INDUSTRIAL SCOPE OF AGGLOMERATION ECONOMIES IN BRAZIL

Ana Maria Bonomi Barufi
Eduardo A. Haddad
Peter Nijkamp

Industrial Scope of Agglomeration Economies in Brazil

Ana Maria Bonomi Barufi, Eduardo A. Haddad, Peter Nijkamp

Abstract. The tendency towards urbanization in the emerging world accompanied by the constant pursuit for higher productivity prompts an urge for studies aiming at understanding agglomeration economies. In the context of Brazil, a country with extremely high regional disparities, exploring this issue is important not only for private stakeholders, but also for public policy practitioners. In the framework of static agglomeration effects, we investigate the industrial scope of agglomeration economies in Brazil. On the basis of identified registration data covering the whole formal labor market in three distinct years (2004, 2008 and 2012), we estimate separate models for the logarithm of the hourly individual wage for five broad economic sectors (S1 – Manufacturing low-tech, S2 – Manufacturing medium-tech, S3 – Manufacturing high-tech, S4 – Services less-knowledge, and S5 – Services high-knowledge). Different estimation strategies are considered in a two-stage mode: with and without individual fixed effects in the first stage, and with and without instrumental variables for population density in the second stage. The main results indicate that there is not a unique optimal local industrial mix to foster productivity in different technological sectors. Comparing possible theoretical approaches (MAR, Jacobs, Porter) related to combinations of diversity, specialization and competition, we find that for S5 only diversity is significant (and positive), suggesting that a Jacobs' perspective is rather adequate. For S1, S2 and S4, the MAR framework seems more adequate to explain the underlying patterns. In the case of S3, there are elements from both Marshall's and Jacobs' perspectives. These results seem to be robust to different specifications and estimation strategies. Finally, the urbanization economies coefficient appears to be positive and significant for all sectors, ranging from 0.0511 to 0.0940 in different specifications, under the simplest estimation (OLS in the first and the second stages). Ordering these effects between the sectors from the highest to the lowest, we find the following sequence: S3, S1, S5, S4 and S2. This can be considered as evidence that high-tech and low-tech manufacturing sectors benefit more from the urban or metropolitan scale in Brazil, followed by services associated with higher knowledge intensity.

1. Introduction

Brazil has recently shown an important decrease in personal and regional inequality. In spite of that, its income inequality level is still one of the highest worldwide (16th highest Gini index among 141 countries[1]), reaching 0.526 in 2012 from 0.603 in 1995. Regional income disparities are also considerably high, summarized by the fact that average personal income was 1.8 times higher in the South-Southeast than in the North-Northeast in 2012.

---

[1] Considering the Brazilian Gini Index of 2012 and the most recent information for other countries, available at https://www.cia.gov/library/publications/the-world-factbook/rankorder/2172rank.html and accessed in 27/06/2014.

Following the tendency towards urbanization observed all over the world (Glaeser, 2011), there is a significant spatial concentration of population and economic activity in Brazil. In 2010, 84.4% of the population was in urban areas, occupying 1.07% of the territory.[2] Regarding economic concentration, in 2012, cities with more than 500,000 inhabitants held a share of 40.8% of total GDP, while concentrating 29.3% of the population.[3] Considering the correlation between the logarithm of wages and the logarithm of population density at the municipal level, it reached 0.06 in 2010 for the whole labor market, and 0.05 for the formal sector.

Moreover, it is widely acknowledged that living costs increase significantly in bigger cities, meaning that only with a positive relationship between wages and the city size it is possible to explain the existence of cities (Duranton and Puga, 2014; van Oort and Lambooy, 2014). In this context, the proximity of economic agents generates agglomeration economies by reducing transportation costs. Different theoretical approaches analyze these costs reductions under different perspectives – of goods, people and ideas (Glaeser and Gottlieb, 2009).

Following the substantial empirical literature on agglomeration economies, this paper aims to understand the nature and magnitude of regional disparities in Brazil by exploring the industrial scope of the extraordinary gains obtained in bigger cities. The strategy followed here presupposes that salaries are directly related to productivity, and investigates whether competition, specialization and/or diversity are more or less relevant to explain the regional residual variation in salaries in different sectors (which is not related to individual characteristics).

In comparison with the previous literature, we provide different estimation strategies, controlling or not for individual unobserved characteristics constant in time (Glaeser and Maré, 2001; Combes, Duranton and Gobillon, 2008; Combes et al., 2010, Groot et al., 2014). Moreover, we consider a two-stage estimation in two different levels (the first stage is at the individual level while the second stage is at the sector-region-year level).

