Southern Region cluster" (Porto Alegre, RS; Florianópolis, SC; Curitiba, PR); the "Southeast Region cluster" (São Paulo, SP; Rio de Janeiro, RJ; Belo Horizonte, MG; Vitória, ES); the "Midwest Region cluster" (Brasília, DF; Goiânia, GO); and the "Northeast Region cluster" (Recife, PE; Salvador, BA, Fortaleza, CE; Terezina, PI; São Luís, MA; Maceió, AL). A cutoff distance of 1,708 kilometers in the Brazilian context allows us to account for industrial localization at very small spatial scales (e.g., within the São Paulo Metropolitan Region (SPMR)), at medium spatial scales (e.g., between SPRM and Belo Horizonte Metropolitan Region (BHMR)), and also at larger interregional scales (e.g., between SPRM and Salvador Metropolitan Region (SMR)) for which the input-output linkages can be more important.

[4] When a particular industry m shows concentration peaks, i.e., $\hat{\mathrm{K}}d_{obs,m}(r)$ exceeds the upper confidence limit $\hat{\mathrm{K}}d_{hi,m}(r)$, other points on the $\hat{\mathrm{K}}d_{obs,m}(r)$ curve may fall below the lower confidence limit, $\hat{\mathrm{K}}d_{lo,m}(r)$, as a form of compensation. This is because the values are normalized to add to 1, but do not imply that the industry in question is dispersed over the associated distance.

events are given respectively by:

$$\Gamma_m = \sum_{r=0}^{\bar{r}} \Gamma_m(r) \qquad \text{and} \qquad \Psi_m = \sum_{r=0}^{\bar{r}} \Psi_m(r) \tag{A.5}$$

A.3 Proxies for covariates, additional figures, tables, and results

This appendix presents additional details about the explanatory variables, other figures, tables and results.

Labor pooling. Following the strategy of Overman and Puga (2010), we use detailed RAIS data for all manufacturing plants in Brazil for the period 2005-2015 to calculate the difference between the variation in plant level employment and the variation in industrial level employment at the 3-digit level. RAIS data allow us to monitor a plant from opening to closing with annually updated information on the structure of the plant and its workers. Formally this is given by:

$$\mathsf{Labor\,pooling}_m = \frac{1}{\mathsf{N}_m} \sum_{i=1}^{\mathsf{N}_m} |\Delta_{im} - \Delta_m| \tag{A.6}$$

where N_m is the number of plants in industry m; Δ_{im} is the percentage change in employment in plant i pertaining to industry m; and Δ_m is the percentage change in employment in industry m.

Input sharing. Official data for input-output tables are not available for the ten years of our panel. We deal with the scarcity of data using the input-output tables made available by the Nucleus of Regional and Urban Economy (NEREUS) of the University of São Paulo,[5] built by the method used by Guilhoto and Sesso Filho (2005, 2010). The classification of activities in the input-output tables is not CNAE 2.0; instead, there is no entirely equivalent disaggregation. Thus, sectoral matches were made at the 2-digit level and we repeated the values in the industries at the 3-digit level that forms each sector at the 2-digit level. As previously mentioned, we recognize that this substitution can suppress specific characteristics of each sector at the 3-digit level, so we estimated the regressions with and without this variable.

The other covariates (transport cost and proxies for spatial heterogeneity) were obtained

[5]Details available at <http://www.usp.br/nereus/?fontes=dados-matrizes>

from the Annual Industrial Survey - Company (*Pesquisa Industrial Anual - Empresa*, or PIA) database, made available by IBGE. The PIA provides economic and financial information (structure of revenues, expenses and investments) of companies in the primary industries and manufacturing industry in Brazil.[6]

K-densities across years. As highlighted by Behrens and Bougna (2015), despite all the advantages of distance-based location measures, it is not clear how these measures can be compared across years and between countries. In order to analyze whether we can compare the estimated measures over the ten years of our sample, we proceeded in an analogous manner to those authors. Figure A.2 plots the 95% confidence bands for the DO index applied to a 5% random sample of all manufacturing plants for the years 2006, 2010 and 2015. The general pattern of location did not change much in the years analyzed, suggesting that the reference distribution did not change much and that we can compare the results across years.

Table A.6 shows how manufacturing location patterns changed among large groups. We rank the 2-digit industries at the top of the panel as those with more industries located at 3-digit and also provide the average of the maximum location distances (in kilometers) per industry, i.e., the average of the maximum distances where the observed K-density intercepts the upper confidence band. Among the more localized industries are *textile products manufacturing* (CNAE 13), *apparel manufacturing* (CNAE 14), *leather and related products* (CNAE 15) and *machinery manufacturing* (CNAE 28). These industries in developing countries (see Brakman *et al.* (2016) for China and Aleksandrova *et al.* (2019) for Russia) and developed ones (see Behrens and Bougna (2015) for Canada) with large territories also show similar patterns of localization. These findings about textile industry are also similar to those for the UK (Duranton and Overman, 2005), Japan (Nakajima *et al.*, 2012), and Germany (Koh and Riedel, 2014), where this industry is among the most localized.

[6]Details available at <https://www.ibge.gov.br/estatisticas/economicas/industria>

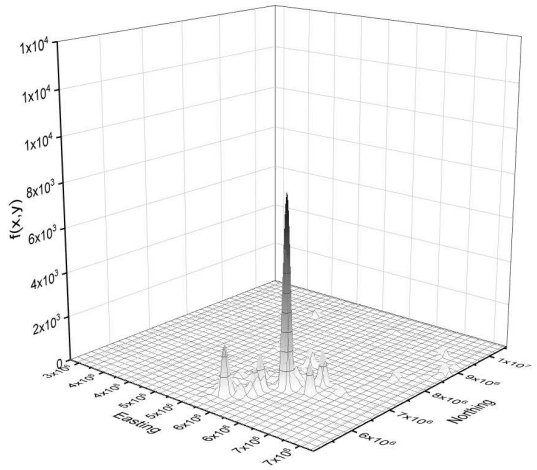
Table A.6 Location patterns of CNAE 3-digit by 2-digit levels in 2006 and 2015 (high-tech sectors in the hatched lines)

CNAE 2-digit	Industry name	2006						2015					
		# of 3-digit	# L.	# D.	# R.	% L.	Location distance	# of 3-digit	# L.	# D.	# R.	% L.	Location distance
22	Plastics & rubber products	2	2	0	0	100	509.758	2	2	0	0	100	516.443
31	Furniture manufacturing	1	1	0	0	100	1313.67	1	1	0	0	100	705.304
32	Miscellaneous manufacturing	6	6	0	0	100	583.297	6	6	0	0	100	738.174
24	Metallurgy	5	5	0	0	100	675.889	5	5	0	0	100	753.439
13	Textile products manufacturing	5	5	0	0	100	881.797	5	5	0	0	100	889.82
28	Machinery manufacturing	6	6	0	0	100	773.829	6	6	0	0	100	893.608
14	Apparel manufacturing	2	2	0	0	100	991.103	2	2	0	0	100	909.207
25	Metal products mfg	6	5	0	1	83	1098.4	6	6	0	0	100	974.947
12	Tobacco products manufacturing	1	0	1	0	0	–	2	2	0	0	100	1002.8
19	Petroleum & biofuels mfg	3	0	1	2	0	–	2	2	0	0	100	1012.83
33	Maintenance of machinery	2	2	0	0	100	991.103	2	2	0	0	100	1156.56
23	Nonmetallic mineral products	5	5	0	0	100	1169.27	5	5	0	0	100	1415.96
15	Leather and related products	4	4	0	0	100	1591.11	4	4	0	0	100	1583.59
11	Beverage production	2	2	0	0	100	1708.11	2	2	0	0	100	1708.11
16	Wood products manufacturing	2	2	0	0	100	1708.11	2	2	0	0	100	1708.11
10	Food processing	9	8	1	0	89	1708.11	9	8	1	0	89	1655.88
20	Chemical manufacturing	8	7	0	1	88	1087.32	8	7	0	1	88	604.546
26	Computer & electronic products	8	7	0	1	88	564.434	7	6	1	0	86	506.972
27	Electrical machinery mfg	6	5	0	1	83	516.777	6	5	1	0	83	653.158
29	Motor vehicle manufacturing	5	5	0	0	100	1109.77	5	4	0	1	80	1115.62
30	Transport exc. motor vehicles	4	3	0	1	75	425.634	4	3	1	0	75	472.431
17	Paper manufacturing	4	3	0	1	75	525.914	4	3	0	1	75	659.621
18	Printing & Related Support activ.	3	3	0	0	100	774.386	3	2	0	1	67	1047.93
21	Pharmaceuticals products mfg	2	1	0	1	50	484.688	2	1	0	1	50	665.192

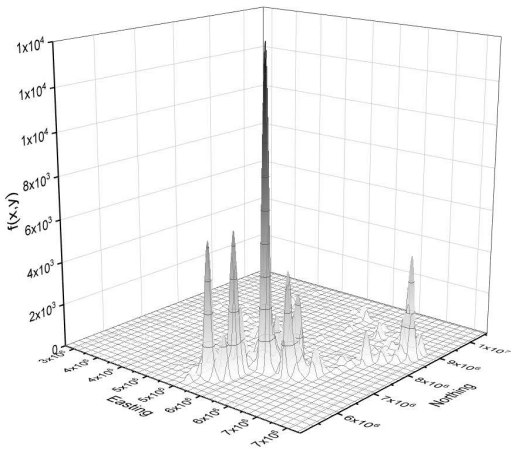
Notes: The manufacturing industry is divided into 24 two-digit groups. # L., # D. and # R. represent the number of sectors at the 3-digit levels localized, dispersed and randomly distributed, respectively. “Location distance" is the average of the maximum distances (in kilometers) per industry where the observed K-density intercepts the upper confidence band. The highlighted lines (in grey) represent the sectors classified as technology intensive (medium-high and high levels) according to the compatibility proposed by Cavalcante (2014) for the OECD technological classification. Source: Prepared by the author based on estimates.

(a) High-tech plants

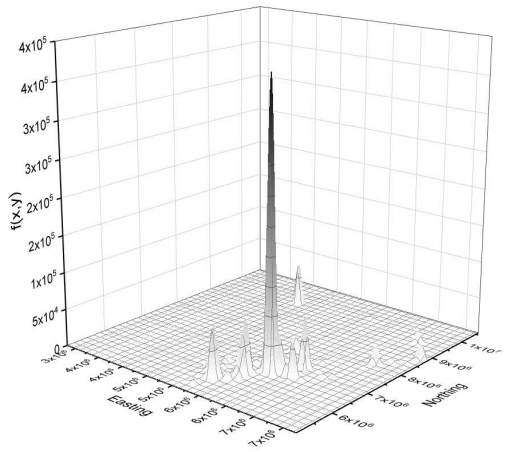
(b) Low-tech plants



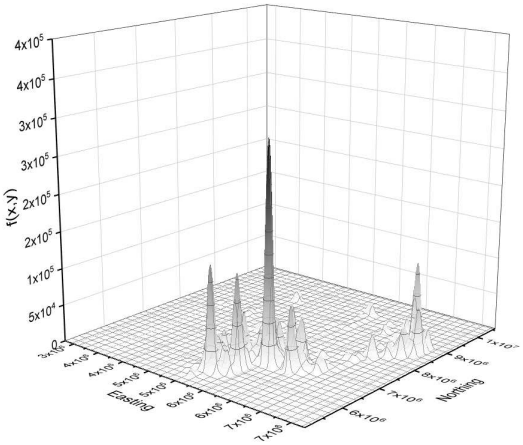
(c) High-tech bivariate kernel density



(d) Low-tech bivariate kernel density



(e) High-tech bivariate kernel density employment weighted



(f) Low-tech bivariate kernel density employment weighted

Figure A.1 Location of manufacturing plants by technology group in 2015

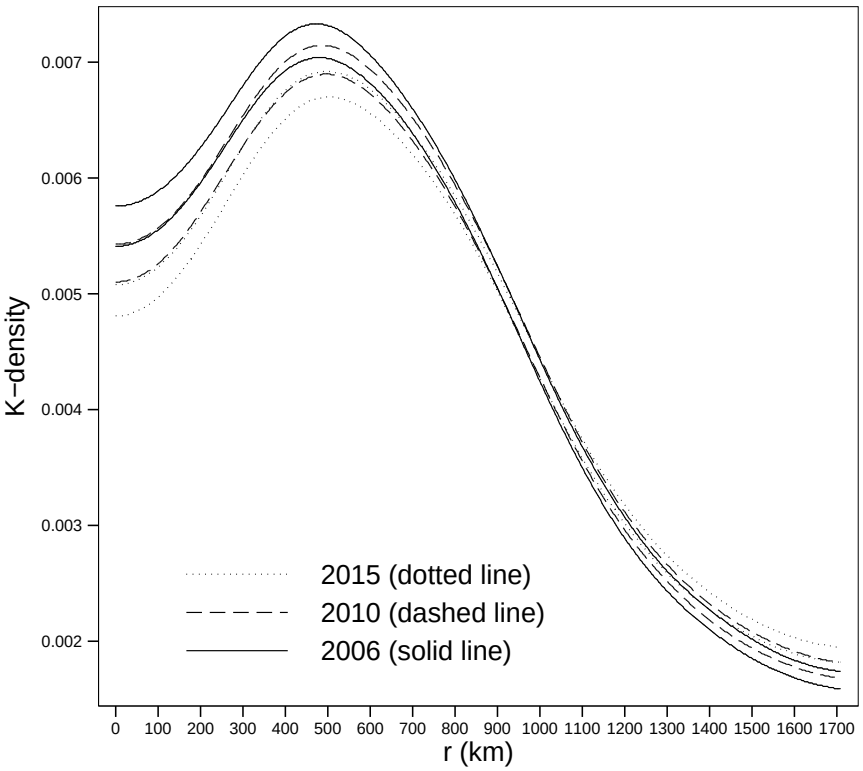
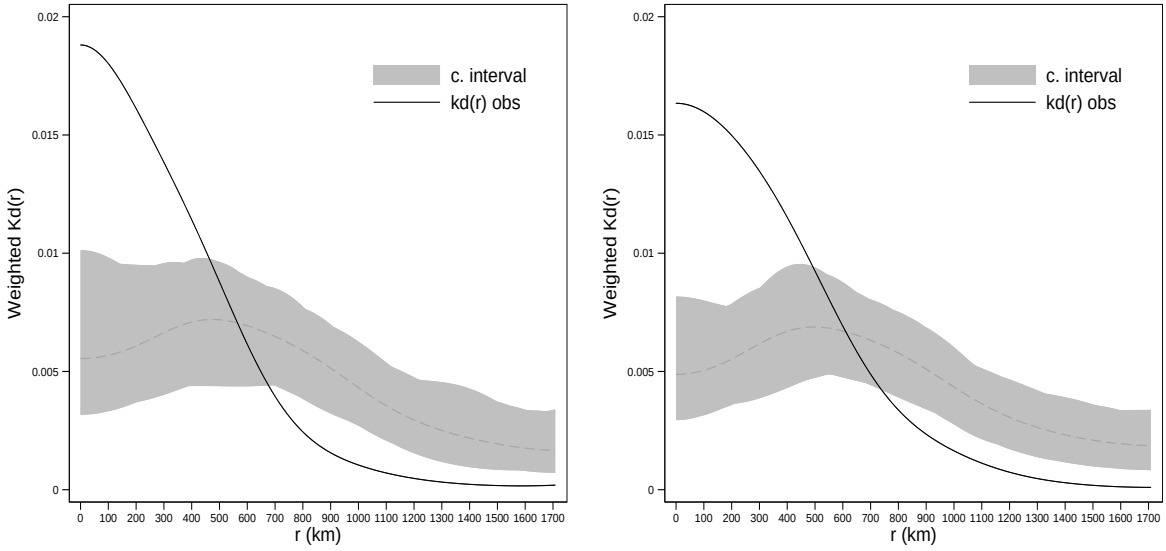
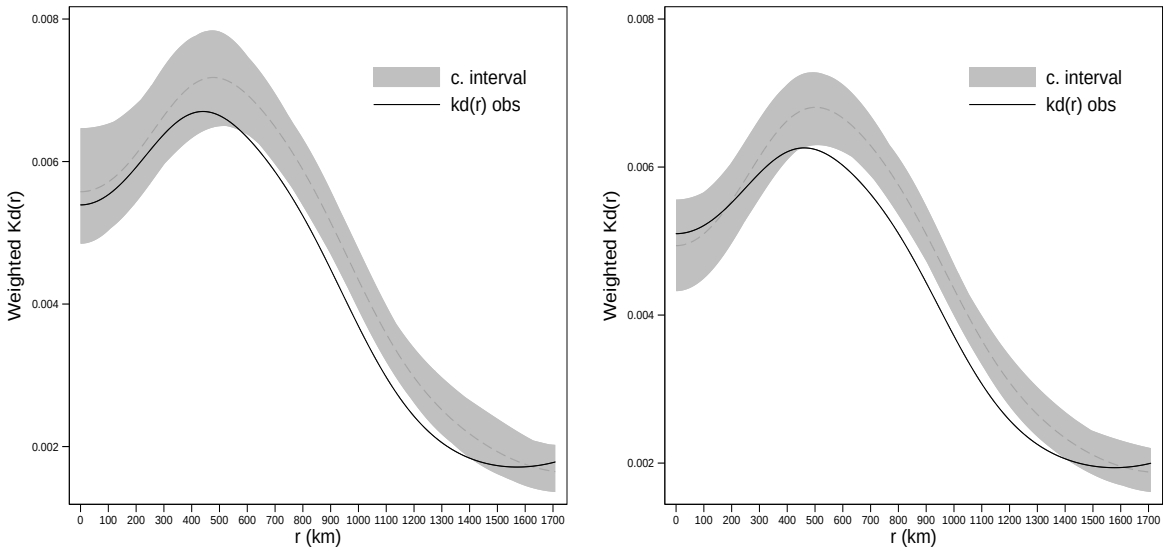


Figure A.2 Distribution of distances between plants, 95% confidence bands



(a) 2006 (b) 2015

M. of electro-medical and electrotherapeutic equipment - CNAE 266



(c) 2006 (d) 2015

M. of other food products - CNAE 109

Figure A.3 Weighted K-density estimates for selected manufacturing sectors (3-digit CNAE) localized (a and b) and dispersed (c and d) in 2006 and 2015

Appendix B

Appendix to Chapter 3

B.1 Data: additional details

In this appendix we provide additional information on the source and treatment of the database used in the research.

B.1.1 Microgeographic data

Table B.1 Births and new-establishment employment

Large regions	2007-2008				2013-2014			
	# of new plants	# of new empl.	share # of plants	Avg. empl.	# of new plants	# of new empl.	share # of plants	Avg. empl.
Southeast	11,125	91,773	42.45	8.25	11,318	81,682	41.63	7.22
São Paulo	6,143	60,844	23.44	9.90	5,675	51,731	20.87	9.12
SPMR	2,577	20,503	9.83	7.96	2,087	15,520	7.68	7.44
South	8,578	45,127	32.73	5.26	8,165	48,988	30.03	6.00
Northeast	3,667	13,044	13.99	3.56	4,251	27,287	15.63	6.42
Midwest	2,035	12,647	7.76	6.21	2,612	16,736	9.61	6.41
North	804	3,939	3.07	4.90	844	6,027	3.10	7.14
Total	26,209	166,530	100	6.35	27,190	180,720	100	6.65

Notes: We consider all new manufacturing establishments in the 2007-2008 and 2013-2014 periods. Source: Author's computations using information from RAIS.