Concerning the empirical literature on agglomeration economies in Brazil, we provide an analysis at the micro-level, controlling for individual characteristics. In addition, we compare

---

[2] Based on data from the Demographic Census of 2010, IBGE.
[3] With information from the Municipal GDP, provided by IBGE.

the potential effects of specific sources of externalities related to the industrial scope of agglomeration economies (diversity, specialization and competition) over sectors with different levels of technology and knowledge intensity. Finally, it is also possible to discuss how the formalization of the Brazilian labor market may have affected all these relations by analyzing the formal sector in two different time spans.

This study aims to contribute to the literature of static agglomeration effects, controlling for individual skills and comparing different estimation strategies. The focus is on the industrial scope of agglomeration economies, a strand of the literature that went back on attracting more attention in the recent period because individual-level longitudinal data became more easily available. Moreover, considering that most empirical studies focused on developed countries, evaluating the relative importance of the local industrial mix for different sectors in the context of a developing economy such as Brazil is another important contribution.

The next sections are organized as follows: Section 2 presents a review of the related literature; Section 3 discusses the methodological approach; Section 4 describes the data; finally, Section 5 brings the results, while Section 6 draws the main conclusions from the analysis.

2. Urban Wage Premium and Increasing Returns

The relationship between city size and the income of firms and workers has been examined under the interpretation that urbanization reflects the gains from agglomeration. The next subsections will briefly discuss the main issues investigated in the literature and explore the advances of the empirical studies in this area.

2.1. Principles of Agglomeration Economies

The urban wage premium has been the subject of analysis in several studies aiming to identify how the density of economic activity affects the productivity of workers (Heuermann et al., 2010). There is usually a positive association between city size and salaries (Combes et al., 2010). However, cities present a trade-off between costs and benefits, which are directly related to dispersion and agglomeration forces (Glaeser, 1998; Combes et al., 2011; Duranton and Puga, 2014). The challenge is to isolate these effects from other explanatory factors of

productivity differentials in space, while investigating the possibility of convergence and divergence between regions, sectors, educational levels, among others dimensions ((Lindley and Machin, 2014).

Spatial wage differentials can be explained by three main sets of variables (Combes, Duranton and Gobillon, 2008): skill composition of the local labor force; geographical characteristics and local factors of production; and interactions between firms or workers. The latter is related to the generation of agglomeration economies.

In fact, under the conditions defined by the Spatial Impossibility Theorem (Starret, 1978), increasing returns constitute one of the possible explanations for the existence of cities with trade flows between them. When there are positive externalities to the urban size, the competitive paradigm in space is no longer valid (Combes, Mayer and Thisse, 2008), creating incentives for the agglomeration of economic agents (Fujita and Thisse, 2012; Glaeser, 1998).

Agglomeration economies are generated by the interactions between firms and individuals. Synthetizing different contributions, Duranton and Puga (2004) and Puga (2010) identify three main micro-foundations: (i) **sharing** of facilities, and gains from individual specialization and variety, and risk sharing, through labor pooling; (ii) higher probability of a **matching**, with higher quality; and (iii) **learning**, related to knowledge generation, diffusion, and accumulation. These externalities are static when they explain the cross-sectional distribution of economic activity, and are dynamic when related to productivity growth and knowledge spillovers (De Groot et al., 2009; Pessoa, 2014). According to Rosenthal and Strange (2004) they may be related to the city size (urbanization economies, associated to Jacobian economies) or to the size of the own economic sector (localization economies, linked to Marshallian economies).

The micro-foundations discussed above come from a reinterpretation of Marshall (1890), who pointed out that proximity gains could be stronger between firms and workers within the same industry (labor market pooling, input sharing and knowledge exchange). Glaeser et al. (1992) summarized further developments in this direction with the MAR model (Marshall-Arrow-Romer), in which industrial specialization could promote knowledge spillovers by favouring the interaction between firms with a similar structure. Moreover, it may be easier

4

for firms to internalize innovation gains when they have a higher local monopolistic power (Capello, 2014). Another perspective is brought by Jacobs (1969), according to whom knowledge externalities are fostered by the interaction between economic actors of different sectors. Therefore, the source of agglomeration economies is external to the industry, allowing the exchange of complementary knowledge, generating urbanization economies. In this setting, competition is seen as an incentive for firms to innovate. Porter (1990) also defends this last aspect, but his theoretical framework advocates that a higher level of specialization reinforce those gains (knowledge exchange is stronger in vertically integrated industries).

Depending on the development stage of each industry in a certain country, they will benefit from different industrial compositions in the local level. For instance, new sectors with a strong focus on innovation efforts may be located in larger and diversified urban areas in which they can profit from a wide range of experiences and ideas (Henderson, 2010). Following De Groot et al. (2009), it is possible to organize the sources of externalities and the expected effects over employment and productivity according to each theoretical view discussed above.

Table 1. The Effect of Agglomeration Externalities on Employment and Productivity

Type		Measure	Effect on employment growth and productivity		
Static	Localisation externality	Geography, infrastructure	+		
	Urbanisation externality	Demand, population size	+		
			MAR (Marshall-Arrow-Romer)	Jacobs	Porter
Dynamic	Knowledge externalities	Specialisation	+	-	+
		Competition	-	+	+
		Diversity	-	+	-

<u>Source</u>: De Groot et al. (2009).

This strand of literature was initially synthetized during the 1960s and the 1970s (Capello, 2014). However, there has been a recent new spur of theoretical and empirical studies covering the relationship of productivity, geographical proximity and local industrial composition. From the perspective of the empirical literature, detailed databases based on micro data at the individual level allowed the brand new theoretical models to be tested by controlling for individual heterogeneity to assess regional-level relationships (Rosenthal and

Strange, 2004). In the Brazilian context, the empirical literature with this recent approach is still very limited, as it will be discussed in the next subsection.

Therefore, the industrial scope of agglomeration economies can be further investigated by controlling for individual characteristics and other confounding factors. This strategy allows the identification of a proper local sector mix, the optimum city size, and the adequate competition degree for different industries. This framework will be considered in the case of a developing economy with continental dimension and huge regional disparities, Brazil.

2.2. Empirical Studies on Agglomeration Externalities

The literature relating productivity and different measures of agglomeration externalities has been reviewed elsewhere (Melo et al., 2009, Beaudry and Schiffauerova, 2009). In general, results are significantly heterogeneous, and one of the explanations is that the outcome related to productivity can vary. It can be either directly derived from the production function or it can be indirectly derived (salaries, rents, job creation, new establishments, among others), according to data availability. When salaries are considered, the main underlying assumption is that they equalize the marginal productivity of labor (under perfect competition), or that at least they are higher in places that are more productive (Rosenthal and Strange, 2004). An important drawback from this strategy is that elasticities in the labor market will determine the extent through which wages capture local productivity (Moretti, 2011).

Apart from that, heterogeneous results found in the literature can be associated to different measures for competition, diversity and specialization (Beaudry and Schiffauerova, 2009). Melo et al. (2009) also make a remark that the estimates of the elasticity of urbanization economies vary among countries, data structure, geographical aggregation and economic sector, in a range between 3% and 8% (an increase of 1% on urban density, for instance, is expected to increase wages from 3% to 8%).

The theoretical framework most commonly applied to this sort of analysis is based on the wage equation of the urban economics literature (Combes, Mayer and Thisse, 2008). One of the first studies aimed at measuring the urban wage premium while controlling for individual heterogeneity was conducted by Glaeser and Maré (2001). They consider longitudinal

individual data for the United States and find evidence that there is an urban premium even when observed and unobserved individual characteristics are taken into account. In a similar setting, Mion and Naticchioni (2009) show that individual abilities and firm size explain a large share of the spatial heterogeneity of wages in Italy.

Combes, Duranton and Gobillon (2008) find that controlling for the sorting of individuals (with individual fixed effects) reduces the estimated coefficient of urbanization externalities by 40 to 50 percent. However, apart from creating a black box for the sorting process, the inclusion of individual fixed effects may introduce a selection bias, as only individuals who have migrated will provide the variation to estimate the agglomeration coefficient. Furthermore, pooled cross-sections will provide an upper bound for the estimates, while the panel estimation will generate a lower bound (Groot et al., 2014).

There is a strong concern for potential endogeneity of urban size measures (endogenous quantity of labor), such as population density (Combes et al., 2011). This is so because cities with higher wages may attract more individuals, leading to an increase in population and consequently, on city size. The approach proposed by Ciccone and Hall (1996) and adopted by Combes, Duranton and Gobillon (2008), Groot et al. (2014), Graham et al. (2010), among others, is based on long lags of the endogenous variable, while Combes et al. (2010) also suggest the inclusion of geographical characteristics as instruments. On the other hand, the individual fixed effects mentioned above aim to control for the endogenous quality of labor, namely the sorting process of the labor force (Glaeser and Maré, 2001; Combes, Duranton and Gobillon, 2008).