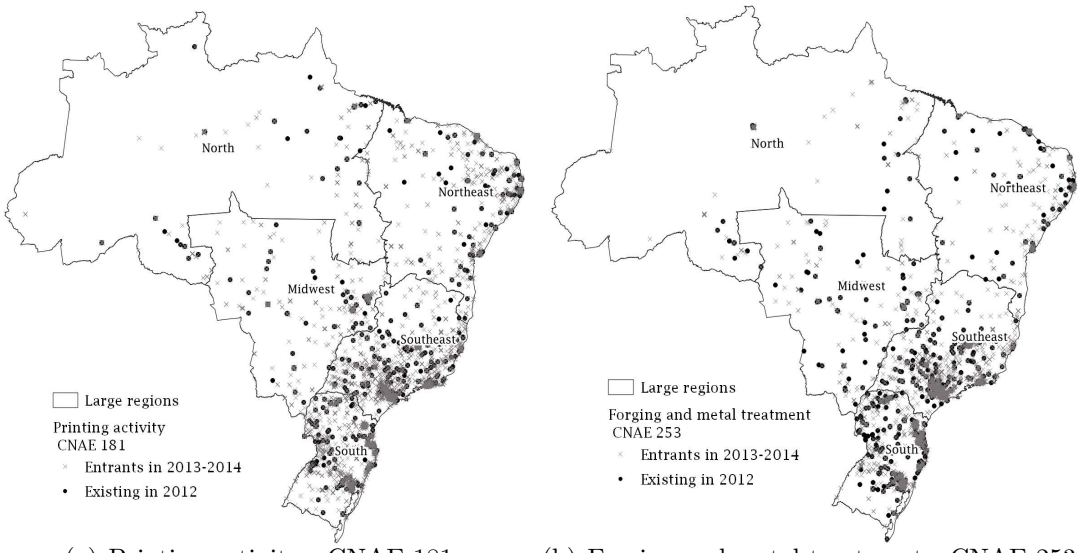
Our main source is the RAIS database. The data used for geocoding contains detailed information about each plant's geographical location, such as address and postal code, but not the geographical coordinates. As this information is updated every year, it incorporates any changes to plant addresses. We only consider plants for which location information was available. In the initial phase of geocoding, the addresses informed in RAIS are compared with Google Maps database to capture the geographic coordinates. In this phase, some plants were not located, often because of incomplete information or typographical errors. In a second stage, the plants not previously located were worked on manually. The last phase included cleaning the data through the geocoding precision. Besides the plants for which the address was not available, we excluded from our sample

those addresses which for some reason were located outside their states of origin in the initial phase. The final result can be seen in Table B.2.

Table B.2 Births and new-establishment employment geocoded by year

Year	New Plants	New Employment	New Plants geocoded	New Empl. geocoded	Percent geo. of plants	Percent geo. of empl.
2007	11,637	84,890	11,578	84,789	99.49	99.88
2008	14,723	81,976	14,656	81,836	99.54	99.83
2009	15,553	77,947	15,478	77,392	99.52	99.29
2010	18,413	117,115	18,304	116,146	99.41	99.17
2011	16,870	99,932	16,796	99,760	99.56	99.83
2012	15,081	109,733	15,031	107,494	99.67	97.96
2013	15,955	94,346	15,894	93,573	99.62	99.18
2014	11,352	87,855	11,309	87,174	99.62	99.22
Total	119,584	753,794	119,046	748,164	99.55	99.25

Source: Prepared by the author using information from RAIS.



(a) Printing activity - CNAE 181 (b) Forging and metal treatment - CNAE 253

Figure B.1 Maps of two illustrative industries

B.1.2 Source of control variables

Here we present more details about our control variables. The transportation controls are constructed with data from Empresa de Planejamento e Logística S.A. (EPL), available at <https://www.epl.gov.br/>. These controls include:

- Distance to the airport: linear distance from the cell's centroid to the nearest airport;
- Distance to the port: linear distance from the cell's centroid to the nearest public port;
- Distance to the railway: linear distance from the cell's centroid to the nearest railway;
- Distance to the federal highway: linear distance from the cell's centroid to the nearest federal highway;
- Distance to the state highway: linear distance from the cell's centroid to the nearest state highway.

The geographic control is the linear distance to the nearest river and is constructed with shapefiles from IBGE, available at <https://portaldemapas.ibge.gov.br/>.

All control variables at the municipal level, are obtained from the Ipeadata, available at <http://www.ipeadata.gov.br/>. In this set the variables are:

- Exports (FOB) per hundred thousand inhabitants;
- Imports (FOB) per hundred thousand inhabitants;
- Capital expenditures (investment) per hundred thousand inhabitants, which include the funds for the planning and execution of construction, including the acquisition of real estate, acquisition of equipment and permanent material, and the constitution or increase of the capital of companies that are not of a commercial or financial nature;
- Housing and town planning expenses per hundred thousand inhabitants;
- Municipal taxes per hundred thousand inhabitants, which include urban property tax, tax on services, and other taxes;
- Homicides per hundred thousand inhabitants;
- Traffic fatalities per hundred thousand inhabitants.

Table B.3 Selected summary statistics

	Pharma. Products - CNAE 212			Food products - CNAE 109			Fruit Canning - CNAE 103			Starch Products - CNAE 106		
	Mean	Std. Dev.	Max	Mean	Std. Dev.	Max	Mean	Std. Dev.	Max	Mean	Std. Dev.	Max
Own Industry												
Births	0.001	0.03	1	0.06	0.25	5	0.004	0.06	3	0.01	0.08	2
New-firm workers	0.04	4.92	1,351	0.23	7.14	1,428	0.04	2.82	582	0.06	2.51	337
# of workers												
Within 0 to 1 km	6.51	81.34	4,012	23.65	120.41	6,174	2.58	36.07	2,236	7.21	43.96	1,977
Within 1 to 5 km	108.44	586.68	14,630	310.02	641.02	7,554	26.25	131.53	2,269	72.51	183.45	2,559
Within 5 to 10 km	273.47	1,217.01	17,823	600.88	1,358.90	12,577	45.22	184.00	2,358	118.62	285.46	3,498
Within 10 to 20 km	780.68	2,878.27	25,512	1,467.62	3,209.54	21,473	94.54	256.11	3,925	262.98	509.29	4,554
Within 20 to 40 km	1,387.88	4,297.06	31,662	2,494.24	5,067.24	37,854	212.76	424.97	4,343	483.84	712.67	5,219
Other Industry												
# of workers												
Within 0 to 1 km	70.80	650.85	33,194	331.06	991.62	34,385	56.66	486.16	34,193	123.02	707.59	34,429
Within 1 to 5 km	4,373.89	12,243.42	151,571	6,130.46	11,803.35	146,205	4,724.05	11,905.34	151,139	5,748.71	12,505.49	153,062
Within 5 to 10 km	10,448.50	29,153.63	338,103	11,998.15	28,663.08	342,369	11,244.88	29,794.97	353,013	11,876.26	29,768.11	353,061
Within 10 to 20 km	30,766.81	77,600.28	558,717	31,161.52	76,040.80	555,609	32,889.75	79,837.74	582,646	32,696.18	78,891.06	569,966
Within 20 to 40 km	60,465.43	131,888.30	876,953	62,958.37	129,582.70	871,144	64,268.44	134,660.60	905,375	64,760.78	134,028.90	902,663
	Furniture - CNAE 310			Prepress Services - CNAE 182			Wood Products - CNAE 162			Textile Articles - CNAE 134		
	Mean	Std. Dev.	Max	Mean	Std. Dev.	Max	Mean	Std. Dev.	Max	Mean	Std. Dev.	Max
Own Industry												
Births	0.05	0.25	8	0.01	0.09	4	0.02	0.13	4	0.01	0.11	5
New-firm workers	0.16	3.01	523	0.02	1.36	347	0.08	4.21	1,101	0.03	0.93	150
# of workers												
Within 0 to 1 km	16.44	82.15	5,957	2.99	19.14	968	7.71	41.77	1,922	3.06	25.26	1,316
Within 1 to 5 km	181.28	456.06	11,163	50.87	172.05	2,865	71.24	171.98	2,745	39.95	144.74	2,266
Within 5 to 10 km	308.25	664.20	9,571	115.23	386.81	4,569	121.03	270.12	3,388	77.06	239.62	2,900
Within 10 to 20 km	779.12	1,634.58	13,029	306.32	944.54	7,522	302.16	603.82	4,313	188.27	503.32	6,101
Within 20 to 40 km	1,506.28	2,681.84	16,127	494.04	1,411.08	10,735	648.38	1,019.53	6,559	424.63	962.88	7,925
Other Industry												
# of workers												
Within 0 to 1 km	304.47	967.77	34,359	187.15	896.91	34,318	225.98	881.43	34,434	113.03	736.42	33,917
Within 1 to 5 km	6,231.69	12,061.93	148,349	5,976.49	12,300.87	147,066	6,331.60	12,349.57	148,675	5,245.18	12,421.22	151,864
Within 5 to 10 km	12,238.71	29,311.92	349,570	11,978.38	29,612.72	350,062	12,389.39	29,659.20	351,893	11,461.00	29,834.48	352,486
Within 10 to 20 km	31,858.90	77,619.78	563,040	32,194.37	78,302.48	564,280	32,317.89	78,572.26	568,452	32,286.89	79,035.55	567,568
Within 20 to 40 km	63,893.80	132,158.40	892,171	64,302.19	133,431.00	89,698	64,706.70	133,676.90	901,987	63,498.59	134,490.60	902,849

Notes: All values calculated at the cell level over 132,072 cells.

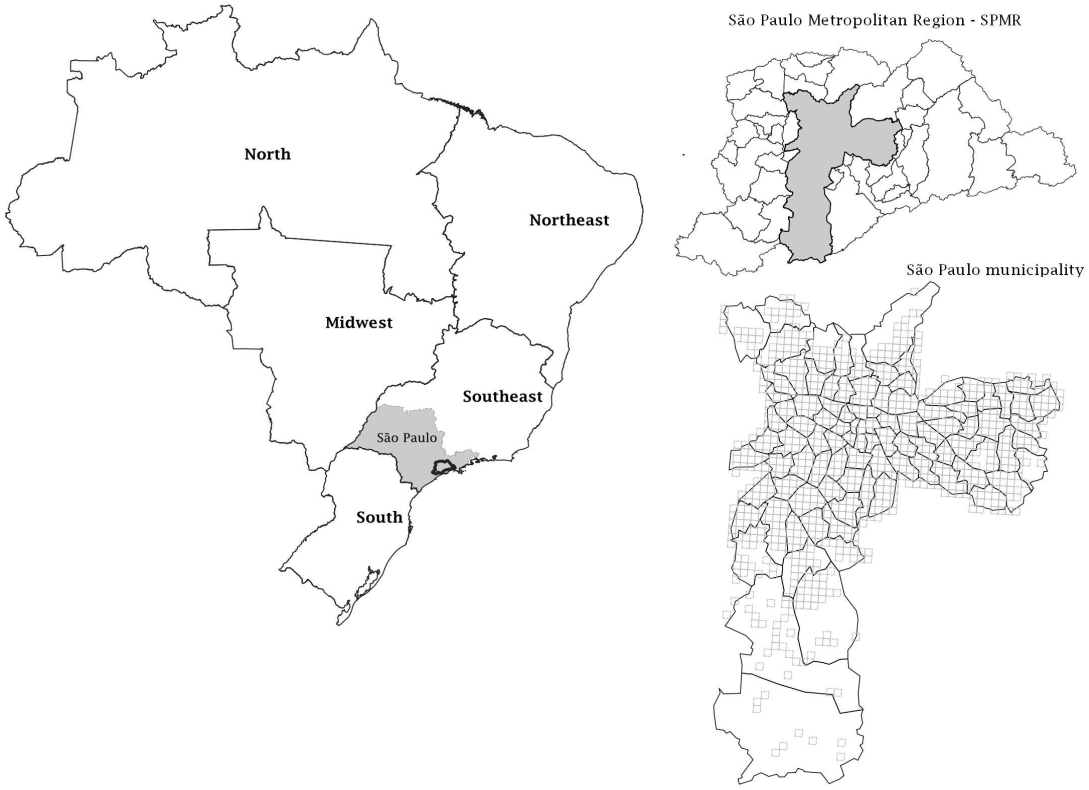


Figure B.2 Districts of the municipality of São Paulo

B.2 Duranton and Overman's nonparametric approach

This appendix provides details on the method of Duranton and Overman (2005; 2008) and the additional results obtained from these measurements.

B.2.1 Methodology

As discussed earlier, we used the Duranton and Overman (2005; 2008)'s nonparametric approach to document the location and colocation patterns of Brazilian manufacturing entrepreneurship. The measurement is carried out in four steps. The way the K-density is calculated follows the same pattern presented in the Appendix to Chapter 1, but we now present only technical details that are different from those presented previously.

Firts step - Obtain the kernel densities. As in Duranton and Overman (2008), among all the establishments of each industry at the 3-digit CNAE level, we distinguish between new entrants and existing establishments. First, we calculate the bilateral distances between each pair of entrants. Using the bilateral distribution of distances between

entrants, we can assess whether they show similar location patterns to incumbents in the same industry. Thus, for n entrants, there are $\frac{n(n-1)}{2}$ unique bilateral distances between entrants. Using a Kernel density function, the density of the bilateral distances at any target distance r can be calculated according to:

$$\hat{\mathrm{K}}d_{obs}(r) = \frac{1}{n(n-1)h} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f\left(\frac{r - r_{i,j}}{h}\right) \tag{B.1}$$

where $r_{i,j}$ is the distance between plants i and j, h is bandwidth, and $f(.)$ is a Gaussian kernel function.

Since our main objective is to investigate the spatial scope of externalities generated by proximity to own-industry employment, we prioritize the K-density employment-weighted version, because the density in equation B.1 does not consider firm size. For comparison purposes only, the unweighted density results are also presented in the next subsection. As explained by Duranton and Overman (2005), in the employment-weighted density the focus is on workers, so that the bilateral distances between all pairs of workers who employed by to different establishments are considered. Formally:

$$\hat{\mathrm{K}}d^{emp}_{obs}(r) = \frac{1}{h \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} e(i)e(j)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} e(i)e(j) f\left(\frac{r - r_{i,j}}{h}\right) \tag{B.2}$$

where $e(i)$ denotes employment of firm i.

As discussed earlier, another interesting question that we also explore from K-densities is if entrants locate near to (or far from) existing establishments. To do this, as in Duranton and Overman (2008), we calculate the distribution of bilateral distances between entrants and all existing establishments in the unweighted (equation B.3) and employment-weighted versions (equation B.4). For example, consider an industry with n entrants and m incumbents in a given period. In this context, there are nm unique bilateral distances and the K-density at any point r is given by:

$$\hat{\mathrm{K}}d_{obs(n,m)}(r) = \frac{1}{nmh} \sum_{i=1}^{n} \sum_{j=1}^{m} f\left(\frac{r - r_{i,j}}{h}\right) \tag{B.3}$$

$$\hat{\mathrm{K}}d^{emp}_{obs(n,m)}(r) = \frac{1}{h \sum_{i=1}^{n} \sum_{j=1}^{m} e(i)e(j)} \sum_{i=1}^{n} \sum_{j=1}^{m} e(i)e(j) f\left(\frac{r - r_{i,j}}{h}\right) \tag{B.4}$$

Second step - Counterfactual densities. A possible counterfactual is to consider a hypothetical industry that is located in the same way as an actual industry and has the number of establishments, both existing and new establishments, but where we know that entrants locate no differently from existing establishments. To do this, we draw (without replacement) the same number of entrants from the population of sites occupied by the specific industry. In other words, we restrict the counterfactual to the locations that contain establishments of the same industry only, whether entrants or incumbents. As in Duranton and Overman (2008), we consider that if two establishments share the same geographic coordinates, two sites are distinguished. As highlighted by the authors, this procedure is equivalent to classifying all establishments as entrants or incumbents while holding the share of each group constant.

Third step - Confidence bands. We consider the distances from 0 to the median, $\bar{r}$ of all bilateral distances in the sample. We compare the actual K-densities to the counterfactuals in each $r \in [0, \bar{r}]$. In the Brazilian case, $\bar{r} = 1708.11$ km (we justify this cutoff in the Appendix of the chapter 2). For each subdivision of the sample in this interval, a $\tilde{\mathrm{K}}d(r)$ is estimated. The lower $\hat{\mathrm{K}}d_{lo}(r)$ and upper $\hat{\mathrm{K}}d_{hi}(r)$ limits are defined so that no less than 95% of the $\tilde{\mathrm{K}}d(r)$ estimated are between $\hat{\mathrm{K}}d_{lo}(r)$ and $\hat{\mathrm{K}}d_{hi}(r)$.

Fourth step - Identification of location and colocation patterns. When we look at the distribution of bilateral distances between entrants (equation B.1), for each industry j, if $\hat{\mathrm{K}}d_{obs,j}(r) > \hat{\mathrm{K}}d_{hi,j}(r)$ for at least one r, the entrants in this industry are localized (at a 5% confidence level). On the other hand, if $\hat{\mathrm{K}}d_{obs,j}(r) < \hat{\mathrm{K}}d_{lo,j}(r)$ for at least one r and $\hat{\mathrm{K}}d_{obs,j}(r) < \hat{\mathrm{K}}d_{hi,j}(r)$ for all r, the entrants in this industry is dispersed. Similarly, when we look at the distribution of bilateral distances between entrants and all existing establishments (equation B.3), in the first case, the entrants in this industry are colocalized (at a 5% confidence level) while in the second case the entrants in this industry are codispersed.