As previously suggested, the industrial scope of agglomeration externalities can be evaluated in order to assess which theoretical framework seems more suitable to explain regional wage disparities. Glaeser et al. (1992) explore city-level data in the United States, finding results consistent with the framework presented by Jacobs (1969) – employment growth is encouraged by urban diversity and local competition. Groot et al. (2014) estimate a two-stage regression, explaining the spatial residual of a mincerian regression. The authors find an employment density elasticity of 4.8 percent for NUTS-3 areas in the Netherlands, with stronger indication of MAR externalities and small negative effects of competition and diversity measures.

Longitudinal data on the labor market in Brazil has become available only recently, and access is still very limited. Therefore, a large part of the empirical literature is based on individual-level cross sections or aggregated data at the area level. In an attempt to measure agglomeration externalities for the country, Henderson's (1986) estimates indicate that localization economies seemed more important than urbanization economies in the 1970s, meaning that manufacturing activities might be less productive in larger urban areas. Moreover, the mean of the coefficients estimated for different sectors is 0.046, ranging from 0.003 to 0.18.

Amaral et al. (2010) estimate a New Economic Geography (NEG) wage equation and find a positive relationship between wages and market potential. Also based on a NEG model, Fally et al. (2010) find that market potential and supply access (intermediary goods) are positively related to individual wages. Silva and Silveira-Neto (2009) explore the determinants of manufacture employment growth at the state-sector level in Brazil between 1994 and 2004. Among the explanatory variables, there are the average wage, average firm size, connections (concentration index based on backward and forward linkages), a proxy for transportation costs, and dynamic externalities (specialization and diversity). They find indication that competition is relevant for employment growth (Jacobs and Porter theoretical arguments), as well as diversity (Jacobs).

Hierarchical wage equations are applied by Fontes et al. (2010) to control for variables in different levels (individual and territorial unit of analysis). Based on census data for 1991 and 2000 (cross-sections), the authors evaluate municipalities with at least 50,000 inhabitants and find significant regional disparities even after controlling for observed individual heterogeneity. There is evidence of a positive effect of urban scale, as well as of a positive relation between industrial concentration and density of modern productive services over wages.

With an urban economics wage equation, Simões and Freitas (2014) apply municipal data to find that urbanization economies are more relevant for high technological intensity sectors, while sectors with low and medium technological intensity are more benefited from mid-sized urban centers, relatively less diversified. Considering data from RAIS (Annual Report of Social Information, from the Ministry of Labor), Freguglia and Menezes-Filho (2012) find that when controlling for individual heterogeneity, almost 63 percent of the total spatial

differential disappears, meaning that local policies should focus on human capital development in order to promote regional development.

The investigation of the industrial scope of agglomeration economies controlling for individual heterogeneity has become the subject of analysis of recent studies, due to an increase in the number of available longitudinal individual-level databases. There is room to contribute with this literature, especially in the context of a developing country such as Brazil, where empirical studies started to cover these possibilities only recently.

3. Analytical Framework

The theoretical framework adopted here is based on the wage equation developed by the urban economics literature. This approach is not new, especially in its aggregated regional version, being present in a variety of studies ever since the 1970s. However, the recent spur of newly available individual-level longitudinal databases allowed a proper control of individual heterogeneity. Following Combes, Mayer and Thisse (2008), profit optimization by a price-taking firm j may generate the following equilibrium wage:

$$w_j = \mu(1-\mu)^{(1-\mu)/\mu} s_j \left(\frac{p_j A_j}{r_j^{1-\mu}}\right)^{1/\mu} = f\big(s_j, A_j, p_j, r_j\big) \qquad (1)$$

In this case, the estimation can be conducted for individual i working in firm j, requiring detailed information at the individual level. While p_j and r_j capture agglomeration and dispersion forces, respectively. The average unit value p_j of the good produced by the firm can be raised by a higher demand, weaker competition or cheaper intermediate goods, leading to more agglomeration. On the other hand, r_j measures the effects of the prices of other production inputs, which can increase with congestion and may provide a dispersion force. The term A_j is related to technological externalities (knowledge and learning spillovers, existing technology associated to workers abilities, among others). Finally, it is necessary to control for individual skills s_j in order to capture the correct effect of local characteristics over salaries.

When wages are regressed against a measure of city size within this context, such as population density, the main underlying assumption is that this relationship happens through the elements discussed above. It is also important to control for all possible confounding variables. Moreover, the industrial mix must be taken into consideration, especially because local wages also vary between sectors, which are heterogeneously affected by the degree of diversity, specialization and concentration. These last factors are embedded in the term measuring technological externalities, namely, A_j.

Combining all these elements, the next sections will discuss the estimation of the equation presented above in two stages. In the first stage (Equation 2), individual-level data is used to explain wages with individual characteristics and a dummy for each combination of region-sector-year, the spatial residual (Groot et al., 2014) – henceforth spatial wage.

$$\ln(w_{i,r,s,t}) = \beta_0 + \beta_1 age_{i,r,s,t} + \beta_2 age^2_{i,r,s,t} + \sum_{edu} \beta_4 D^{edu}_{i,r,s} + \sum_{factor} \theta_{factor} Skill^{factor}_{i,r,s,t}$$
$$+ \sum_{firm\_size} \beta_5 D^{firm\_size}_{i,r,s,t} + \sum_r \sum_s \sum_t \lambda_{r,s,t} + \varepsilon_{i,t} \quad (2)$$

Then, this estimated spatial wage $\hat{\lambda}_{r,s,t}$ will be used as the dependent variable in the second stage (Equation 3), being associated with measures of agglomeration externalities and geographical characteristics.

$$\hat{\lambda}_{r,s,t} = \delta_0 + \delta_1 \ln(density_{r,t}) + \delta_2 area_{r,t} + \delta_3 specialization_{r,s,t} + \delta_4 diversity_{r,t}$$
$$+ \delta_5 competition_{r,s,t} + \sum_{geog} \delta_{geog} geography_{r,t} + \sum_s \delta_6 D^{sector}_s \quad (3)$$
$$+ \sum_t \delta_7 D^{year}_t + \sum_{macro-region} \delta_8 D^{macro-region}_{mr} + \xi_{r,s,t}$$

Such a strategy in two stages is based on the argument that individuals residing in the same labor market and working in the same sector may share some specific characteristics that are not entirely captured by the controls included in Equation 2 and Equation 3. In this case, the error term of a single-stage estimation will be positively correlated across individuals from the same sector-region, generating downwardly biased standard errors for the regional and sectoral level variables, generating higher t statistics (Combes, Duranton and Gobillon, 2008;

Moulton, 1990, Bell et al., 2002 and Card, 1995, in a different context). A two-stage estimation procedure aims to overcome this issue.

A descriptive analysis of all variables included in the model will be presented in the next section. Among the individual characteristics considered in Equation 2, $w_{i,r,s,t}$ is the hourly wage, $age_{i,r,s,t}$ is measured in years, and $D_{i,r,s}^{edu}$ assumes value 1 when the individual has attained a certain educational level (edu = {illiterate, incomplete primary school, complete primary school to incomplete high school, complete high school to incomplete college, college degree or more}) or 0 otherwise. Moreover, $Skill_{i,r,s,t}^{factor}$ measures the degree of certain skills that are required for the job, following Maciente (2013).[4] Finally, $D_{i,r,s,t}^{firm\_size}$ represents a group of dummies for each firm size in which the individual works (aiming to control for technological patterns and competition differences between firms, Glaeser et al., 1992), and $\lambda_{r,s,t}$ is the sector-spatial wage.

Equation 3 presents a set of independent variables aiming to measure how the city size and the industrial mix affect the spatial wage. Urbanization externalities are measured by the logarithm of employment density, and the area of the region[5] is included to control for the scale effect. In addition, different measures of specialization, diversity and competition are considered in alternative specifications (see Table A.1). Other controls include two-digit-sector dummies, year and macro-region dummies, as well as geographical characteristics (altitude and distance to the Equator line).

The estimation strategies have the following variations: in the first stage, an individual fixed effect may be included to control for the unobserved individual heterogeneity constant in time. Furthermore, in the second stage, the potential endogeneity of employment density is taken into account with instrumental variables techniques. The instruments considered here are the following: population density in 1940, distance to the coast, and dummies for sugarcane and coffee production, and gold exploration during the colonial period.[6] Table A.2

---

[4] See Annex 4.

[5] Regions are labor market areas (REGIC areas), which aggregate municipalities in regions of immediate articulation (482 regions containing all 5,565 municipalities), defined by the Brazilian Institute of Geography and Statistics for 2007 (IBGE, 2013), taking into account all daily commuting and transportation connections among municipalities.

[6] These last three variables were generated and kindly provided by Naritomi et al. (2012). In the 1940s, the Brazilian economy was much less industrialized and productivity differentials were more related to agriculture. Therefore, population density in that decade is expected to be exogenous in relation to wages in the twenty-first century. Concerning distance to the coast, the colonization process in Brazil was mostly concentrated in the coast, and a simple analysis of the distribution of cities in Brazil shows that there is still a high population

provides a detailed description of all variables considered, including methods of calculation and data sources.

In relation to specialization, diversity and competition, there is a handful of measures applied in the literature. Even if it is possible to classify their suitability according to the problem at hand, it is interesting to compare the results of different combinations of measures. Table A.3 in Annex 3 provides a list of the alternative variables considered here, with their respective method of calculation, interpretation and source. Evaluating all of them increases the robustness of the results.