For each industry j, the location ($\Gamma_j(r)$) and dispersion ($\Psi_j(r)$) indices are defined, respectively, by:

$$\Gamma_j(r) \equiv \max\left\{\hat{\mathrm{K}}d_{obs,j}(r) - \hat{\mathrm{K}}d_{hi,j}(r), 0\right\} \tag{B.5}$$

$$\Psi_j(r) \equiv \begin{cases} \max\left\{\hat{\mathrm{K}}d_{lo,j}(r) - \hat{\mathrm{K}}d_{obs,j}(r), 0\right\} & \text{if } \sum_{r=0}^{\bar{r}} \Gamma_j(r) = 0 \\ 0 \qquad \text{otherwise} \end{cases} \tag{B.6}$$

In the graphical analysis (examples in Figure 3.2), for an industry to be considered localized, it suffices that the estimated distribution of bilateral distances be above the

upper confidence range for at least a distance r. However, for an industry to be considered dispersed, the distances distribution must be below the lower confidence band for at least one distance r and never above the upper limit. The indices of localization and dispersion for all the events are given respectively by:

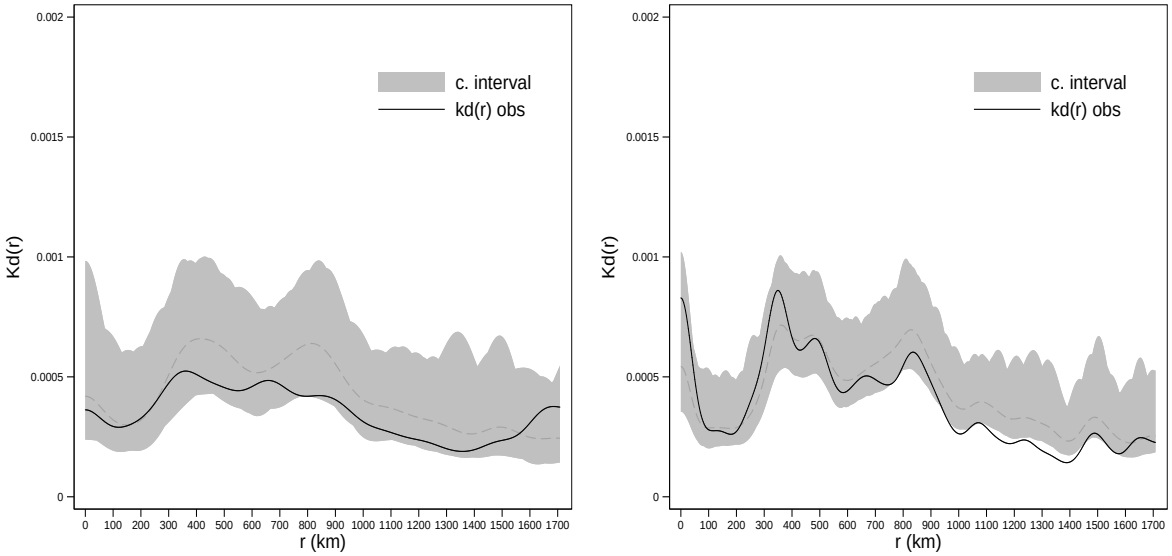
$$\Gamma_j = \sum_{r=0}^{\bar{r}} \Gamma_j(r) \qquad \text{and} \qquad \Psi_j = \sum_{r=0}^{\bar{r}} \Psi_j(r) \tag{B.7}$$

B.2.2 Additional results

Table B.4 Localization and Colocalization of new establishments (unweighted version)

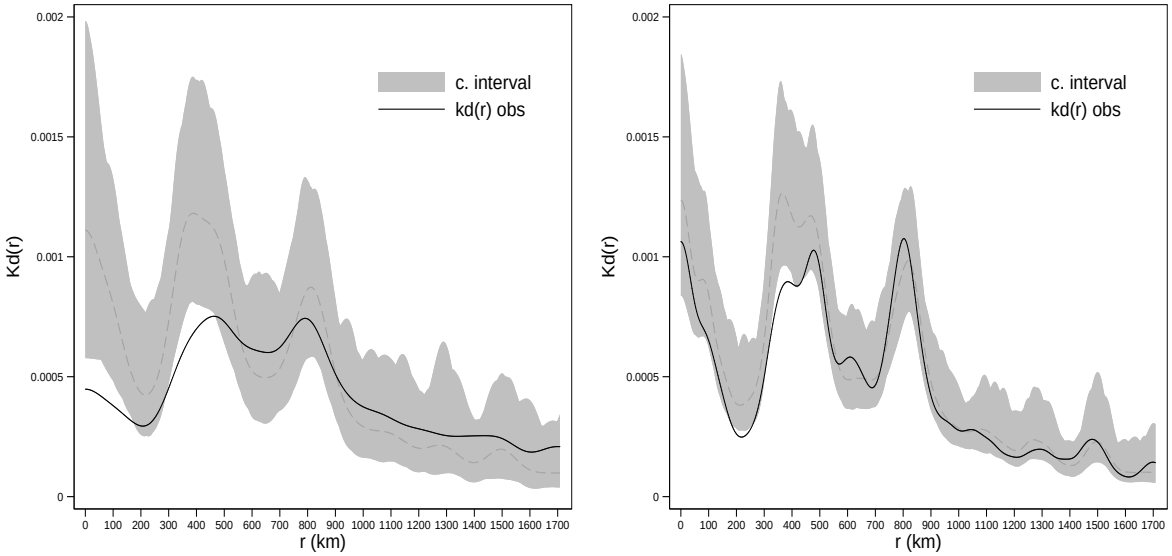
	Localization				Colocalization			
	2007-2008		2013-2014		2007-2008		2013-2014	
	# of ind.	%	# of ind.	%	# of ind.	%	# of ind.	%
Localized	46	52.87	39	50	56	64.37	40	51.28
Dispersed	12	13.79	9	11.54	11	12.64	12	15.38
Random	29	33.33	30	38.46	20	22.99	26	33.33
	87[a]	100	78[b]	100	87	100	78	100

Notes: After the restrictions imposed by the minimum of 10 plants in each sector: [a] 16 industries were dropped and [b] 15 industries were dropped. Source: Prepared by the author based on estimates.



(a) Localization (b) Colocalization

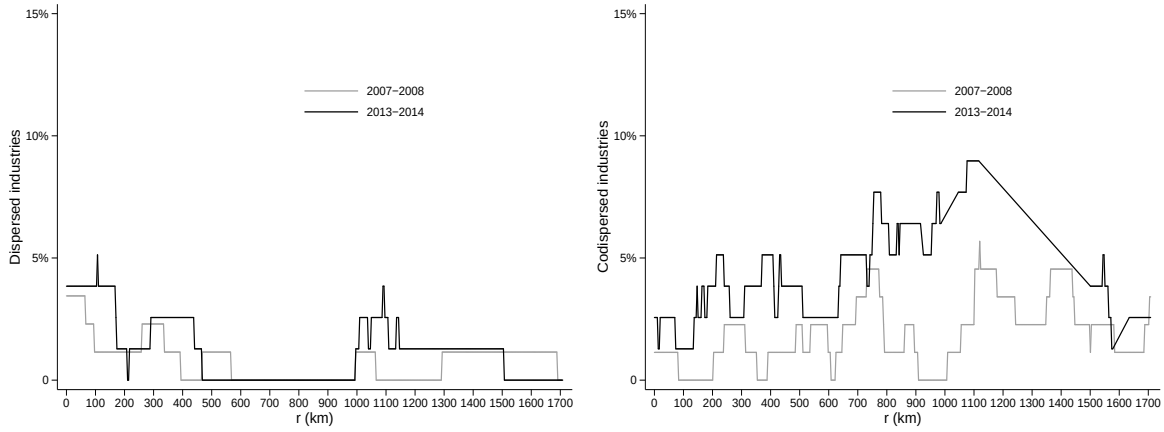
Printing activity - CNAE 181



(c) Localization (d) Colocalization

Forging and metal treatment - CNAE 253

Figure B.3 K-density estimates for printing and metal treatment activities



(a) Significant dispersion by distance (b) Significant codispersion by distance

Figure B.4 Share of industries for which entrants are dispersed and codispersed with existing establishments

Table B.5 Localization measurement (weighted and unweighted by employment) by industry 2006-2012

	Group (three digits CNAE 2.0)	2006					2012				
		New	**Weighted**		**Unweighted**		New	**Weighted**		**Unweighted**	
		plants	$\Gamma_{j,2006}$	$\Psi_{j,2006}$	$\Gamma_{j,2006}$	$\Psi_{j,2006}$	plants	$\Gamma_{j,2012}$	$\Psi_{j,2012}$	$\Gamma_{j,2012}$	$\Psi_{j,2012}$
112	M. of non-alcoholic beverages	63	0	0	0	0	56	0.00696	0	0	0.00013
321	M. of jewellery and related articles	173	0	0	0	0	176	0.00482	0	0.00103	0
285	M. of machinery for use in mineral extraction	31	0	0	0	0	18	0.00132	0	0	0
265	M. of measuring, testing and control devices	65	0	0	0.00146	0	44	0.00119	0	0.01130	0
309	M. of transport equipment not otherwise specified	45	0	0	0	0	42	0.00035	0	0	0
212	M. of pharmaceutical products	31	0	0	0.00364	0	20	0.00034	0	0	0
325	M. of instruments for medical use	238	0	0	0.00284	0	457	0.00017	0	0.00447	0
329	M. of miscellaneous products	564	0	0	0.00946	0	450	0.00016	0	0.01158	0
233	M. of articles of concrete	859	0	0.00399	0.00666	0	947	0.00005	0	0.00468	0
273	M. of equip. for distribution of electrical energy	105	0	0	0.00145	0	82	0.00001	0	0.00246	0
253	Forging and metal treatment services	641	0	0.00051	0.00967	0	727	0	0.00929	0.00977	0
153	Footwear manufacturing	770	0.00688	0	0.00012	0	744	0	0.00711	0.00012	0
284	M. of machine tools	105	0	0	0.00004	0	45	0	0.00401	0	0.00050
221	M. of rubber products	184	0	0	0.00077	0	93	0	0.00288	0.00217	0
141	M. of wearing apparel and accessories	4607	0	0.00128	0	0.00209	5299	0	0.00062	0.00012	0
135	M. of textile articles, except apparel	467	0	0	0.00616	0	430	0	0.00005	0.00154	0
111	M. of alcoholic beverages	61	0	0	0	0	71	0	0.00002	0	0
239	M. of other non-metallic mineral products	564	0	0	0.00758	0	676	0	0	0.00308	0
103	M. of tinned fruit, vegetables and other vegetables	169	0.02167	0	0	0	152	0	0	0	0
242	Steel	35	0.00834	0	0.00008	0	26	0	0	0	0
106	Grinding, M. of starch products and animal feed	330	0.00242	0	0.00098	0	262	0	0	0	0
263	M. of communication equipment	20	0.00157	0	0	0	16	0	0	0	0
234	M. of ceramic products	495	0.00107	0	0.00889	0	276	0	0	0.00703	0
282	M. of general-purpose machinery and equipment	388	0.00041	0	0.00003	0	255	0	0	0.00157	0
105	Dairy products	464	0.00014	0	0.00413	0	351	0	0	0.00013	0
206	M. of soaps and personal care products	213	0.00012	0	0.00043	0	115	0	0	0	0.00039
259	M. of metal products not otherwise specified	662	0.00006	0	0.00194	0	504	0	0	0.00542	0
332	Installation of machinery and equipment	337	0.00003	0	0	0	769	0	0	0.00162	0
142	M. of knitted and crocheted articles	143	0	0.00202	0.00024	0	88	0	0	0	0
154	M. of parts for footwear, of any material	145	0	0.00145	0.00751	0	245	0	0	0	0.00021
281	M. of engines and transmission equipment	102	0	0.00058	0.00213	0	62	0	0	0	0.00119
254	M. of cutlery, locksmiths' wares and tools	585	0	0.00023	0	0	743	0	0	0.00329	0

Continued on next page

Table B.5 – continued from previous page

	Group (three digits CNAE 2.0)	2006					2012				
		New	**Weighted**		**Unweighted**		New	**Weighted**		**Unweighted**	
		plants	$\Gamma_{j,2006}$	$\Psi_{j,2006}$	$\Gamma_{j,2006}$	$\Psi_{j,2006}$	plants	$\Gamma_{j,2012}$	$\Psi_{j,2012}$	$\Gamma_{j,2012}$	$\Psi_{j,2012}$
294	M. of parts and accessories for motor vehicles	213	0	0	0.00034	0	137	0	0	0.00032	0
109	M. of other food products	1890	0	0	0.00149	0	2567	0	0	0.00193	0
181	Print activity	822	0	0	0.00235	0	702	0	0	0.00047	0
108	Coffee roasting and grinding	59	0	0	0	0	47	0	0	0.00165	0
133	M. of knitted and crocheted fabrics	76	0	0	0.01711	0	33	0	0	0.00035	0
134	Finishing of yarns, fabrics and textile articles	291	0	0	0	0.00027	317	0	0	0.00051	0
201	M. of inorganic chemicals	86	0	0	0	0.00002	85	0	0	0	0
244	Non-ferrous metal metallurgy	82	0	0	0	0	55	0	0	0.00413	0
274	M. of lamps and other lighting equipment	59	0	0	0	0.00001	24	0	0	0.00008	0
152	M. of travel goods and miscellaneous leather goods	207	0	0	0	0.00001	177	0	0	0.00034	0
251	M. of metal structures	1064	0	0	0.00523	0	1152	0	0	0.00089	0
173	M. of paper and corrugated board packaging	151	0	0	0	0	111	0	0	0.00073	0
151	Tanning and other leather preparations	35	0	0	0	0	31	0	0	0	0
245	Foundry	114	0	0	0.00007	0	59	0	0	0	0
104	M. of vegetable and animal oils and fats	44	0	0	0.00491	0	15	0	0	0.00063	0
324	M. of toys and recreational games	33	0	0	0	0	43	0	0	0	0
174	M. of miscellaneous paper	135	0	0	0	0.00378	107	0	0	0	0
295	Recovery of engines for motor vehicles	127	0	0	0	0.00130	96	0	0	0	0
323	M. of fishing and sporting goods	35	0	0	0	0	38	0	0	0	0
262	M. of computer and peripheral equipment	53	0	0	0.00001	0	16	0	0	0	0
286	M. of machinery for industrial uses	311	0	0	0.00262	0	215	0	0	0.00036	0
252	M. of tanks, metal containers and boilers	50	0	0	0	0	30	0	0	0	0.00104
271	M. of electric generators, transformers and motors	31	0	0	0.02559	0	25	0	0	0	0
209	M. of miscellaneous chemical products	130	0	0	0.00062	0	63	0	0	0	0.00067
279	M. of apparatus not otherwise specified	73	0	0	0.00752	0	49	0	0	0.00037	0
231	M. of glass and glass products	89	0	0	0	0.00008	95	0	0	0.00038	0
283	M. of agricultural machinery and equipment	133	0	0	0.00253	0	98	0	0	0	0
203	M. of resins and elastomers	17	0	0	0	0	15	0	0	0	0
293	M. of motor vehicle cabins, bodies and trailers	161	0	0	0.00114	0	118	0	0	0	0
161	Wood unfolding	525	0	0	0.00221	0	379	0	0	0	0
301	Shipbuilding	45	0	0	0	0.00052	36	0	0	0	0.00183
331	Repair of machinery and equipment	1148	0	0	0.00089	0	1872	0	0	0	0
162	M. of products of wood except furniture	640	0	0	0	0	546	0	0	0.00314	0
202	M. of organic chemicals	36	0	0	0	0	13	0	0	0	0

Continued on next page

Table B.5 – continued from previous page

	Group (three digits CNAE 2.0)	2006					2012				
		New	**Weighted**		**Unweighted**		New	**Weighted**		**Unweighted**	
		plants	$\Gamma_{j,2006}$	$\Psi_{j,2006}$	$\Gamma_{j,2006}$	$\Psi_{j,2006}$	plants	$\Gamma_{j,2012}$	$\Psi_{j,2012}$	$\Gamma_{j,2012}$	$\Psi_{j,2012}$
310	Furniture manufacturing	1479	0	0	0.00610	0	2193	0	0	0.00964	0
182	Pre-press services and graphic finishing	285	0	0	0.00031	0	175	0	0	0	0
131	Preparation and spinning of textile fibres	50	0	0	0	0.00136	32	0	0	0	0
222	M. of plastic products	873	0	0	0.00014	0	586	0	0	0	0.00995
264	M. of reception and apparatus for audio and video	22	0	0	0	0	22	0	0	0	0
261	M. of electronic components	90	0	0	0.00271	0	36	0	0	0.00079	0
102	Preservation of fish and M. of fish products	29	0	0	0	0	25	0	0	0	0
207	M. of paints and related products	91	0	0	0	0.00004	71	0	0	0.00141	0
101	Slaughtering and production of meat products	326	0	0	0.00011	0	252	0	0	0.00142	0
172	M. of paper, paperboard and paperboard	19	0	0	0	0	17	0	0	0.00889	0
275	M. of household appliances	30	0	0	0	0	28	0	0	0	0
132	Weaving, not knitted or crocheted	63	0	0	0.00151	0	44	0	0	0	0
107	M. and refining of sugar	27	0.04674	0	0.02515	0	-	-	-	-	-
243	Production of steel tubes other than seamless tubes	17	0.02720	0	0.03205	0	-	-	-	-	-
266	M. of electromedical equipment	21	0	0	0	0	-	-	-	-	-
193	M. of biofuels	41	0	0	0	0.00156	-	-	-	-	-
122	M. of tobacco products	11	0	0	0	0	-	-	-	-	-
241	Production of pig iron and ferroalloys	12	0	0	0	0	-	-	-	-	-
192	M. of petroleum products	14	0	0	0	0	-	-	-	-	-
205	M. of pesticides and household disinfectants	16	0	0	0	0	-	-	-	-	-
183	Reproduction of recorded materials on any medium	11	0	0	0	0	-	-	-	-	-

Source: Prepared by the author based on estimates.

Table B.6 Colocalization measurement (weighted and unweighted by employment) by industry 2006-2012

	Group (three digits CNAE 2.0)	2006					2012				
		New	**Weighted**		**Unweighted**		New	**Weighted**		**Unweighted**	
		plants	$\Gamma_{j,2006}$	$\Psi_{j,2006}$	$\Gamma_{j,2006}$	$\Psi_{j,2006}$	plants	$\Gamma_{j,2012}$	$\Psi_{j,2012}$	$\Gamma_{j,2012}$	$\Psi_{j,2012}$
212	M. of pharmaceutical products	31	0.01697	0	0.00023	0	20	0.02409	0	0	0
321	M. of jewellery and related articles	173	0	0	0.00007	0	176	0.00513	0	0.00012	0
109	M. of other food products	1890	0.00697	0	0.00072	0	2567	0.00469	0	0.00080	0
162	M. of products of wood except furniture	640	0.00010	0	0	0	546	0.00372	0	0.00209	0
273	M. of equip. for distribution of electrical energy	105	0	0.00002	0.00032	0	82	0.00127	0	0.00052	0
281	M. of engines and transmission equipment	102	0	0	0.00223	0	62	0.00125	0	0	0.00054
182	Pre-press services and graphic finishing	285	0.00005	0	0.00021	0	175	0.00121	0	0	0
103	M. of tinned fruit, vegetables and other vegetables	169	0.00934	0	0	0	152	0.00100	0	0	0
283	M. of agricultural machinery and equipment	133	0.00025	0	0.00081	0	98	0.00098	0	0	0
329	M. of miscellaneous products	564	0.00034	0	0.00525	0	450	0.00086	0	0.00427	0
209	M. of miscellaneous chemical products	130	0	0	0	0	63	0.00067	0	0	0.00197
310	Furniture manufacturing	1479	0	0	0.00363	0	2193	0.00058	0	0.00467	0
102	Preservation of fish and M. of fish products	29	0	0	0.00004	0	25	0.00034	0	0	0.00011
112	M. of non-alcoholic beverages	63	0	0	0	0.00009	56	0.00019	0	0	0.00037
106	Grinding, M. of starch products and animal feed	330	0.00240	0	0.00012	0	262	0.00014	0	0	0
262	M. of computer and peripheral equipment	53	0	0.00268	0.00564	0	16	0.00012	0	0	0
293	M. of motor vehicle cabins, bodies and trailers	161	0.00162	0	0.00026	0	118	0.00008	0	0	0
279	M. of apparatus not otherwise specified	73	0.00122	0	0.00032	0	49	0.00004	0	0	0
134	Finishing of yarns, fabrics and textile articles	291	0.00068	0	0	0	317	0.00003	0	0.00009	0
154	M. of parts for footwear, of any material	145	0.00064	0	0.00559	0	245	0.00001	0	0.00092	0
101	Slaughtering and production of meat products	326	0.00042	0	0.00005	0	252	0.00001	0	0.00014	0
153	Footwear manufacturing	770	0.00397	0	0.00008	0	744	0	0.01692	0.00014	0
233	M. of articles of concrete	859	0	0.00241	0.00398	0	947	0	0.00723	0.00302	0
264	M. of reception and apparatus for audio and video	22	0	0	0.00001	0	22	0	0.00718	0	0
141	M. of wearing apparel and accessories	4607	0	0.00324	0	0.00105	5299	0	0.00414	0.00005	0
181	Print activity	822	0.00052	0	0.00058	0	702	0	0.00305	0	0.00251
221	M. of rubber products	184	0.00014	0	0.00010	0	93	0	0.00234	0.00001	0
253	Forging and metal treatment services	641	0.00116	0	0.00479	0	727	0	0.00212	0.00612	0
222	M. of plastic products	873	0	0	0.00043	0	586	0	0.00210	0.00001	0
133	M. of knitted and crocheted fabrics	76	0	0.00049	0.00288	0	33	0	0.00128	0	0.00017
234	M. of ceramic products	495	0.00058	0	0.00736	0	276	0	0.00091	0.00287	0
286	M. of machinery for industrial uses	311	0	0	0.00099	0	215	0	0.00085	0.00078	0

Continued on next page

Table B.6 – continued from previous page

	Group (three digits CNAE 2.0)	2006					2012				
		New	**Weighted**		**Unweighted**		New	**Weighted**		**Unweighted**	
		plants	$\Gamma_{j,2006}$	$\Psi_{j,2006}$	$\Gamma_{j,2006}$	$\Psi_{j,2006}$	plants	$\Gamma_{j,2012}$	$\Psi_{j,2012}$	$\Gamma_{j,2012}$	$\Psi_{j,2012}$
131	Preparation and spinning of textile fibres	50	0.00005	0	0	0.00303	32	0	0.00079	0	0
309	M. of transport equipment not otherwise specified	45	0	0.00010	0.00025	0	42	0	0.00067	0	0
206	M. of soaps and personal care products	213	0.00195	0	0	0	115	0	0.00055	0	0.00091
285	M. of machinery for use in mineral extraction	31	0.00085	0	0	0.00007	18	0	0.00033	0.00002	0
132	Weaving, not knitted or crocheted	63	0	0	0.00147	0	44	0	0.00024	0.00085	0
174	M. of miscellaneous paper	135	0	0.00360	0	0.00314	107	0	0.00021	0.00047	0
239	M. of other non-metallic mineral products	564	0.00046	0	0.00366	0	676	0	0.00020	0.00304	0
135	M. of textile articles, except apparel	467	0	0	0.00473	0	430	0	0.00020	0.00090	0
245	Foundry	114	0	0	0.00034	0	59	0	0.00018	0	0
259	M. of metal products not otherwise specified	662	0.00217	0	0.00109	0	504	0	0.00001	0.00342	0
201	M. of inorganic chemicals	86	0	0	0	0	85	0	0	0	0
251	M. of metal structures	1064	0	0.00009	0.00206	0	1152	0	0	0.00059	0
295	Recovery of engines for motor vehicles	127	0	0	0	0.00113	96	0	0	0	0
254	M. of cutlery, locksmiths' wares and tools	585	0.00466	0	0	0	743	0	0	0.00230	0
172	M. of paper, paperboard and paperboard	19	0.00134	0	0	0	17	0	0	0.00859	0
104	M. of vegetable and animal oils and fats	44	0.00012	0	0.00161	0	15	0	0	0	0
244	Non-ferrous metal metallurgy	82	0.00430	0	0	0	55	0	0	0.00140	0
323	M. of fishing and sporting goods	35	0.00502	0	0	0.00142	38	0	0	0	0
231	M. of glass and glass products	89	0.00179	0	0.00009	0	95	0	0	0.00001	0
325	M. of instruments for medical use	238	0	0	0.00130	0	457	0	0	0.00171	0
332	Installation of machinery and equipment	337	0	0	0.00007	0	769	0	0	0.00115	0
161	Wood unfolding	525	0	0.00086	0.00076	0	379	0	0	0	0
207	M. of paints and related products	91	0.00036	0	0	0.00001	71	0	0	0.00154	0
151	Tanning and other leather preparations	35	0.00204	0	0	0	31	0	0	0	0
274	M. of lamps and other lighting equipment	59	0	0.00003	0	0	24	0	0	0	0
271	M. of electric generators, transformers and motors	31	0	0	0.00123	0	25	0	0	0.00004	0
261	M. of electronic components	90	0	0.00270	0	0	36	0	0	0.00009	0
282	M. of general-purpose machinery and equipment	388	0.00023	0	0.00055	0	255	0	0	0.00042	0
252	M. of tanks, metal containers and boilers	50	0	0	0	0	30	0	0	0.00012	0
275	M. of household appliances	30	0	0	0	0.00025	28	0	0	0	0
105	Dairy products	464	0.00055	0	0.00403	0	351	0	0	0.00001	0
173	M. of paper and corrugated board packaging	151	0	0.00009	0	0	111	0	0	0.00001	0
284	M. of machine tools	105	0	0	0	0	45	0	0	0	0.00280
331	Repair of machinery and equipment	1148	0.00013	0	0.00042	0	1872	0	0	0	0

Continued on next page

Table B.6 – continued from previous page

	Group (three digits CNAE 2.0)	2006					2012			
		New	**Weighted**		**Unweighted**		New	**Weighted**		**Unweighted**	
		plants	$\Gamma_{j,2006}$	$\Psi_{j,2006}$	$\Gamma_{j,2006}$	$\Psi_{j,2006}$	plants	$\Gamma_{j,2012}$	$\Psi_{j,2012}$	$\Gamma_{j,2012}$	$\Psi_{j,2012}$
108	Coffee roasting and grinding	59	0	0	0	0	47	0	0	0	0
203	M. of resins and elastomers	17	0	0	0	0	15	0	0	0	0
142	M. of knitted and crocheted articles	143	0	0.00446	0.00134	0	88	0	0	0	0.00009
202	M. of organic chemicals	36	0	0.00027	0	0	13	0	0	0	0
242	Steel	35	0	0	0.00131	0	26	0	0	0	0
152	M. of travel goods and miscellaneous leather goods	207	0	0	0	0	177	0	0	0.00127	0
294	M. of parts and accessories for motor vehicles	213	0	0.00191	0.00029	0	137	0	0	0.00041	0
301	Shipbuilding	45	0	0	0.00120	0	36	0	0	0	0.00491
263	M. of communication equipment	20	0	0	0	0.00013	16	0	0	0	0.00001
324	M. of toys and recreational games	33	0.00022	0	0	0	43	0	0	0	0
265	M. of measuring, testing and control devices	65	0	0	0.00217	0	44	0	0	0	0.00310
111	M. of alcoholic beverages	61	0.00068	0	0.00027	0	71	0	0	0	0
192	M. of petroleum products	14	0	0	0	0	-	-	-	-	-
266	M. of electromedical equipment	21	0.00273	0	0.00040	0	-	-	-	-	-
241	Production of pig iron and ferroalloys	12	0	0	0	0	-	-	-	-	-
107	M. and refining of sugar	27	0	0	0	0	-	-	-	-	-
193	M. of biofuels	41	0	0.00005	0	0.00324	-	-	-	-	-
243	Production of steel tubes other than seamless tubes	17	0	0.00001	0.00510	0	-	-	-	-	-
205	M. of pesticides and household disinfectants	16	0	0	0	0	-	-	-	-	-
183	Reproduction of recorded materials on any medium	11	0	0.00001	0	0	-	-	-	-	-
122	M. of tobacco products	11	0	0	0	0	-	-	-	-	-

Source: Prepared by the author based on estimates.

B.3 Regression analysis: additional results

Table B.7 Other key industries: localization effects

	(1)	(2)	(3)	(4)	(5)
	Apparel CNAE 141	Machinery Repair CNAE 331	Metallic Structures CNAE 251	Footwear CNAE 153	Machinery CNAE 282
Panel A: The dependent variable is births of new establishments					
0 to 1 km	3.35e-04***	4.58e-04***	2.04e-04**	4.01e-04***	1.08e-03***
	(1.99e-05)	(1.05e-04)	(9.00e-05)	(2.69e-05)	(2.32e-04)
1 to 5 km	2.71e-05**	3.34e-05	1.04e-04**	4.73e-05***	-9.69e-05
	(1.14e-05)	(5.09e-05)	(4.89e-05)	(1.73e-05)	(1.37e-04)
5 to 10 km	5.37e-06	-9.68e-05**	2.03e-05	-5.86e-06	-2.64e-05
	(9.09e-06)	(3.98e-05)	(5.16e-05)	(1.62e-05)	(1.16e-04)
10 to 20 km	-9.54e-06	-1.35e-05	-2.82e-05	-1.22e-05	-1.62e-05
	(8.11e-06)	(2.57e-05)	(5.40e-05)	(1.62e-05)	(9.31e-05)
20 to 40 km	-1.74e-05***	-4.26e-05*	-2.01e-04***	7.55e-06	6.33e-05
	(6.52e-06)	(2.31e-05)	(5.67e-05)	(1.61e-05)	(7.33e-05)
	Average Change in Localization Effect per km				
0.5 to 3 km	-1.23e-04	-1.70e-04	-3.99e-05	-1.42e-04	-4.72e-04
3 to 7.5 km	-4.82e-06	-2.89e-05	-1.86e-05	-1.18e-05	1.57e-05
7.5 to 15 km	-1.99e-06	1.11e-05	-6.47e-06	-8.51e-07	1.37e-06
15 to 30 km	-5.25e-07	-1.94e-06	-1.15e-05	1.32e-06	5.30e-06
# of district FE	2,009	1,337	1,452	426	501
Pseudo R^2	0.2959	0.0815	0.0699	0.4439	0.0925
Pseudo-LL	-47,279.58	-22,967.6	-17,550.86	-7,008.711	-5,679.481
Observations	120,892	107,496	107,141	42,202	67,344
Panel B: The dependent variable is new-establishment employment					
0 to 1 km	4.26e-04***	1.26e-03***	-3.25e-05	6.73e-04***	1.17e-03*
	(4.40e-05)	(3.28e-04)	(1.88e-04)	(1.03e-04)	(6.16e-04)
1 to 5 km	4.24e-05*	-8.61e-04**	4.04e-06	1.24e-04**	4.67e-06
	(2.37e-05)	(3.66e-04)	(7.69e-05)	(5.09e-05)	(4.25e-04)
5 to 10 km	2.85e-05	-5.76e-04**	-4.86e-05	2.75e-05	-3.43e-04
	(2.16e-05)	(2.48e-04)	(7.29e-05)	(4.11e-05)	(3.48e-04)
10 to 20 km	1.40e-05	-1.35e-04	-3.41e-04**	8.65e-05*	-4.12e-04
	(1.85e-05)	(1.62e-04)	(1.52e-04)	(4.48e-05)	(2.65e-04)
20 to 40 km	-7.15e-06	9.91e-05	-3.28e-04**	7.82e-05*	1.17e-04
	(1.52e-05)	(1.11e-04)	(1.58e-04)	(4.01e-05)	(1.37e-04)
	Average Change in Localization Effect per km				
0.5 to 3 km	-1.53e-04	-8.49e-04	1.46e-05	-2.19e-04	-4.66e-04
3 to 7.5 km	-3.09e-06	6.34e-05	-1.17e-05	-2.14e-05	-7.73e-05
7.5 to 15 km	-1.93e-06	5.88e-05	-3.90e-05	7.87e-06	-9.13e-06
15 to 30 km	-1.41e-06	1.56e-05	8.61e-07	-5.53e-07	3.52e-05
# of district FE	1,696	1,012	1,034	318	325
Pseudo R^2	0.3514	0.2624	0.1958	0.4563	0.2077
Pseudo-LL	-226,917.3	-85,603.98	-46,975.85	-90,018.62	-21,469.72
Observations	112,724	94,731	90,548	32,969	51,607

Notes: This table reports the localization effects in other key industries. All columns report the results of Poisson regressions where the dependent variable is the births of new establishments (Panel A) and the new-establishment employment (Panel B) and the variable of interest is the number of workers in the same industry in each concentric ring. All columns include the diversification and competition control variables, transport and geographic controls, municipal level controls, and district fixed effects. The transportation controls include the distance to the nearest airport, public port, railway, federal highway, and state highway interacted with time effects. The geographic control is the distance to the nearest river interacted with time effects. The municipal level controls include proxies for insertion in international trade (exports and imports), municipal taxes, capital investments, housing and town planning expenses, homicides and traffic fatalities. Change per kilometer is computed by subtracting the adjacent localization coefficients and dividing by the number of kilometers between the midpoints. Significance level: $^*p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: Prepared by the author based on estimates.

Table B.8 Poisson regression without district fixed effects and without controls - plant birth

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pharma. Products CNAE 212	Food Products CNAE 109	Fruit & vegetable Canning CNAE 103	Starch Products CNAE 106	Furniture CNAE 310	Prepress Services CNAE 182	Wood Products CNAE 162	Fin. of textile Artifacts CNAE 134
				Localization Effects				
0 to 1 km	1.28e-03***	4.27e-04***	1.48e-03***	1.98e-03***	6.58e-04***	3.49e-03***	2.48e-03***	2.51e-03***
	(2.21e-04)	(3.70e-05)	(2.64e-04)	(1.51e-04)	(5.46e-05)	(4.34e-04)	(1.68e-04)	(3.64e-04)
1 to 5 km	3.08e-04**	2.22e-04***	6.08e-04***	5.03e-04***	2.23e-04***	8.52e-04***	7.93e-04***	1.33e-03***
	(1.26e-04)	(1.34e-05)	(1.64e-04)	(1.42e-04)	(1.27e-05)	(1.61e-04)	(6.86e-05)	(1.11e-04)
5 to 10 km	2.05e-04*	1.13e-04***	6.31e-05	9.56e-05	2.09e-04***	-4.80e-07	2.89e-04***	2.53e-05
	(1.10e-04)	(1.21e-05)	(1.98e-04)	(1.27e-04)	(1.73e-05)	(1.36e-04)	(1.06e-04)	(1.11e-04)
10 to 20 km	1.03e-04	6.34e-05***	3.76e-04***	-9.27e-05	8.65e-05***	3.21e-04***	-2.15e-04***	6.73e-04***
	(6.62e-05)	(9.48e-06)	(1.30e-04)	(9.92e-05)	(1.28e-05)	(8.69e-05)	(7.58e-05)	(4.62e-05)
20 to 40 km	5.20e-05	-1.00e-05	1.12e-04	8.69e-05	8.04e-05***	-1.86e-04**	2.25e-04***	6.45e-05**
	(6.04e-05)	(9.83e-06)	(1.18e-04)	(8.29e-05)	(6.92e-06)	(9.42e-05)	(3.53e-05)	(3.02e-05)
				Urbanization Effects				
0 to 1 km	1.11e-04***	8.09e-05***	1.05e-04***	1.32e-04***	9.88e-05***	4.74e-05***	1.33e-04***	8.06e-05***
	(2.91e-05)	(5.75e-06)	(1.70e-05)	(1.20e-05)	(6.47e-06)	(1.02e-05)	(8.63e-06)	(1.24e-05)
1 to 5 km	1.07e-05	5.04e-06***	5.71e-06	-5.21e-06	2.25e-06	9.25e-06***	-7.12e-06**	6.74e-06**
	(1.16e-05)	(1.46e-06)	(5.50e-06)	(5.40e-06)	(1.53e-06)	(2.82e-06)	(3.58e-06)	(3.18e-06)
5 to 10 km	-1.85e-07	-9.10e-07	5.33e-06	-9.38e-06**	-1.12e-06	1.35e-06	-5.92e-06**	4.29e-06**
	(4.83e-06)	(1.05e-06)	(3.36e-06)	(4.50e-06)	(9.91e-07)	(1.75e-06)	(2.35e-06)	(1.96e-06)
10 to 20 km	-9.04e-06**	-3.69e-06***	-2.95e-06	6.17e-06***	-2.52e-06***	-1.40e-06	2.62e-06***	-3.35e-06***
	(4.18e-06)	(5.93e-07)	(1.85e-06)	(1.61e-06)	(5.23e-07)	(1.51e-06)	(9.07e-07)	(1.18e-06)
20 to 40 km	-5.47e-07	-1.81e-06***	-1.97e-06**	-6.50e-06***	-3.23e-06***	4.17e-07	-3.45e-06***	-1.50e-06***
	(2.14e-06)	(4.25e-07)	(8.50e-07)	(1.39e-06)	(2.43e-07)	(1.01e-06)	(4.77e-07)	(4.57e-07)
Pseudo R^2	0.0475	0.0231	0.0117	0.0189	0.0377	0.0867	0.0267	0.0766
Pseudo-LL	-738.4323	-28,198.32	-3,411.104	-5,444.411	-26,714.01	-5,593.45	-10,478.48	-6,355.135

Notes: This table reports the localization and urbanization effects when the dependent variable is the number of new establishments in each cell. All models are estimated with 132,072 observations. Heteroscedasticity-robust standard errors are reported in parentheses. Significance level: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: Prepared by the author based on estimates.

Table B.9 Poisson regression without district fixed effects and without controls - new employment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pharma. Products CNAE 212	Food Products CNAE 109	Fruit & vegetable Canning CNAE 103	Starch Products CNAE 106	Furniture CNAE 310	Prepress Services CNAE 182	Wood Products CNAE 162	Fin. of textile Articles CNAE 134
Localization Effects								
0 to 1 km	1.94e-03***	9.02e-04***	2.53e-03***	3.24e-03***	1.00e-03***	5.62e-03***	3.92e-03***	2.98e-03***
	(2.60e-04)	(9.61e-05)	(4.32e-04)	(3.02e-04)	(9.10e-05)	(1.66e-03)	(5.04e-04)	(6.83e-04)
1 to 5 km	3.06e-04	3.61e-04***	7.67e-04	1.32e-03***	1.93e-04***	9.55e-04**	1.17e-03***	1.53e-03***
	(2.56e-04)	(5.40e-05)	(5.02e-04)	(3.11e-04)	(4.88e-05)	(4.79e-04)	(4.45e-04)	(2.93e-04)
5 to 10 km	3.28e-04**	1.63e-04*	4.83e-04	2.66e-04	2.63e-04***	-5.14e-04	9.28e-04**	-8.38e-05
	(1.43e-04)	(8.81e-05)	(5.26e-04)	(2.35e-04)	(5.70e-05)	(6.50e-04)	(3.95e-04)	(2.22e-04)
10 to 20 km	1.95e-04**	1.84e-04*	8.38e-04**	2.81e-04*	1.98e-04***	1.98e-04	-4.33e-04**	6.92e-04***
	(9.17e-05)	(1.06e-04)	(3.52e-04)	(1.46e-04)	(5.69e-05)	(3.76e-04)	(1.91e-04)	(6.86e-05)
20 to 40 km	9.41e-05	-1.07e-04	6.29e-04**	3.30e-04*	8.35e-05***	-1.14e-03**	3.09e-04*	3.01e-05
	(1.81e-04)	(8.59e-05)	(2.85e-04)	(1.78e-04)	(2.35e-05)	(4.50e-04)	(1.60e-04)	(6.44e-05)
Urbanization Effects								
0 to 1 km	9.80e-05*	1.29e-04***	6.89e-05	1.51e-04**	1.32e-04***	7.63e-05***	1.85e-04***	7.67e-05***
	(5.78e-05)	(2.15e-05)	(1.31e-04)	(5.87e-05)	(1.05e-05)	(2.92e-05)	(1.62e-05)	(2.43e-05)
1 to 5 km	-3.77e-06	-1.01e-05	1.19e-05	3.58e-06	-5.75e-06	3.63e-06	-5.36e-05**	1.68e-05***
	(1.45e-05)	(6.92e-06)	(2.14e-05)	(1.47e-05)	(6.64e-06)	(9.10e-06)	(2.34e-05)	(5.86e-06)
5 to 10 km	-2.93e-06	-1.99e-05*	-2.59e-05	-2.16e-05	-9.19e-06**	5.49e-06	5.36e-06	9.27e-06***
	(8.51e-06)	(1.10e-05)	(1.68e-05)	(1.88e-05)	(3.75e-06)	(7.28e-06)	(1.74e-05)	(3.37e-06)
10 to 20 km	-1.21e-05**	-5.76e-07	-1.85e-05*	8.84e-06*	-1.00e-06	2.59e-06	-2.79e-06	-7.66e-06***
	(5.79e-06)	(4.73e-06)	(9.63e-06)	(4.95e-06)	(2.21e-06)	(5.44e-06)	(7.58e-06)	(2.26e-06)
20 to 40 km	1.13e-06	1.56e-06	4.97e-07	-6.64e-06***	-3.92e-06***	6.85e-06**	-1.07e-07	-1.29e-06
	(5.32e-06)	(3.29e-06)	(1.76e-06)	(2.07e-06)	(8.70e-07)	(2.81e-06)	(2.91e-06)	(1.17e-06)
Pseudo R^2	0.1829	0.0627	0.0748	0.0721	0.1271	0.1284	0.1609	0.0891
Pseudo-LL	-38,296.45	-14,4717.5	-36,389.17	-49,815.39	-88,028.24	-18,097.02	-57,440.77	-23,685.46

Notes: This table reports the localization and urbanization effects when the dependent variable is the number of new employments in each cell. All models are estimated with 132,072 observations. Heteroscedasticity-robust standard errors are reported in parentheses. Significance level: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: Prepared by the author based on estimates.