The last methodological issue to be highlighted is that the regressions presented in Equations 2 and 3 are estimated for five aggregated sectors separately, defined in Annex 1. The reason of doing this is that it is possible to compare the effect of agglomeration externalities and the industrial mix over different development stages of manufacture and service industries, based on their technological and knowledge intensity. It also reduces the computational complexity of the estimation of the spatial wage.[7] The major drawback here is that the balanced panels consider only individuals who do not move between these five aggregated sectors over time, only within the industries belonging to each of them. Therefore, part of individuals who move between sectors is not observed. As the group which moves is usually associated to a higher productivity, it is possible that the estimations will be downwardly biased.

4. Data

The estimation of the wage equation discussed above is based on data of the Brazilian formal labor market. For this purpose, the main database considered here is the Identified RAIS (Annual Report of Social Information, from the Ministry of Labor), which consists of identified registration data of all formal firms and their employees, focusing on the characteristics of the work contract. It provides annual information on all formally employed

---

concentration in this part of the country. Finally, the main economic activities developed during the colonial period helped defining the location of many urban centers, but are not necessarily related to the determinants of productivity nowadays.

[7] In the case when the combination region-sector-year is considered for all industries simultaneously, the limit of Stata MP and R in a computer with RAM of 64G is reached.

individuals in the private sector (or part of the public sector, depending on the type of contract), with a significant regional disaggregation (municipal level).[8]

Among the main advantages of this database, there is the fact that it has a longitudinal structure at both the firm and individual levels. Moreover, as it is a mandatory report, all firms are supposed to fill in the required information, covering the entire formal sector. This is also the reason why, differently from individual self-reporting surveys, there is a smaller risk of wage under-reporting. A potential drawback of using this database is that it covers only the formal sector and it does not characterize entrepreneurs (who are self-employed) or employers. Another relevant detail is that the reporting process is more accurate for bigger firms, usually located in larger urban areas, and sometimes firms with multiple branches may report incorrectly the municipality of each employee.

Following the process of formalization of the labor market, the size of RAIS has increased significantly in the past decade. In 2004, 30.3% of total jobs[9] were formal, while in 2012 this percentage increased to 39.3%. It is also important to notice that there is a significant regional heterogeneity in this process. In 2004, only 17.5% of total jobs were formal in the North region of the country and 17.1% in the Northeast, while 39.4% in the Southeast, 35.1% in the South and 28.6% in the Center-West. Then, in 2012, these percentages went up to 23.6% in the North and 25.4% in the Northeast, while reaching 48.2% in the Southeast, 44.9% in the South and 40.2% in the Center-West.[10]

This issue is taken into account by comparing the relationships studied here over time (2004, 2008 and 2012). In order to explore the longitudinal structure of the database, a few steps were conducted to generate a balanced panel for each period (2004-2008; 2008-2012; 2004-2008-2012), and the comparison between the first two datasets allows the discussion of the potential effect of formalization over the characteristics of the formal labor force.[11] Finally,

---

[8] The authors are thankful for FIPE (The Economic Research Foundation Institute), Prof. Helio Zylberstajn and Eduardo Zylberstajn for giving access to the database. All confidentiality requirements were respected.

[9] Total jobs include public sector and military, formal and informal employees, self-employment, entrepreneurs, and unpaid work. Therefore, the percentages discussed here refer to the representativeness of RAIS in the labor market.

[10] Information obtained from the National Household Sample Survey (PNAD), from the Brazilian Insititute of Geography and Statistics (IBGE).

[11] The original database contains individual information for each contract, meaning that individuals can appear more than once if they work for more than one firm. A few steps were taken in order to select the desired observations (ending up with only one contract for worker). Firstly, a filter selected only active contracts in

database tractability was obtained by generating random samples of size 20% for each of the five technological sectors.[12]

Descriptive statistics of the main variables of the first stage for the sample based on the period 2004-2008-2012 in 2012 are presented in Table 2, at the individual level. The average wage increases for sectors with higher technological or knowledge intensities. While Manufacturing high-tech (S3) and Services high-knowledge (S5) concentrate relatively more workers with higher education attainment, Manufacturing medium-tech (S2) workers are more present in the groups of incomplete and complete middle school. As expected, Manufacturing low-tech (S1) and Services less-knowledge (S4) are characterized by less qualified workers (up to incomplete primary school).