Table B.10 Negative binomial regression without district fixed effects and without controls - plant birth

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pharma. Products CNAE 212	Food Products CNAE 109	Fruit & vegetable Canning CNAE 103	Starch Products CNAE 106	Furniture CNAE 310	Prepress Services CNAE 182	Wood Products CNAE 162	Fin. of textile Articles CNAE 134
				Localization Effects				
0 to 1 km	1.77e-03***	6.13e-04***	2.33e-03***	2.35e-03***	2.30e-03***	7.95e-03***	4.46e-03***	6.73e-03***
	(6.86e-04)	(7.15e-05)	(6.76e-04)	(2.88e-04)	(1.85e-04)	(1.62e-03)	(3.19e-04)	(1.13e-03)
1 to 5 km	2.36e-04	2.42e-04***	6.03e-04***	5.01e-04***	2.51e-04***	1.14e-03***	8.09e-04***	1.35e-03***
	(2.14e-04)	(1.59e-05)	(1.73e-04)	(1.47e-04)	(2.24e-05)	(1.97e-04)	(8.25e-05)	(1.51e-04)
5 to 10 km	1.99e-04**	1.11e-04***	3.86e-05	8.83e-05	2.02e-04***	-8.47e-05	1.84e-04*	8.07e-05
	(9.59e-05)	(1.28e-05)	(2.01e-04)	(1.30e-04)	(2.31e-05)	(1.61e-04)	(1.09e-04)	(1.26e-04)
10 to 20 km	9.13e-05	5.89e-05***	3.72e-04***	-9.48e-05	5.47e-05***	3.29e-04***	-2.13e-04***	6.86e-04***
	(6.67e-05)	(9.35e-06)	(1.31e-04)	(1.01e-04)	(1.42e-05)	(9.45e-05)	(7.66e-05)	(5.85e-05)
20 to 40 km	3.85e-05	-1.20e-05	1.31e-04	8.66e-05	7.36e-05***	-1.46e-04	2.05e-04***	3.18e-05
	(6.39e-05)	(9.91e-06)	(1.15e-04)	(8.38e-05)	(7.13e-06)	(9.73e-05)	(3.69e-05)	(3.52e-05)
				Urbanization Effects				
0 to 1 km	1.98e-04	1.33e-04***	2.12e-04**	1.62e-04***	1.54e-04***	1.01e-04***	2.11e-04***	3.25e-04***
	(2.37e-04)	(1.21e-05)	(9.98e-05)	(2.64e-05)	(1.35e-05)	(2.11e-05)	(2.58e-05)	(4.56e-05)
1 to 5 km	9.34e-06	4.40e-06***	5.32e-06	-7.00e-06	1.08e-06	9.77e-06***	-8.70e-06**	6.99e-06**
	(1.21e-05)	(1.51e-06)	(5.41e-06)	(6.01e-06)	(1.64e-06)	(3.20e-06)	(3.71e-06)	(3.34e-06)
5 to 10 km	9.34e-07	-1.34e-06	5.44e-06	-8.28e-06*	-8.45e-07	1.30e-06	-4.79e-06**	2.36e-06
	(4.85e-06)	(1.06e-06)	(3.36e-06)	(4.66e-06)	(1.03e-06)	(2.08e-06)	(2.27e-06)	(2.18e-06)
10 to 20 km	-8.26e-06*	-3.34e-06***	-3.05e-06	5.87e-06***	-2.05e-06***	-1.67e-06	2.31e-06**	-2.74e-06**
	(4.72e-06)	(5.72e-07)	(1.92e-06)	(1.71e-06)	(5.21e-07)	(1.55e-06)	(9.46e-07)	(1.15e-06)
20 to 40 km	-3.46e-07	-1.84e-06***	-2.00e-06**	-6.43e-06***	-3.12e-06***	-3.96e-08	-3.29e-06***	-1.62e-06***
	(2.14e-06)	(4.27e-07)	(8.50e-07)	(1.41e-06)	(2.45e-07)	(1.05e-06)	(4.87e-07)	(4.93e-07)
Pseudo R^2	0.0494	0.0221	0.0126	0.0191	0.0365	0.0832	0.0288	0.0753
Pseudo-LL	-736.9458	-27,922.64	-3,392.883	-5,432.249	-26,186.56	-5,521.758	-10,353.51	-6,137.321

Notes: This table reports the localization and urbanization effects when the dependent variable is the number of new establishments in each cell. All models are estimated with 132,072 observations. Heteroscedasticity-robust standard errors are reported in parentheses. Significance level: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: Prepared by the author based on estimates.

Table B.11 Negative binomial regression without district fixed effects and without controls - new employment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pharma. Products CNAE 212	Food Products CNAE 109	Fruit & vegetable Canning CNAE 103	Starch Products CNAE 106	Furniture CNAE 310	Prepress Services CNAE 182	Wood Products CNAE 162	Fin. of textile Articles CNAE 134
				Localization Effects				
0 to 1 km	9.46e-03***	9.21e-03***	2.07e-02***	2.18e-02***	1.06e-02***	3.83e-02***	1.74e-02***	2.56e-02***
	(2.84e-03)	(8.88e-04)	(7.47e-03)	(2.71e-03)	(1.05e-03)	(7.43e-03)	(2.04e-03)	(4.82e-03)
1 to 5 km	1.46e-04	3.32e-04***	1.35e-03	2.67e-03***	7.97e-04***	1.14e-03	2.05e-03***	3.15e-03***
	(6.07e-04)	(1.05e-04)	(1.45e-03)	(8.18e-04)	(1.73e-04)	(1.14e-03)	(5.27e-04)	(8.71e-04)
5 to 10 km	3.97e-04*	1.99e-04*	1.17e-03	3.68e-04	2.16e-04**	3.21e-03**	-8.33e-05	3.19e-04
	(2.27e-04)	(1.03e-04)	(7.10e-04)	(5.47e-04)	(9.32e-05)	(1.25e-03)	(3.28e-04)	(3.73e-04)
10 to 20 km	9.29e-05	1.08e-04***	5.19e-04	-1.17e-05	7.77e-05*	5.22e-04	-3.74e-04**	9.53e-04***
	(1.82e-04)	(2.75e-05)	(3.85e-04)	(1.98e-04)	(4.70e-05)	(5.70e-04)	(1.68e-04)	(2.41e-04)
20 to 40 km	3.53e-04**	4.27e-05	2.02e-03*	1.80e-04	6.06e-05*	-7.57e-04**	2.02e-04	-1.14e-04
	(1.53e-04)	(2.76e-05)	(1.16e-03)	(2.26e-04)	(3.53e-05)	(3.29e-04)	(1.23e-04)	(1.01e-04)
				Urbanization Effects				
0 to 1 km	8.75e-04*	3.24e-04***	9.78e-04***	6.06e-04***	4.50e-04***	2.77e-04**	3.82e-04***	7.61e-04***
	(5.12e-04)	(5.95e-05)	(3.45e-04)	(1.48e-04)	(6.37e-05)	(1.31e-04)	(7.53e-05)	(1.44e-04)
1 to 5 km	-1.10e-05	4.19e-06	7.80e-06	-1.41e-06	-1.24e-05**	5.99e-05***	-2.40e-05**	2.15e-05**
	(2.16e-05)	(5.88e-06)	(1.23e-05)	(1.79e-05)	(4.89e-06)	(1.56e-05)	(9.55e-06)	(1.02e-05)
5 to 10 km	1.87e-05	-1.12e-05***	2.66e-06	-2.30e-06	-8.57e-06***	-4.70e-05***	2.48e-06	-3.38e-07
	(1.79e-05)	(3.79e-06)	(9.51e-06)	(8.16e-06)	(2.99e-06)	(1.45e-05)	(6.03e-06)	(5.52e-06)
10 to 20 km	-2.09e-05**	-1.98e-06	-1.39e-05**	5.26e-06	-1.64e-06	-2.15e-06	3.70e-06*	-6.86e-06***
	(9.30e-06)	(1.57e-06)	(6.28e-06)	(3.33e-06)	(1.60e-06)	(6.83e-06)	(2.05e-06)	(2.23e-06)
20 to 40 km	-2.93e-06	-4.40e-06***	-3.58e-06	-4.32e-06**	-2.37e-06**	7.47e-06*	-3.86e-06***	-1.93e-07
	(6.76e-06)	(1.08e-06)	(2.58e-06)	(1.92e-06)	(9.68e-07)	(3.88e-06)	(1.08e-06)	(9.88e-07)
Pseudo R^2	0.0369	0.0235	0.0197	0.0171	0.0268	0.0412	0.0264	0.0314
Pseudo-LL	-587.1157	-28,551.9	-2,723.142	-4,607.305	-25,110.61	-4,070.659	-9,547.105	-5,930.928

Notes: This table reports the localization and urbanization effects when the dependent variable is the number of new employments in each cell. All models are estimated with 132,072 observations. Heteroscedasticity-robust standard errors are reported in parentheses. Significance level: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: Prepared by the author based on estimates.

Table B.12 Poisson regression with district fixed effects and without controls - plant birth

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pharma. Products CNAE 212	Food Products CNAE 109	Fruit & vegetable Canning CNAE 103	Starch Products CNAE 106	Furniture CNAE 310	Prepress Services CNAE 182	Wood Products CNAE 162	Fin. of textile Articles CNAE 134
				Localization Effects				
0 to 1 km	1.16e-03***	4.78e-04***	1.20e-03**	1.72e-03***	6.98e-04***	4.64e-03***	2.35e-03***	2.14e-03***
	(3.73e-04)	(5.15e-05)	(5.96e-04)	(3.78e-04)	(7.31e-05)	(4.69e-04)	(2.78e-04)	(5.16e-04)
1 to 5 km	2.35e-04	1.91e-04***	2.12e-04	-3.60e-04	1.45e-04***	1.10e-03***	2.68e-04	4.46e-04**
	(3.03e-04)	(3.67e-05)	(4.70e-04)	(3.42e-04)	(4.65e-05)	(1.81e-04)	(2.05e-04)	(2.00e-04)
5 to 10 km	2.11e-04	1.05e-04***	-2.48e-04	-4.23e-04	1.56e-05	2.31e-04	3.07e-04*	-3.64e-04**
	(1.93e-04)	(3.18e-05)	(4.53e-04)	(2.84e-04)	(4.85e-05)	(1.63e-04)	(1.72e-04)	(1.62e-04)
10 to 20 km	7.19e-05	7.20e-05***	-1.82e-04	-4.46e-04*	-4.81e-05	3.46e-04***	1.79e-04	6.05e-05
	(1.61e-04)	(2.76e-05)	(3.77e-04)	(2.47e-04)	(3.48e-05)	(1.24e-04)	(1.44e-04)	(9.47e-05)
20 to 40 km	1.44e-04	-1.01e-05	1.88e-04	-9.90e-05	2.97e-06	1.87e-05	-2.61e-05	-1.85e-04**
	(1.70e-04)	(2.36e-05)	(2.93e-04)	(2.16e-04)	(2.40e-05)	(1.50e-04)	(9.60e-05)	(7.97e-05)
				Urbanization Effects				
0 to 1 km	1.77e-04***	1.38e-04***	1.58e-04***	1.96e-04***	1.53e-04***	6.93e-05***	1.69e-04***	1.76e-04***
	(3.50e-05)	(8.33e-06)	(3.12e-05)	(2.18e-05)	(8.50e-06)	(1.41e-05)	(1.45e-05)	(1.59e-05)
1 to 5 km	3.42e-05**	2.22e-05***	1.43e-05	1.87e-05**	2.30e-05***	2.03e-05***	6.03e-06	2.27e-05***
	(1.63e-05)	(2.99e-06)	(9.56e-06)	(8.31e-06)	(2.66e-06)	(4.50e-06)	(5.14e-06)	(5.62e-06)
5 to 10 km	-6.02e-06	-4.42e-06**	4.70e-06	-5.44e-06	5.10e-06**	-8.66e-06***	-7.11e-06*	-1.57e-06
	(1.09e-05)	(2.15e-06)	(6.14e-06)	(6.49e-06)	(2.08e-06)	(3.25e-06)	(4.11e-06)	(3.89e-06)
10 to 20 km	-8.96e-06	-3.71e-06**	-2.76e-06	6.90e-06*	1.81e-06	-7.39e-06***	-3.59e-06	-4.60e-06**
	(8.88e-06)	(1.70e-06)	(5.07e-06)	(3.85e-06)	(1.39e-06)	(2.54e-06)	(2.35e-06)	(2.32e-06)
20 to 40 km	-3.66e-06	-7.12e-07	-3.74e-06	1.08e-06	3.97e-06***	-8.60e-06***	-5.50e-06***	7.22e-07
	(3.99e-06)	(1.33e-06)	(4.32e-06)	(3.63e-06)	(1.06e-06)	(2.31e-06)	(2.10e-06)	(1.82e-06)
# of district FE	81	1,904	391	670	1,590	398	968	430
Pseudo R^2	0.1126	0.0709	0.1102	0.1141	0.0968	0.1109	0.1099	0.1569
Pseudo-LL	-545.7164	-27,578.92	-2,793.952	-4,651.204	-25,133.54	-4,752.476	-9,259.15	-5,128.462
Observations	19,151	121,077	49,100	63,055	113,939	58,297	85,019	59,910

Notes: This table reports the localization and urbanization effects when the dependent variable is the number of new establishments in each cell. Heteroscedasticity-robust standard errors are reported in parentheses. All columns include the district fixed effects. Significance level: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: Prepared by the author based on estimates.

Table B.13 Poisson regression with district fixed effects and without controls - new employment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pharma. Products CNAE 212	Food Products CNAE 109	Fruit & vegetable Canning CNAE 103	Starch Products CNAE 106	Furniture CNAE 310	Prepress Services CNAE 182	Wood Products CNAE 162	Fin. of textile Artifacts CNAE 134
Localization Effects								
0 to 1 km	1.28e-03	1.42e-03***	7.77e-04	3.72e-03***	8.93e-04***	2.41e-03***	5.69e-03***	1.71e-03
	(8.14e-04)	(2.73e-04)	(9.50e-04)	(1.02e-03)	(1.81e-04)	(6.12e-04)	(7.43e-04)	(1.25e-03)
1 to 5 km	-2.34e-05	3.82e-04**	-1.50e-03	-1.02e-03	-4.79e-06	5.64e-04	1.59e-03**	1.32e-05
	(4.15e-04)	(1.72e-04)	(1.20e-03)	(8.39e-04)	(1.59e-04)	(6.03e-04)	(7.93e-04)	(4.23e-04)
5 to 10 km	5.84e-04***	2.10e-04	-2.96e-04	-1.34e-03	2.84e-05	-7.96e-04	5.48e-04	-6.09e-04*
	(1.74e-04)	(1.73e-04)	(1.03e-03)	(8.55e-04)	(1.20e-04)	(6.23e-04)	(7.95e-04)	(3.42e-04)
10 to 20 km	4.13e-04	1.27e-04	-1.56e-04	-1.52e-03*	4.57e-05	-1.05e-04	-4.69e-04	-8.71e-05
	(3.35e-04)	(1.30e-04)	(9.48e-04)	(8.04e-04)	(9.84e-05)	(2.76e-04)	(5.80e-04)	(2.46e-04)
20 to 40 km	5.92e-04	-3.29e-05	-2.46e-03**	-1.62e-04	9.86e-05	-1.12e-03**	-5.48e-04	-3.30e-05
	(4.72e-04)	(8.78e-05)	(1.23e-03)	(6.00e-04)	(1.12e-04)	(4.77e-04)	(3.90e-04)	(1.41e-04)
Urbanization Effects								
0 to 1 km	1.60e-04**	2.05e-04***	3.06e-04**	3.03e-04***	1.98e-04***	5.92e-05**	1.94e-04***	1.90e-04***
	(7.14e-05)	(2.97e-05)	(1.41e-04)	(7.57e-05)	(1.87e-05)	(2.33e-05)	(2.90e-05)	(5.45e-05)
1 to 5 km	7.93e-05	1.60e-05	6.35e-05*	6.58e-05***	2.50e-05***	1.88e-05	-1.23e-05	4.21e-05***
	(5.13e-05)	(1.09e-05)	(3.35e-05)	(1.84e-05)	(9.24e-06)	(1.22e-05)	(1.58e-05)	(1.40e-05)
5 to 10 km	-1.96e-05	-1.41e-05	6.36e-06	2.59e-05**	9.80e-07	8.05e-06	7.30e-06	1.05e-05
	(2.30e-05)	(1.20e-05)	(1.28e-05)	(1.07e-05)	(6.41e-06)	(9.33e-06)	(1.96e-05)	(9.49e-06)
10 to 20 km	6.71e-06	-1.02e-05	-2.71e-05*	3.95e-06	-3.20e-07	2.20e-06	1.28e-05*	-2.13e-05***
	(2.05e-05)	(1.14e-05)	(1.47e-05)	(1.47e-05)	(3.85e-06)	(6.34e-06)	(7.05e-06)	(4.94e-06)
20 to 40 km	1.58e-06	6.62e-07	-1.40e-05	2.14e-05**	2.75e-06	4.20e-07	1.46e-05**	-4.29e-07
	(1.58e-05)	(8.72e-06)	(1.33e-05)	(9.51e-06)	(4.19e-06)	(5.15e-06)	(7.34e-06)	(4.97e-06)
# of district FE	41	1,396	216	392	1,181	247	689	311
Pseudo R^2	0.3285	0.2271	0.4705	0.3081	0.2442	0.2364	0.3735	0.2931
Pseudo-LL	-32,913.21	-119,296	-17,795.52	-34,611.65	-74,001.93	-13,693.42	-39,837.76	-16,795.75
Observations	11,089	107,556	34,290	43,877	100,668	43,598	68,522	48,411

Notes: This table reports the localization and urbanization effects when the dependent variable is the number of new employments in each cell. Heteroscedasticity-robust standard errors are reported in parentheses. All columns include the district fixed effects. Significance level: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: Prepared by the author based on estimates.

Table B.14 Poisson regression without urbanization variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pharma. Products CNAE 212	Food Products CNAE 109	Fruit & vegetable Canning CNAE 103	Starch Products CNAE 106	Furniture CNAE 310	Prepress Services CNAE 182	Wood Products CNAE 162	Fin. of textile Articles CNAE 134
Panel A: The dependent variable is births of new establishments								
0 to 1 km	1.06e-03***	5.77e-04***	1.99e-03***	2.51e-03***	7.89e-04***	3.91e-03***	2.48e-03***	3.13e-03***
	(3.77e-04)	(4.62e-05)	(6.64e-04)	(3.96e-04)	(7.66e-05)	(4.32e-04)	(2.90e-04)	(5.37e-04)
1 to 5 km	1.24e-04	2.50e-04***	6.79e-04	2.58e-04	2.16e-04***	8.55e-04***	2.51e-04	6.48e-04***
	(2.90e-04)	(2.90e-05)	(5.05e-04)	(3.44e-04)	(4.80e-05)	(1.76e-04)	(2.08e-04)	(2.01e-04)
5 to 10 km	1.23e-04	4.85e-05*	2.73e-05	2.83e-05	5.87e-05	-2.40e-04*	1.69e-04	-4.26e-04***
	(1.80e-04)	(2.48e-05)	(4.74e-04)	(2.88e-04)	(4.55e-05)	(1.39e-04)	(1.72e-04)	(1.57e-04)
10 to 20 km	2.29e-05	3.27e-06	5.44e-05	4.77e-05	-3.61e-05	1.74e-06	1.88e-05	-9.83e-05
	(1.62e-04)	(2.08e-05)	(4.03e-04)	(2.40e-04)	(3.26e-05)	(1.10e-04)	(1.41e-04)	(1.08e-04)
20 to 40 km	1.26e-04	-4.87e-05***	3.54e-04	2.48e-04	1.09e-05	-3.21e-04***	-1.44e-04	-2.59e-04***
	(1.47e-04)	(1.76e-05)	(3.18e-04)	(2.01e-04)	(2.27e-05)	(1.13e-04)	(9.71e-05)	(9.10e-05)
# of district FE	81	1,874	381	663	1,579	395	957	426
Pseudo R^2	0.1367	0.0691	0.1156	0.1194	0.0961	0.1199	0.1104	0.1516
Pseudo-LL	-530.8668	-26,985.07	-2,701.629	-4,533.395	-24,718.3	-4,537.803	-9,075.686	-5,084.437
Observations	19,151	11,8376	47,495	61,350	111,668	56,812	82,976	58,389
Panel B: The dependent variable is new-establishment employment								
0 to 1 km	9.18e-04	1.59e-03***	4.67e-03***	4.97e-03***	8.72e-04***	3.67e-03**	5.81e-03***	3.33e-03***
	(8.45e-04)	(2.98e-04)	(1.30e-03)	(1.06e-03)	(1.93e-04)	(1.50e-03)	(8.39e-04)	(1.08e-03)
1 to 5 km	-5.81e-04	3.89e-04**	1.83e-03	-1.83e-04	-4.12e-05	9.21e-04*	1.48e-03**	4.91e-04
	(6.63e-04)	(1.76e-04)	(1.34e-03)	(9.53e-04)	(1.62e-04)	(5.42e-04)	(6.33e-04)	(3.45e-04)
5 to 10 km	2.54e-04	1.32e-04	1.60e-03	-8.16e-04	-2.27e-05	-8.75e-04*	5.25e-04	-4.53e-04
	(5.10e-04)	(1.41e-04)	(1.17e-03)	(9.34e-04)	(1.17e-04)	(4.72e-04)	(6.68e-04)	(4.18e-04)
10 to 20 km	1.18e-04	9.23e-05	7.86e-04	-1.14e-03*	-1.89e-05	-2.38e-04	-5.22e-04	-5.35e-04**
	(5.93e-04)	(1.44e-04)	(1.10e-03)	(6.31e-04)	(8.61e-05)	(3.48e-04)	(5.43e-04)	(2.53e-04)
20 to 40 km	1.68e-04	-2.28e-05	-5.44e-04	-1.03e-04	4.57e-05	-1.20e-03**	-3.61e-04	-1.36e-04
	(7.79e-04)	(5.96e-05)	(1.20e-03)	(5.78e-04)	(9.73e-05)	(4.93e-04)	(3.39e-04)	(1.76e-04)
# of district FE	41	1,374	214	385	1,175	245	682	308
Pseudo R^2	0.4885	0.2402	0.5506	0.3211	0.2469	0.2892	0.3965	0.2973
Pseudo-LL	-25,070.53	-115,061.8	-14,973.66	-33,003.63	-72,582.97	-12,548.98	-37,999.85	-16,421.54
Observations	11,089	105,114	33,052	42,262	98,583	42,243	66,745	46,925

Notes: This table reports the localization effects when we exclude agglomeration variables. All columns report the results of Poisson regressions where the dependent variable is the births of new establishments (Panel A) and the new-establishment employment (Panel B) and the variable of interest is the number of workers in the same industry in each concentric ring. All columns include the district fixed effects, diversification and competition variables, transport and geographic controls. Significance level: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: Prepared by the author based on estimates.

Table B.15 Endogeneity test for the localization variables

	(1) Pharma. Products CNAE 212	(2) Food Products CNAE 109	(3) Fruit & vegetable Canning CNAE 103	(4) Starch Products CNAE 106	(5) Furniture CNAE 310	(6) Prepress Services CNAE 182	(7) Wood Products CNAE 162	(8) Fin. of textile Articles CNAE 134
Panel A: The dependent variable is births of new establishments								
Residuals of equation 1 (0 to 1 km)	-7.93e-03**	-7.96e-04***	-5.48e-03**	-4.08e-03***	-3.43e-03***	-2.78e-03	-2.20e-03*	-1.11e-02**
	(3.52e-03)	(2.95e-04)	(2.68e-03)	(1.23e-03)	(5.02e-04)	(3.07e-03)	(1.22e-03)	(4.53e-03)
Residuals of equation 2 (1 to 5 km)	7.45e-04	-1.58e-04**	1.66e-04	-4.46e-04	7.90e-05	7.17e-04	-2.96e-04	-1.71e-03***
	(5.43e-04)	(6.55e-05)	(7.82e-04)	(4.00e-04)	(9.32e-05)	(4.68e-04)	(3.67e-04)	(6.51e-04)
Residuals of equation 3 (5 to 10 km)	-6.14e-05	9.00e-05**	2.02e-04	1.41e-05	-1.93e-04**	-3.05e-04	-2.76e-04	4.24e-04
	(2.55e-04)	(4.49e-05)	(6.97e-04)	(4.03e-04)	(8.62e-05)	(3.03e-04)	(3.43e-04)	(4.98e-04)
Residuals of equation 4 (10 to 20 km)	-6.98e-05	4.81e-05	-4.26e-04	-3.00e-04	-5.34e-05	-3.87e-04*	1.61e-04	-7.50e-04***
	(1.58e-04)	(3.37e-05)	(3.65e-04)	(2.49e-04)	(4.69e-05)	(1.98e-04)	(2.00e-04)	(1.84e-04)
Residuals of equation 5 (20 to 40 km)	1.92e-04	6.26e-05**	1.63e-04	1.17e-05	-4.57e-05	2.28e-04	-2.04e-04	-2.09e-04
	(1.57e-04)	(2.90e-05)	(3.26e-04)	(2.02e-04)	(2.79e-05)	(1.87e-04)	(1.40e-04)	(1.65e-04)
Panel B: The dependent variable is new-establishment employment								
Residuals of equation 1 (0 to 1 km)	-2.63e-02*	-1.83e-03**	-1.90e-02**	-3.13e-03	-5.50e-03***	-1.19e-03	-3.44e-03	-2.58e-02***
	(1.56e-02)	(9.33e-04)	(7.70e-03)	(2.83e-03)	(1.02e-03)	(1.40e-02)	(4.00e-03)	(7.87e-03)
Residuals of equation 2 (1 to 5 km)	1.88e-03	-1.36e-04	5.62e-03	-2.87e-03***	2.91e-04	-6.05e-04	2.44e-03*	-3.77e-03***
	(2.00e-03)	(2.60e-04)	(4.64e-03)	(9.45e-04)	(2.59e-04)	(1.93e-03)	(1.26e-03)	(1.28e-03)
Residuals of equation 3 (5 to 10 km)	-9.11e-04	4.56e-04	1.10e-03	-6.88e-04	-5.13e-04**	-2.70e-03	-2.15e-03	1.50e-03
	(7.37e-04)	(3.26e-04)	(2.54e-03)	(9.05e-04)	(2.24e-04)	(2.07e-03)	(1.76e-03)	(1.16e-03)
Residuals of equation 4 (10 to 20 km)	-4.06e-04	2.93e-05	-3.74e-04	-1.81e-03***	4.36e-05	-1.30e-03	-5.98e-04	-1.47e-03**
	(5.12e-04)	(1.52e-04)	(1.56e-03)	(6.06e-04)	(1.21e-04)	(8.44e-04)	(6.16e-04)	(6.11e-04)
Residuals of equation 5 (20 to 40 km)	-7.23e-05	2.08e-04	2.47e-04	-9.42e-04	2.12e-05	-5.75e-04	-8.45e-05	2.13e-04
	(4.93e-04)	(1.48e-04)	(1.23e-03)	(6.08e-04)	(1.17e-04)	(5.56e-04)	(6.41e-04)	(3.23e-04)

Notes: This table reports the estimated coefficients for the first-stage residuals. All columns report the results of Poisson regressions where the dependent variable is the births of new establishments (Panel A) and the new-establishment employment (Panel B). All columns include the urbanization, diversification and competition variables, transport and geographic controls as presented in the previous section. Standard errors based on 400 bootstrap replications are reported in parentheses. Significance level: *$p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: Prepared by the author based on estimates.

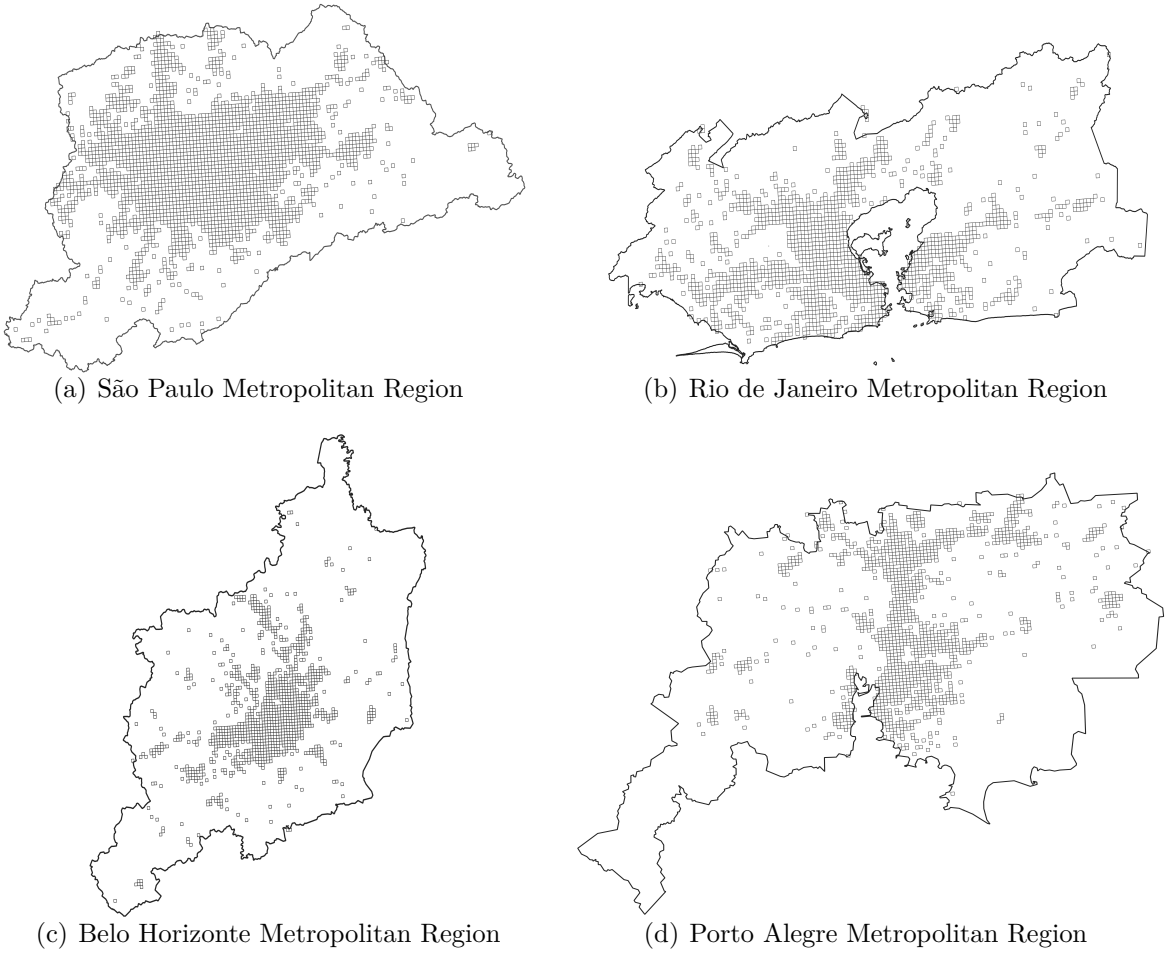
Appendix C

Appendix to Chapter 4

C.1 Empirical Strategy: additional details

In this Appendix we describe additional details of our empirical strategy. The next two subsections present additional figures mentioned in the text and a brief description of our strategy to control for worker-plant matching fixed effects, respectively.

C.1.1 Additional figures



(a) São Paulo Metropolitan Region

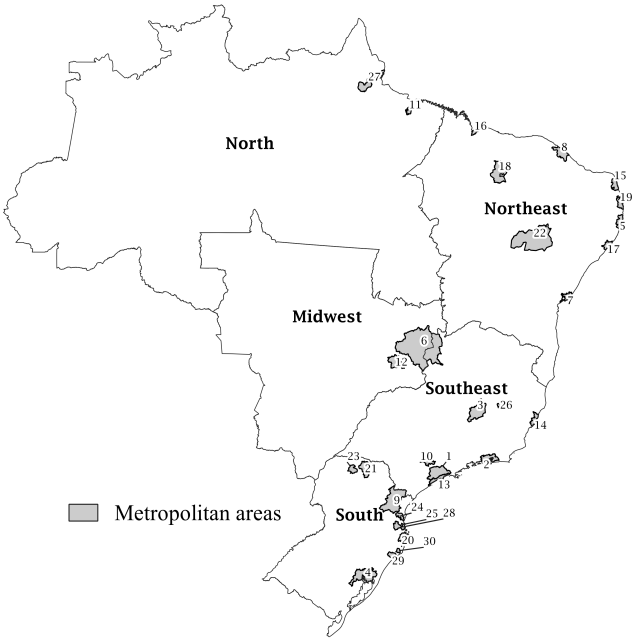
(b) Rio de Janeiro Metropolitan Region

(c) Belo Horizonte Metropolitan Region

(d) Porto Alegre Metropolitan Region

Figure C.1 Grid boundaries for selected metropolitan areas in 2014

Notes: The grids were exogenously defined based on the territorial limits of Brazil. We present only the cells which have at least one plant in their territory.



id	MR name
1	São Paulo
2	Rio de Janeiro
3	Belo Horizonte
4	Porto Alegre
5	Recife
6	RIDE Distrito Federal*
7	Salvador
8	Fortaleza
9	Curitiba
10	Campinas
11	Belém
12	Goiânia
13	Baixada Santista
14	Grande Vitória
15	Natal
16	Grande São Luís
17	Maceió
18	RIDE Teresina*
19	João Pessoa
20	Florianópolis
21	Londrina
22	RIDE Petrolina/Juazeiro*
23	Maringá
24	North Catarinense
25	Vale do Itajaí
26	Vale do Aço
27	Macapá
28	Foz do Rio Itajaí
29	Carbonífera
30	Tubarão

* RIDE: located in more than one federative unit.