In terms of skills required for the job, Cognitive skills are more relevant for sectors with higher intensity of knowledge or technology (S3 and S5). The same pattern is observed for Telecommunication, Attention, On-the-job experience, and Team-work skills. In the case of service sectors (S4 and S5), they concentrate more Assistance, Transportation, Artistic, Conflict management, Sales and Monitoring skills, while manufacturing sectors (S1, S2 and S3) are associated to Maintenance and operation, Design and engineering, Accuracy and automation and Independence skills. Age does not seem to change significantly among these groups of sectors, except for S4, which seems slightly bigger, but still with a considerably high standard deviation.

December of each year for individuals working for private companies in permanent jobs. Then, contracts with missing individual ids, wages equal to zero or with less than 20 weekly hours were excluded. The next steps for individuals with multiple contracts included keeping those with 5 or less contracts, dropping those with different gender in each contract, and keeping the contracts with the highest number of weekly hours and with the oldest hiring date. Finally, for the remaining cases of multiple contracts, only one of them was randomly selected, leading to one contract per individual for the whole database. After all these procedures and keeping only manufacture and service sectors, the database size ranged from 8.9 to 12.6 million between 2004 and 2012 (the increase is related to the formal sector expansion previously mentioned). Only individuals observed in each period (2004 and 2008; 2008 and 2012; and 2004, 2008 and 2012) were kept, with the same gender and birth date in every year. Finally, following a common practice in the related literature, the analysis was conducted for men with 18 to 56 years old in 2004.

[12] These samples are representative for the following characteristics: age group (less than 30 years old, 30 to 45 years old, 46 or more), population size (less than 100 thousand, 100 thousand to less than 500 thousand, 500 thousand to less than 1 million, 1 million or more), firm size (up to 4 employees, 5 to 9, 10 to 19, 20 to 49, 50 to 99, 100 to 249, 250 to 499, 500 to 999, 1,000 employees or more), macro-region, educational level (illiterate, incomplete primary school, complete primary school to incomplete high school, complete high school to incomplete college, college degree or more), and technological sector (see Annex 1).

Table 2. Descriptive Statistics of the Main Variables of the First Stage of the Model, 2012 (for the database containing individuals present in the whole period, 2004-08-12)