Figure C.2 Brazilian Metropolitan Areas

Table C.1 Descriptive statistics by metropolitan region

id	MR name		Wage	Illiterate	Incomplete primary school	Incomplete high school	Incomplete college	College degree or more	Age	Tenure (months)
1	São Paulo	Obs.	904011	904011	904011	904011	904011	904011	904011	904011
		Mean	30.79	0.002	0.15	0.22	0.48	0.15	37.59	103.89
		Std. Dev.	32.08	0.047	0.36	0.41	0.50	0.35	8.18	82.01
		Min.	1.11	0	0	0	0	0	18	0
		Max.	739.91	1	1	1	1	1	56	511.5
2	Rio de Janeiro	Obs.	129537	129537	129537	129537	129537	129537	129537	129537
		Mean	24.57	0.002	0.18	0.29	0.43	0.10	39.29	108.29
		Std. Dev.	29.61	0.048	0.38	0.45	0.49	0.30	8.14	84.20
		Min.	1.22	0	0	0	0	0	18	0
		Max.	725.79	1	1	1	1	1	56	489.9
3	Belo Horizonte	Obs.	176687	176687	176687	176687	176687	176687	176687	176687
		Mean	25.82	0.002	0.13	0.28	0.44	0.14	37.08	98.50
		Std. Dev.	27.61	0.041	0.34	0.45	0.50	0.35	8.03	80.08
		Min.	1.11	0	0	0	0	0	18	0
		Max.	677.17	1	1	1	1	1	56	502.6
4	Porto Alegre	Obs.	179918	179918	179918	179918	179918	179918	179918	179918
		Mean	22.15	0.001	0.21	0.24	0.47	0.08	36.76	89.05
		Std. Dev.	22.21	0.033	0.41	0.43	0.50	0.27	8.46	79.31
		Min.	1.09	0	0	0	0	0	18	0
		Max.	900.29	1	1	1	1	1	56	505.9
5	Recife	Obs.	53245	53245	53245	53245	53245	53245	53245	53245
		Mean	17.40	0.04	0.18	0.16	0.55	0.07	37.97	101.63
		Std. Dev.	20.66	0.19	0.39	0.37	0.50	0.25	7.92	82.29
		Min.	1.11	0	0	0	0	0	18	0.1
		Max.	465.73	1	1	1	1	1	56	502
6	RIDE - Distrito Federal	Obs.	11045	11045	11045	11045	11045	11045	11045	11045
		Mean	16.67	0.01	0.27	0.26	0.40	0.07	35.48	75.95
		Std. Dev.	19.62	0.07	0.44	0.44	0.49	0.25	7.87	61.92
		Min.	1.60	0	0	0	0	0	18	0
		Max.	348.31	1	1	1	1	1	56	450.5
7	Salvador	Obs.	51776	51776	51776	51776	51776	51776	51776	51776
		Mean	33.80	0.002	0.08	0.11	0.68	0.13	36.97	93.20
		Std. Dev.	37.86	0.043	0.27	0.32	0.47	0.33	7.95	75.19
		Min.	1.17	0	0	0	0	0	18	0.2
		Max.	550.20	1	1	1	1	1	56	470.5
8	Fortaleza	Obs.	75279	75279	75279	75279	75279	75279	75279	75279
		Mean	11.31	0.01	0.20	0.28	0.48	0.04	36.40	90.34
		Std. Dev.	14.84	0.08	0.40	0.45	0.50	0.19	8.01	70.00

Continued on next page

Table C.1 – continued from previous page

id	MR name		Wage	Illiterate	Incomplete primary school	Incomplete high school	Incomplete college	College degree or more	Age	Tenure (months)
		Min.	1.26	0	0	0	0	0	18	0
		Max.	585.26	1	1	1	1	1	56	462.3
		Obs.	152916	152916	152916	152916	152916	152916	152916	152916
		Mean	27.866	0.001	0.112	0.199	0.537	0.150	36.793	93.879
		Std. Dev.	26.743	0.035	0.316	0.399	0.499	0.357	7.934	74.765
9	Curitiba	Min.	1.16	0	0	0	0	0	18	0
		Max.	727.54	1	1	1	1	1	56	486.9
		Obs.	206879	206879	206879	206879	206879	206879	206879	206879
		Mean	31.027	0.001	0.127	0.214	0.528	0.130	36.791	96.474
		Std. Dev.	30.586	0.035	0.333	0.410	0.499	0.336	8.320	79.241
10	Campinas	Min.	1.18	0	0	0	0	0	18	0
		Max.	711.11	1	1	1	1	1	56	515.3
		Obs.	10203	10203	10203	10203	10203	10203	10203	10203
		Mean	13.62	0.01	0.23	0.34	0.38	0.04	37.60	94.07
		Std. Dev.	15.96	0.10	0.42	0.47	0.49	0.19	7.99	71.74
11	Belém	Min.	1.25	0	0	0	0	0	18	0.2
		Max.	379.97	1	1	1	1	1	56	429.9
		Obs.	29578	29578	29578	29578	29578	29578	29578	29578
		Mean	13.813	0.004	0.236	0.293	0.415	0.053	35.919	79.248
		Std. Dev.	14.662	0.065	0.424	0.455	0.493	0.224	8.116	64.279
12	Goiânia	Min.	1.14	0	0	0	0	0	18	0.2
		Max.	334.72	1	1	1	1	1	56	454
		Obs.	21695	21695	21695	21695	21695	21695	21695	21695
		Mean	38.079	0.001	0.057	0.088	0.677	0.177	39.165	102.834
		Std. Dev.	32.405	0.023	0.232	0.283	0.468	0.382	7.839	89.904
13	Baixada Santista	Min.	1.29	0	0	0	0	0	18	0.3
		Max.	690.62	1	1	1	1	1	56	452.9
		Obs.	30702	30702	30702	30702	30702	30702	30702	30702
		Mean	29.617	0.001	0.088	0.161	0.575	0.174	37.199	73.469
		Std. Dev.	27.450	0.032	0.284	0.368	0.494	0.379	8.350	69.227
14	Grande Vitória	Min.	1.59	0	0	0	0	0	18	0.1
		Max.	507.99	1	1	1	1	1	56	454.9
		Obs.	18999	18999	18999	18999	18999	18999	18999	18999
		Mean	10.008	0.006	0.238	0.308	0.423	0.025	36.453	80.288
		Std. Dev.	10.887	0.076	0.426	0.462	0.494	0.155	7.895	59.556
15	Natal	Min.	1.09	0	0	0	0	0	18	0.1
		Max.	314.84	1	1	1	1	1	56	411.2
		Obs.	6306	6306	6306	6306	6306	6306	6306	6306
		Mean	25.904	0.004	0.066	0.110	0.734	0.087	38.992	128.971
		Std. Dev.	20.846	0.062	0.248	0.313	0.442	0.281	7.487	92.753
16	Grande São Luís	Min.	1.28	0	0	0	0	0	18	0.5

Continued on next page

Table C.1 – continued from previous page

id	MR name		Wage	Illiterate	Incomplete primary school	Incomplete high school	Incomplete college	College degree or more	Age	Tenure (months)
		Max.	350.34	1	1	1	1	1	56	413.9
		Obs.	19229	19229	19229	19229	19229	19229	19229	19229
		Mean	13.64	0.12	0.49	0.17	0.18	0.04	37.43	104.92
		Std. Dev.	20.91	0.33	0.50	0.37	0.39	0.19	8.25	93.73
17	Maceió	Min.	1.18	0	0	0	0	0	18	0.3
		Max.	440.87	1	1	1	1	1	56	469.1
		Obs.	11535	11535	11535	11535	11535	11535	11535	11535
		Mean	8.65	0.01	0.35	0.31	0.30	0.03	37.32	93.15
		Std. Dev.	6.63	0.12	0.48	0.46	0.46	0.16	8.03	69.52
18	RIDE - Teresina	Min.	1.41	0	0	0	0	0	18	0.4
		Max.	185.14	1	1	1	1	1	56	367.9
		Obs.	19933	19933	19933	19933	19933	19933	19933	19933
		Mean	10.18	0.04	0.37	0.28	0.29	0.02	37.11	93.32
		Std. Dev.	9.38	0.19	0.48	0.45	0.46	0.15	8.07	80.41
19	João Pessoa	Min.	1.28	0	0	0	0	0	18	0.1
		Max.	206.48	1	1	1	1	1	56	531.5
		Obs.	9458	9458	9458	9458	9458	9458	9458	9458
		Mean	16.085	0.004	0.166	0.275	0.460	0.095	35.902	91.929
		Std. Dev.	14.349	0.062	0.372	0.446	0.498	0.294	8.261	71.917
20	Florianópolis	Min.	1.88	0	0	0	0	0	18	0.1
		Max.	232.89	1	1	1	1	1	56	416.2
		Obs.	31835	31835	31835	31835	31835	31835	31835	31835
		Mean	15.416	0.002	0.210	0.279	0.444	0.065	36.532	80.912
		Std. Dev.	13.845	0.046	0.407	0.448	0.497	0.246	8.313	66.974
21	Londrina	Min.	1.37	0	0	0	0	0	18	0.1
		Max.	522.21	1	1	1	1	1	56	488.9
		Obs.	2686	2686	2686	2686	2686	2686	2686	2686
		Mean	10.87	0.01	0.345	0.257	0.356	0.032	37.348	102.033
	RIDE	Std. Dev.	14.58	0.10	0.476	0.437	0.479	0.175	8.023	75.813
22	Petrolina/PE	Min.	1.97	0	0	0	0	0	18	0.1
	Juazeiro/BA	Max.	593.98	1	1	1	1	1	56	372.7
		Obs.	20098	20098	20098	20098	20098	20098	20098	20098
		Mean	13.271	0.003	0.225	0.299	0.427	0.046	36.891	79.703
		Std. Dev.	11.774	0.056	0.418	0.458	0.495	0.209	8.563	66.460
23	Maringá	Min.	1.20	0	0	0	0	0	18	0
		Max.	377.78	1	1	1	1	1	56	419.9
		Obs.	76024	76024	76024	76024	76024	76024	76024	76024
		Mean	22.753	0.002	0.064	0.244	0.570	0.121	36.286	108.530
	North/Northeast	Std. Dev.	19.351	0.045	0.244	0.429	0.495	0.326	8.254	88.261
24	Catarinense	Min.	1.32	0	0	0	0	0	18	0.1
		Max.	582.69	1	1	1	1	1	56	496

Continued on next page

Table C.1 – continued from previous page

id	MR name		Wage	Illiterate	Incomplete primary school	Incomplete high school	Incomplete college	College degree or more	Age	Tenure (months)
		Obs.	63972	63972	63972	63972	63972	63972	63972	63972
		Mean	17.935	0.002	0.186	0.324	0.417	0.072	36.109	90.939
		Std. Dev.	14.476	0.039	0.389	0.468	0.493	0.259	8.595	80.042
25	Vale do Itajaí	Min.	1.20	0	0	0	0	0	18	0.1
		Max.	310.05	1	1	1	1	1	56	486.6
		Obs.	33628	33628	33628	33628	33628	33628	33628	33628
		Mean	33.919	0.001	0.049	0.191	0.612	0.146	36.142	146.066
		Std. Dev.	24.849	0.031	0.217	0.393	0.487	0.353	8.101	103.407
26	Vale do Aço	Min.	1.21	0	0	0	0	0	18	0
		Max.	612.09	1	1	1	1	1	56	476.9
		Obs.	338	338	338	338	338	338	338	338
		Mean	8.34	0.02	0.19	0.29	0.49	0.00	37.52	74.36
		Std. Dev.	4.94	0.14	0.39	0.46	0.50	0.05	7.80	65.09
27	Macapá	Min.	2.79	0	0	0	0	0	18	0.7
		Max.	39.29	1	1	1	1	1	56	353.9
		Obs.	8571	8571	8571	8571	8571	8571	8571	8571
		Mean	21.078	0.003	0.229	0.307	0.387	0.074	36.511	76.640
		Std. Dev.	22.700	0.051	0.420	0.461	0.487	0.262	8.730	62.002
28	Foz do Rio Itajaí	Min.	1.37	0	0	0	0	0	18	0.2
		Max.	672.06	1	1	1	1	1	56	455.5
		Obs.	25649	25649	25649	25649	25649	25649	25649	25649
		Mean	16.799	0.002	0.204	0.261	0.464	0.069	35.358	85.203
		Std. Dev.	15.223	0.043	0.403	0.439	0.499	0.254	8.512	67.120
29	Carbonífera	Min.	1.09	0	0	0	0	0	18	0
		Max.	311.59	1	1	1	1	1	56	471.9
		Obs.	5843	5843	5843	5843	5843	5843	5843	5843
		Mean	14.317	0.001	0.159	0.240	0.551	0.049	34.910	94.964
		Std. Dev.	10.939	0.037	0.366	0.427	0.497	0.215	8.110	70.727
30	Tubarão	Min.	1.74	0	0	0	0	0	18	0
		Max.	368.25	1	1	1	1	1	56	392.9
		Obs.	2387575	2387575	2387575	2387575	2387575	2387575	2387575	2387575
		Mean	26.307	0.004	0.157	0.233	0.486	0.120	37.222	98.998
		Std. Dev.	28.518	0.065	0.363	0.423	0.500	0.325	8.221	80.623
	Total	Min.	1.09	0	0	0	0	0	18	0
		Max.	900.29	1	1	1	1	1	56	531.5

Notes: Illiterate, incomplete primary school, complete primary school to incomplete high school, complete high school to incomplete college and college degree or more are dummies variables. Source: Author' computations using information from RAIS.

C.1.2 Worker-plant matching fixed effects

Consider a more general version of linear model in equation 4.9:

$$\mathsf{w}_{it} = \mathbf{X}_{it}\lambda + \mathbf{H}_{j(i,t)t}\gamma + \sum_r \beta_r \mathsf{S}_{zrt} + \alpha_i + \phi_{j(i,t)} + \mu_c + \psi_{pt} + \varepsilon_{izt} \tag{C.1}$$

where $j(i,t)$ is function that maps worker i to plant j at year t; $\phi_{j(i,t)}$ are observed and unobserved heterogeneities fixed over time at plant level; and ε_{izt} is the error component. All other variables are defined above.

Estimation of equation C.1 is computationally challenging when working with a large database (around 90,000 plants per year). To deal with this problem, we utilized the "spell fixed effects" method proposed by Andrews *et al.* (2006) and combined the worker and plant fixed effects into a single effect: $\eta_s = \alpha_i + \phi_{j(i,t)}$. This combined effect represents each unique worker-plant match (spell-level heterogeneity). If the match-specific effect is correlated with the number of college-educated workers in each ring, $\text{Cov}(\eta_s, \mathsf{S}_{zrt}) \neq 0$, our estimates may be capturing this effect. This effect is removed by subtracting averages at the match level, so that both α_i and $\phi_{j(i,t)}$ have disappear:

$$\begin{aligned}\mathsf{w}_{it} - \bar{\mathsf{w}}_s = (\mathbf{X}_{it} - \bar{\mathbf{X}}_s)\lambda + (\mathbf{H}_{j(i,t)t} - \bar{\mathbf{H}}_s)\gamma + \sum_r \beta_r(\mathsf{S}_{zrt} - \bar{\mathsf{S}}_s) + (\mu_c - \bar{\mu}_s) + \\ (\psi_{pt} - \bar{\psi}_s) + (\varepsilon_{izt} - \bar{\varepsilon}_s)\end{aligned} \tag{C.2}$$

C.2 Additional Results

This Appendix present all the additional results mentioned in the text.

Table C.2 Sample percentiles for concentric ring employment variables

	Sample percentile				
# of workers	10th	25th	50th	75th	90th
Within 0 to 1 km	288	686	1,610	3,627	6,668
Within 1 to 5 km	1,384	3,929	10,046	26,304	51,735
Within 5 to 10 km	2,670	7,444	20,564	51,266	104,457
Within 10 to 20 km	5,994	19,379	45,115	196,072	267,098
Within 20 to 40 km	12,918	27,319	87,958	262,766	373,024
College-or-more, 0 to 1 km	8	30	115	386	980
College-or-more, 1 to 5 km	51	195	778	2,507	6,808
College-or-more, 5 to 10 km	101	375	1,741	6,140	14,556
College-or-more, 10 to 20 km	324	1,168	4,194	23,960	36,210
College-or-more, 20 to 40 km	403	1,640	7,445	25,837	46,274
# of plants					
Within 0 to 1 km	5	14	34	71	131
Within 1 to 5 km	54	182	473	1,078	2,122
Within 5 to 10 km	108	368	987	2,372	5,666
Within 10 to 20 km	297	945	2,483	11,239	14,030
Within 20 to 40 km	459	1,274	4,053	10,539	17,301

Source: Author' computations using informations from RAIS.

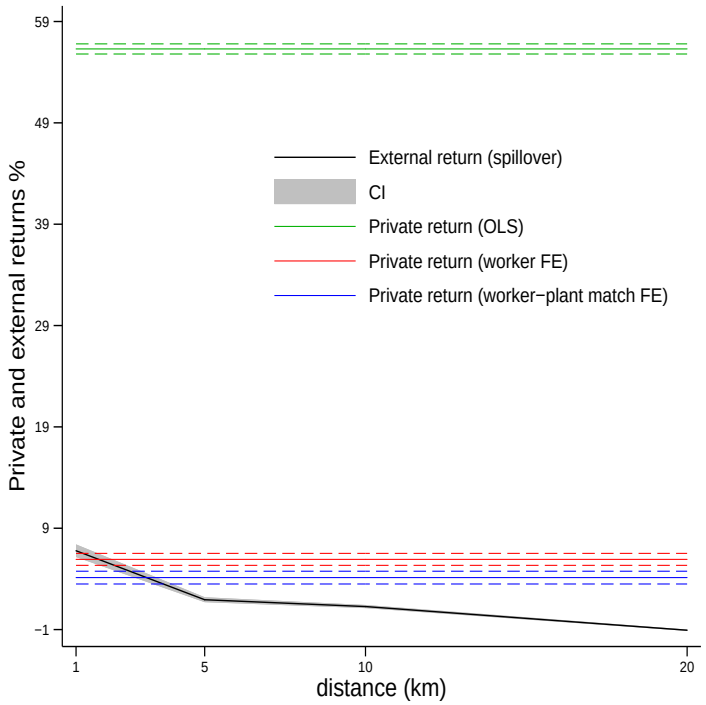


Figure C.3 External versus private returns with OLS results

Table C.3 Spatial scope of human capital externalities using unrestricted sample

Dependent variable: individual hourly wage (in log)						
# of workers with college-or-more	OLS (1)	OLS (2)	OLS (3)	OLS (4)	FE (5)	FE + IV (6)
0 to 1 km	4.90e-05***	2.86e-05***	1.55e-05***	1.95e-05***	4.62e-06***	5.56e-05***
	(3.81e-07)	(3.44e-07)	(3.24e-07)	(3.17e-07)	(2.52e-07)	(4.83e-06)
1 to 5 km	6.14e-06***	4.96e-06***	4.67e-06***	5.96e-06***	-3.84e-07**	2.38e-05***
	(1.18e-07)	(1.13e-07)	(1.10e-07)	(1.10e-07)	(1.60e-07)	(1.86e-06)
5 to 10 km	-4.92e-07***	3.65e-07***	3.24e-07***	5.71e-07***	-4.51e-08	9.05e-08
	(6.43e-08)	(6.08e-08)	(5.91e-08)	(5.85e-08)	(9.41e-08)	(8.03e-07)
10 to 20 km	2.21e-06***	2.89e-06***	2.54e-06***	-1.17e-07***	-8.07e-07***	-5.11e-06***
	(2.64e-08)	(2.51e-08)	(2.45e-08)	(3.14e-08)	(5.32e-08)	(2.87e-07)
Worker-level controls	Yes	Yes	Yes	Yes	Yes	Yes
Plant-level controls	No	Yes	Yes	Yes	Yes	Yes
Industry × year effect	No	No	Yes	Yes	Yes	Yes
Metropolitan region FE	No	No	No	Yes	Yes	Yes
Kleibergen-Paap rk F[a]						1,421.11
Kleibergen-Paap rk LM[b]						5,048.33
1 st stage F-stat. 0 to 1 km						2,805.74
1 st stage F-stat. 1 to 5 km						10,179.67
1 st stage F-stat. 5 to 10 km						30,048.87
1 st stage F-stat. 10 to 20 km						77,366.28
F-stat.	48,167.72	50,274.55	21,272.01	21,643.06	–	–
R squared	0.5265	0.5651	0.5985	0.6347	0.3828	–

Notes: This table presents the estimates obtained from equation 4.9 using the unrestricted sample to calculate the agglomeration variables, i.e., all workers in the manufacturing industry are considered in each ring regardless of gender, age and type of contract. All models are estimated with 2,387,434 observations. Worker-level controls include all the individual characteristics detailed above. Plant-level controls are dummies for plant size. Industry × year effects are dummies for each 2-digit × year combination. Metropolitan region FE are metropolitan region fixed effects. [a]: H_0 - weakly identified model. [b]: H_0 - under-identified model. Standard errors adjusted for clustering are in parentheses. Significance levels: *** $p < 0.01$. Source: Prepared by the author based on estimates.

Table C.4 Basic models: spatial scope of human capital spillovers

Dependent variable: individual hourly wage (in log)			
# of workers with college-or-more	OLS		
	(1)	(2)	(3)
0 to 1 km	6.39e-05***	3.65e-05***	1.90e-05***
	(5.38e-07)	(4.81e-07)	(4.42e-07)
1 to 5 km	1.10e-05***	8.24e-06***	6.76e-06***
	(1.74e-07)	(1.66e-07)	(1.61e-07)
5 to 10 km	1.25e-06***	1.86e-06***	1.50e-06***
	(1.03e-07)	(9.74e-08)	(9.48e-08)
10 to 20 km	3.20e-06***	4.75e-06***	4.32e-06***
	(4.33e-08)	(4.13e-08)	(4.02e-08)
Worker-level controls	Yes	Yes	Yes
Plant-level controls	No	Yes	Yes
Industry × year effect	No	No	Yes
F-stat.	48,138.22	50,366.39	21,295.98
R squared	0.5267	0.5655	0.5988

Notes: This table presents the estimates obtained from equation 4.9 when we consider the number of college educated worker in each ring. All models are estimated with 2,387,434 observations. Worker-level controls include all the individual characteristics detailed above. Plant-level controls are dummies for plant size. Industry × year effects are dummies for each 2-digit × year combination. Standard errors adjusted for clustering are in parentheses. Significance levels: *** $p < 0.01$. Source: Prepared by the author based on estimates.

Table C.5 Basic models: spatial scope of heterogeneity of human capital externalities by education groups

Dependent variable: individual hourly wage (in log)			
# of workers with college-or-more	OLS (1)	OLS (2)	OLS (3)
Panel A: Less than college degree			
0 to 1 km	1.12e-04***	7.37e-05***	5.04e-05***
	(8.19e-07)	(6.84e-07)	(6.18e-07)
1 to 5 km	7.19e-06***	4.07e-06***	2.25e-06***
	(1.89e-07)	(1.79e-07)	(1.75e-07)
5 to 10 km	1.72e-06***	2.17e-06***	1.39e-06***
	(1.10e-07)	(1.03e-07)	(1.01e-07)
10 to 20 km	3.25e-06***	5.07e-06***	4.81e-06***
	(4.59e-08)	(4.39e-08)	(4.30e-08)
R squared	0.3738	0.4247	0.4665
Panel B: College degree or more			
0 to 1 km	6.85e-05***	4.41e-05***	2.62e-05***
	(1.21e-06)	(1.13e-06)	(1.10e-06)
1 to 5 km	1.44e-05***	1.60e-05***	1.43e-05***
	(4.98e-07)	(4.82e-07)	(4.72e-07)
5 to 10 km	-2.19e-08	1.03e-06***	2.59e-06***
	(3.22e-07)	(3.12e-07)	(3.01e-07)
10 to 20 km	8.23e-07***	1.03e-06***	2.44e-07*
	(1.45e-07)	(1.39e-07)	(1.35e-07)
R squared	0.3246	0.3735	0.4248
Controls to Panel A and B			
Worker-level controls	Yes	Yes	Yes
Plant-level controls	No	Yes	Yes
Industry × year effect	No	No	Yes

Notes: All models in Panel A are estimated with 2,003,730 observations and in Panel B with 187,511 observations. The industry × year effect is computed at the 2-digit level. All the controls shown at the bottom of this table are included in both panel A and panel B. Robust standard errors in parentheses. Significance level: ** $p < 0.05$, *** $p < 0.01$. Source: Prepared by the author based on estimates.

Table C.6 First-stage results: spatial scope of human capital spillovers

Dependent variable:	# of workers with college-or-more			
	0 to 1 km	1 to 5 km	5 to 10 km	10 to 20 km
Shift-Share IV				
0 to 5 km	7.87e-02***	9.38e-01***	3.85e+00***	7.58e+00***
	(2.98e-03)	(9.13e-03)	(1.55e-02)	(2.49e-02)
20 to 40 km	-1.25e-02***	7.76e-03***	-2.94e-02***	1.07e-01***
	(2.29e-04)	(5.47e-04)	(8.95e-04)	(1.36e-03)
40 to 80 km	3.43e-03***	-5.07e-02***	-6.69e-02***	-1.50e-01***
	(2.81e-04)	(5.90e-04)	(8.90e-04)	(1.71e-03)
80 to 120 km	-1.79e-02***	8.49e-03***	4.46e-02***	1.30e-01***
	(2.41e-04)	(3.64e-04)	(4.95e-04)	(8.27e-04)
Worker-level controls	Yes	Yes	Yes	Yes
Plant-level controls	Yes	Yes	Yes	Yes
Industry × year FE	Yes	Yes	Yes	Yes
Metropolitan region FE	Yes	Yes	Yes	Yes
F-stat.	285.03	831.60	1552.72	4253.71
R squared	0.1904	0.2264	0.4181	0.6461

Notes: This table presents the first stage estimates of the model in column 3 of Table 4.3. All models are estimated with 2,387,434 observations. Worker-level controls include all the individual characteristics detailed above. Plant-level controls are dummies for plant size. Industry × year effect are dummies for each 2-digit × year combination. Metropolitan region FE are metropolitan region fixed effects. Standard errors adjusted for clustering are in parentheses. Significance levels: *** $p < 0.01$. Source: Prepared by the author based on estimates.

Table C.7 Leave-one-out by metropolitan region

Dependent variable: individual hourly wage (in log)										
# of workers with college-or-more	FE + IV									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
0 to 1 km	6.95e-05***	6.35e-05***	3.46e-04***	6.09e-05***	6.31e-05***	6.71e-05***	5.70e-05***	6.18e-05***	4.60e-05***	6.80e-04***
	(2.17e-05)	(6.90e-06)	(2.83e-05)	(5.96e-06)	(6.17e-06)	(6.25e-06)	(6.04e-06)	(6.12e-06)	(5.46e-06)	(3.39e-05)
1 to 5 km	6.13e-05**	1.10e-06	1.67e-04***	1.63e-05***	1.52e-05***	1.93e-05***	1.52e-05***	1.75e-05***	9.94e-06***	6.37e-05***
	(2.92e-05)	(1.93e-06)	(1.21e-05)	(2.28e-06)	(2.39e-06)	(2.44e-06)	(2.39e-06)	(2.41e-06)	(2.31e-06)	(7.93e-06)
5 to 10 km	-3.61e-05***	3.17e-05***	-2.73e-05***	1.43e-05***	1.27e-05***	1.28e-05***	1.39e-05***	1.31e-05***	1.02e-05***	-1.03e-04***
	(7.35e-06)	(1.55e-06)	(4.73e-06)	(1.52e-06)	(1.55e-06)	(1.55e-06)	(1.53e-06)	(1.55e-06)	(1.64e-06)	(7.33e-06)
10 to 20 km	1.83e-05***	-1.73e-05***	-2.26e-05***	-1.07e-05***	-9.97e-06***	-1.06e-05***	-1.04e-05***	-1.04e-05***	-8.04e-06***	3.76e-05***
	(2.02e-06)	(7.12e-07)	(9.03e-07)	(6.68e-07)	(6.64e-07)	(6.62e-07)	(6.65e-07)	(6.61e-07)	(7.06e-07)	(2.66e-06)
Kleibergen-Paap rk F[a]	47.551	1,448.921	245.5571	1,455.679	1,467.317	1,465.781	1,483.188	1,478.196	1,558.431	119.064
Kleibergen-Paap rk LM[b]	34.465	5,326.061	942.3623	5,370.696	5,356.963	5,345.846	5,441.904	5,404.029	5,703.539	447.5052
Observations	1,460,531	2,257,504	2,210,465	2,207,179	2,334,010	2,376,349	2,335,500	2,312,067	2,233,967	2,178,681
# of workers with college-or-more	FE + IV									
	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
0 to 1 km	6.81e-05***	6.41e-05***	6.46e-05***	6.61e-05***	6.96e-05***	6.78e-05***	6.82e-05***	6.81e-05***	6.53e-05***	6.73e-05***
	(6.24e-06)	(6.22e-06)	(6.24e-06)	(6.21e-06)	(6.23e-06)	(6.24e-06)	(6.24e-06)	(6.25e-06)	(6.16e-06)	(6.24e-06)
1 to 5 km	1.95e-05***	1.65e-05***	2.13e-05***	1.80e-05***	2.01e-05***	1.94e-05***	2.00e-05***	1.96e-05***	1.98e-05***	1.88e-05***
	(2.44e-06)	(2.41e-06)	(2.47e-06)	(2.41e-06)	(2.44e-06)	(2.44e-06)	(2.44e-06)	(2.43e-06)	(2.44e-06)	(2.43e-06)
5 to 10 km	1.27e-05***	1.28e-05***	1.38e-05***	1.23e-05***	1.28e-05***	1.27e-05***	1.27e-05***	1.27e-05***	1.33e-05***	1.26e-05***
	(1.55e-06)	(1.54e-06)	(1.54e-06)	(1.55e-06)	(1.55e-06)	(1.55e-06)	(1.55e-06)	(1.55e-06)	(1.55e-06)	(1.55e-06)
10 to 20 km	-1.06e-05***	-1.01e-05***	-1.13e-05***	-1.02e-05***	-1.07e-05***	-1.06e-05***	-1.07e-05***	-1.06e-05***	-1.09e-05***	-1.05e-05***
	(6.62e-07)	(6.62e-07)	(6.58e-07)	(6.65e-07)	(6.63e-07)	(6.62e-07)	(6.64e-07)	(6.61e-07)	(6.62e-07)	(6.62e-07)
Kleibergen-Paap rk F	1,466.988	1,467.078	1,460.665	1,468.929	1,469.203	1,466.576	1,467.98	1,465.098	1,470.813	1,467.47
Kleibergen-Paap rk LM	5,352.056	5,352.578	5,331.501	5,363.413	5,363.563	5,350.059	5,354.204	5,342.915	5,373.265	5,351.807
Observations	2,377,214	2,357,819	2,365,598	2,356,643	2,368,403	2,381,113	2,368,169	2,375,889	2,367,455	2,377,951
# of workers with college-or-more	FE + IV									
	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
0 to 1 km	6.51e-05***	6.77e-05***	6.69e-05***	4.84e-05***	7.07e-05***	6.95e-05***	6.78e-05***	6.70e-05***	6.80e-05***	6.79e-05***
	(6.18e-06)	(6.25e-06)	(6.23e-06)	(6.39e-06)	(6.24e-06)	(6.28e-06)	(6.25e-06)	(6.24e-06)	(6.24e-06)	(6.25e-06)
1 to 5 km	1.75e-05***	1.96e-05***	1.75e-05***	2.92e-05***	1.91e-05***	2.20e-05***	1.95e-05***	1.88e-05***	1.92e-05***	1.90e-05***
	(2.41e-06)	(2.44e-06)	(2.42e-06)	(2.46e-06)	(2.42e-06)	(2.41e-06)	(2.44e-06)	(2.44e-06)	(2.42e-06)	(2.44e-06)
5 to 10 km	1.27e-05***	1.27e-05***	1.26e-05***	1.81e-05***	1.23e-05***	1.30e-05***	1.27e-05***	1.28e-05***	1.28e-05***	1.27e-05***
	(1.55e-06)	(1.55e-06)	(1.55e-06)	(1.61e-06)	(1.55e-06)	(1.55e-06)	(1.55e-06)	(1.55e-06)	(1.55e-06)	(1.55e-06)
10 to 20 km	-1.02e-05***	-1.06e-05***	-1.02e-05***	-1.49e-05***	-1.03e-05***	-1.12e-05***	-1.06e-05***	-1.05e-05***	-1.06e-05***	-1.05e-05***
	(6.63e-07)	(6.62e-07)	(6.63e-07)	(7.00e-07)	(6.65e-07)	(6.55e-07)	(6.62e-07)	(6.61e-07)	(6.63e-07)	(6.62e-07)
Kleibergen-Paap rk F	1,468.375	1,466.532	1,465.868	1,413.149	1,466.74	1,459.126	1,466.286	1,468.888	1,466.255	1,467.021
Kleibergen-Paap rk LM	5,362.496	5,348.973	5,345.429	5,135.42	5,350.901	5,318.742	5,346.958	5,355.808	5,350.127	5,349.067
Observations	2,355,534	2,384,741	2,367,305	2,311,224	2,323,345	2,353,721	2,387,096	2,378,811	2,361,722	2,381,569

Notes: [a]: H_0 - weakly identified model. [b]: H_0 - under-identified model. Standard errors adjusted for clustering are in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$. Source: Prepared by the author based on estimates.

Table C.8 Leave-one-out by 2-digit CNAE code

Dependent variable: individual hourly wage (in log)								
# of workers with college-or-more	FE + IV							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0 to 1 km	4.99e-05***	6.95e-05***	6.73e-05***	7.83e-05***	6.96e-05***	6.70e-05***	6.88e-05***	6.41e-05***
	(5.93e-06)	(6.34e-06)	(6.25e-06)	(6.12e-06)	(6.30e-06)	(6.23e-06)	(6.24e-06)	(6.02e-06)
1 to 5 km	1.27e-05***	1.95e-05***	1.94e-05***	1.69e-05***	1.69e-05***	2.07e-05***	1.89e-05***	1.49e-05***
	(2.36e-06)	(2.43e-06)	(2.44e-06)	(2.47e-06)	(2.43e-06)	(2.44e-06)	(2.44e-06)	(2.47e-06)
5 to 10 km	2.22e-05***	1.29e-05***	1.26e-05***	1.18e-05***	1.28e-05***	1.26e-05***	1.27e-05***	1.34e-05***
	(1.60e-06)	(1.57e-06)	(1.55e-06)	(1.56e-06)	(1.55e-06)	(1.55e-06)	(1.55e-06)	(1.53e-06)
10 to 20 km	-1.43e-05***	-1.08e-05***	-1.06e-05***	-9.78e-06***	-1.01e-05***	-1.07e-05***	-1.06e-05***	-1.05e-05***
	(7.02e-07)	(6.71e-07)	(6.63e-07)	(6.76e-07)	(6.77e-07)	(6.61e-07)	(6.68e-07)	(6.49e-07)
Kleibergen-Paap rk F[a]	1,478.47	1,405.731	1,465.474	1,447.726	1,425.33	1473.488	1,462.115	1,536.334
Kleibergen-Paap rk LM[b]	5,376.086	5,139.266	5,343.461	5,296.554	5,222.444	5,369.944	5,332.828	5,549.549
Observations	2,188,503	2,341,365	2,381,599	2,259,154	2,333,385	2,338,107	2,365,773	2,317,499
# of workers with college-or-more	FE + IV							
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
0 to 1 km	6.69e-05***	7.57e-05***	5.27e-05***	6.18e-05***	4.69e-05***	6.94e-05***	6.07e-05***	5.91e-05***
	(6.21e-06)	(7.04e-06)	(6.26e-06)	(6.27e-06)	(6.29e-06)	(5.95e-06)	(6.02e-06)	(6.09e-06)
1 to 5 km	1.83e-05***	2.25e-05***	1.55e-05***	1.55e-05***	1.76e-05***	1.73e-05***	2.52e-05***	1.68e-05***
	(2.51e-06)	(2.39e-06)	(2.43e-06)	(2.44e-06)	(2.62e-06)	(2.50e-06)	(2.76e-06)	(2.68e-06)
5 to 10 km	1.36e-05***	1.05e-05***	1.78e-05***	1.72e-05***	1.37e-05***	1.35e-05***	1.27e-05***	1.53e-05***
	(1.57e-06)	(1.68e-06)	(1.47e-06)	(1.60e-06)	(1.49e-06)	(1.56e-06)	(1.75e-06)	(1.71e-06)
10 to 20 km	-1.08e-05***	-1.02e-05***	-1.23e-05***	-1.19e-05***	-1.09e-05***	-1.07e-05***	-1.18e-05***	-1.16e-05***
	(6.65e-07)	(7.01e-07)	(6.76e-07)	(6.89e-07)	(6.84e-07)	(6.52e-07)	(7.20e-07)	(7.60e-07)
Kleibergen-Paap rk F	1,464.801	1,369.016	1,345.695	1,382.187	1,276.398	1,572.162	1,467.074	1,355.237
Kleibergen-Paap rk LM	5,348.778	5,022.005	5,027.295	5,087.883	4,715.451	5,653.652	5,354.587	4,961.148
Observations	2,348,869	2,372,417	2,225,615	2,342,922	2,164,619	2,271,078	2,202,423	2,161,786
# of workers with college-or-more	FE + IV							
	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
0 to 1 km	6.19e-05***	5.81e-05***	5.99e-05***	6.56e-04***	7.52e-05***	6.88e-05***	6.73e-05***	7.22e-05***
	(6.31e-06)	(6.05e-06)	(6.17e-06)	(4.96e-05)	(6.02e-06)	(6.20e-06)	(6.25e-06)	(6.41e-06)
1 to 5 km	1.61e-05***	2.02e-05***	1.87e-05***	2.38e-04***	1.72e-05***	1.81e-05***	1.89e-05***	1.58e-05***
	(2.45e-06)	(2.54e-06)	(2.66e-06)	(1.71e-05)	(2.43e-06)	(2.45e-06)	(2.43e-06)	(2.43e-06)
5 to 10 km	1.50e-05***	1.51e-05***	1.51e-05***	-8.40e-05***	1.23e-05***	1.13e-05***	1.20e-05***	1.35e-05***
	(1.60e-06)	(1.66e-06)	(1.73e-06)	(7.64e-06)	(1.57e-06)	(1.56e-06)	(1.55e-06)	(1.45e-06)
10 to 20 km	-1.11e-05***	-1.20e-05***	-1.20e-05***	-4.74e-07	-1.01e-05***	-9.63e-06***	-1.02e-05***	-1.05e-05***
	(6.84e-07)	(6.96e-07)	(7.50e-07)	(1.11e-06)	(6.43e-07)	(6.65e-07)	(6.70e-07)	(6.39e-07)
Kleibergen-Paap rk F	1,385.285	1,471.315	1,373.077	80.201	1,595.829	1,457.471	1,453.045	1,399.921
Kleibergen-Paap rk LM	5,073.521	5,347.26	5,000.703	312.784	5,721.016	5,312.199	5,296.876	5,179.779
Observations	2,345,822	2,294,024	2,184,389	2,011,557	2,367,543	2,336,968	2,352,190	2,354,927

Notes: [a]: H_0 - weakly identified model. [b]: H_0 - under-identified model. Standard errors adjusted for clustering are in parentheses. Significance levels: *** $p < 0.01$. Source: Prepared by the author based on estimates.

Table C.9 Private returns to education - including agglomeration variables

Dependent variable: individual hourly wage (in log)			
	OLS	Worker FE	Worker-plant FE
	(1)	(2)	(3)
Illiterate (reference category)			
Incomplete primary school	0.0596***	-0.0121**	-0.0096
	(0.0046)	(0.0057)	(0.0061)
Incomplete high school	0.1788***	-0.0137**	-0.0113*
	(0.0047)	(0.0057)	(0.0061)
Complete high school	0.3119***	-0.0153***	-0.0100
	(0.0047)	(0.0057)	(0.0061)
Incomplete college	0.6300***	0.0035	0.0026
	(0.0050)	(0.0061)	(0.0065)
College degree or more	0.8600***	0.0391***	0.0273***
	(0.0049)	(0.0059)	(0.0063)
Worker-level controls	Yes	Yes	Yes
Plant-level controls	Yes	Yes	Yes
Industry × year FE	Yes	Yes	Yes
Metropolitan region FE	Yes	Yes	Yes
F-stat.	21,947.88	3,399.19	–
R squared	0.6407	0.3828	0.3626

Notes: This table presents the estimates obtained from equation 4.9. All models are estimated with 2,387,434 observations. Worker-level controls include all the individual characteristics detailed above. Plant-level controls are dummies for plant size. Industry × year effect are dummies for each 2-digit × year combination. Metropolitan region FE are metropolitan region fixed effects. Standard errors adjusted for clustering are in parentheses. Significance levels: *** $p < 0.01$.
Source: Prepared by the author based on estimates.