	S1 Manufacturing low-tech		S2 Manufacturing medium-tech		S3 Manufacturing high-tech		S4 Services less-knowledge		S5 Services high-knowledge	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
ln(real hourly wage)	2.36	0.66	2.67	0.69	3.08	0.73	2.24	0.58	2.60	0.92
age	40.26	8.93	41.03	8.63	40.86	8.49	43.36	9.17	41.38	8.72
Factor 1 - Cognitive skills	-0.63	0.74	-0.59	0.67	-0.34	0.77	-0.81	0.70	-0.23	0.98
Factor 2 - Maintenance and operation skills	0.47	0.93	0.69	0.87	0.48	0.95	-0.11	0.91	-0.51	0.77
Factor 3 - Assistance skills	-0.43	0.45	-0.47	0.44	-0.45	0.55	0.21	0.66	0.08	0.74
Factor 4 - Management skills	-0.16	0.90	-0.24	0.83	-0.21	0.96	-0.28	0.96	-0.05	1.17
Factor 5 - Design and engineering skills	0.03	0.69	0.20	0.75	0.39	0.93	-0.50	0.57	-0.42	1.06
Factor 6 - Transportation skills	-0.17	1.11	-0.17	1.00	-0.38	0.78	1.08	1.84	0.33	1.08
Factor 7 - Artistic skills	-0.32	0.68	-0.38	0.66	-0.40	0.67	-0.21	0.65	-0.16	0.83
Factor 8 - Accuracy and automation skills	0.05	0.72	0.12	0.69	0.02	0.73	-0.13	0.61	-0.05	0.86
Factor 9 - Supervised work skills	0.00	0.89	0.31	0.83	0.40	0.80	0.03	0.91	0.26	0.86
Factor 10 - Teaching and social science skills	-0.23	0.48	-0.18	0.44	-0.25	0.47	-0.20	0.50	-0.33	0.77
Factor 11 - Physical strength	0.15	0.80	0.16	0.82	-0.07	0.79	0.35	0.91	0.05	0.86
Factor 12 - Telecommunication skills	-0.29	0.86	-0.25	0.90	-0.02	0.94	-0.19	0.72	0.08	1.07
Factor 13 - Independence skills	-0.10	1.01	0.12	0.90	0.19	0.84	-0.62	1.22	-0.52	1.41
Factor 14 - Natural science skills	-0.16	0.81	-0.08	0.80	-0.15	0.79	-0.29	0.57	-0.12	0.87
Factor 15 - Attention skills	-0.19	0.81	-0.16	0.82	-0.06	0.89	-0.30	0.77	-0.04	0.91
Factor 16 - On-the-job experience	-0.32	0.90	-0.19	0.92	-0.11	0.89	-0.44	0.72	-0.10	0.84
Factor 17 - Conflict management skills	-0.30	0.83	-0.29	0.76	-0.26	0.75	0.28	0.76	0.50	1.04
Factor 18 - Team-work skills	-0.22	0.91	-0.13	0.85	0.00	0.74	-0.13	0.90	0.17	0.68
Factor 19 - Sales skills	-0.51	0.89	-0.58	0.82	-0.40	0.86	0.15	0.77	0.01	0.90
Factor 20 - Monitoring and compliance skills	-0.22	0.95	-0.11	0.82	-0.24	0.83	-0.08	1.21	-0.25	0.86
Factor 21 - Clerical skills	-0.38	1.13	-0.49	1.12	-0.17	1.04	-0.02	0.90	0.22	1.01
Education attainment										
Illiterate	0.5%		0.3%		0.1%		0.4%		0.2%	
Incomplete primary school	27.8%		21.1%		10.9%		31.3%		11.7%	
Complete primary school - incompl. middle school	26.0%		23.7%		18.4%		30.7%		18.1%	
Complete middle school - incomplete college	39.2%		46.2%		50.2%		33.0%		41.3%	
College degree or more	6.6%		8.8%		20.5%		4.7%		28.8%	
Firm size										
Up to 4 employees	2.7%		2.0%		0.8%		11.2%		2.5%	
5 to 9	4.0%		3.7%		1.4%		15.8%		3.8%	
10 to 19	6.6%		7.6%		3.6%		10.9%		5.7%	
20 to 49	10.6%		12.4%		6.8%		14.8%		9.6%	
50 to 99	9.4%		12.1%		8.2%		9.6%		7.5%	
100 to 249	14.8%		15.7%		14.9%		11.1%		11.3%	
250 to 499	14.2%		15.3%		14.9%		8.8%		11.8%	
500 to 999	12.7%		13.8%		13.6%		8.4%		12.8%	
1,000 or more employees	25.0%		17.3%		35.8%		9.6%		35.1%	
Macro-region										
North	1.7%		1.3%		1.6%		2.1%		2.6%	
Northeast	14.7%		7.0%		2.8%		13.9%		13.2%	
Southeast	49.6%		70.2%		76.0%		66.8%		67.7%	
South	29.6%		20.0%		18.9%		12.8%		11.1%	
Center-West	4.4%		1.5%		0.7%		4.4%		5.4%	
Variation in the panel structure										
Individuals in the same industry and Area REGIC	85.4%		86.5%		83.7%		80.7%		69.1%	
Individuals in different industries	5.5%		4.1%		4.3%		7.2%		13.4%	
Individuals in different Areas REGIC	5.4%		6.9%		8.9%		6.7%		11.2%	
Individuals in different industries and Areas REGIC	3.7%		2.5%		3.2%		5.4%		6.3%	
Percentage of areas REGIC (total = 482)	89.2%		81.1%		56.0%		94.0%		88.4%	
Number of observations	45,902		34,487		37,840		72,014		58,350	

Source: Elaboration of the authors.

In relation to their regional distribution, both service sectors (S4 and S5) present a similar distribution of workers in the sample in all regions, and the biggest disparities from this distribution are observed for manufacturing sectors. S3 and S2 are relatively more concentrated in the Southeast of the country (the richest region in terms of GDP), while S1 is more present in the Northeast. Comparing services and manufacturing regional distribution, the South concentrates more workers for S1, S2 and S3, while the Center-West is relatively less important for S2 and S3.

Firm size also varies among sectors, as there seems to exist a larger share of workers in bigger firms in S3 and S5, what indicates a need of scale to develop productive activities in these sectors, at least in 2012. Furthermore, the panel structure is extremely relevant to understand the results found in the next section. The share of individuals who do not change industry[13] or area REGIC is very high, over eighty percent for most sectors (except S5, in which it is slightly smaller than seventy percent). This is an indication that the sample to obtain the variation to identify the spatial wage when individual fixed effects are included will be relatively small. Table A.5 in Annex 5 presents a comparison of selected descriptive statistics. Apart from the fact that individuals who change industry (within the technological sectors) and/or area REGIC are more educated, their mean salaries can be even smaller than the full sample. This last element goes against the literature, which states that workers who move are positively selected and should be more productive.

Finally, it is not possible to observe individuals in the sample for all 482 areas REGIC of the country, because the sample of twenty percent restricts the analysis and the formal sector is not necessarily present in all areas of the country (especially when public administration is excluded). The share of areas is even smaller for S3, an indication that high-technology manufacturing industries are more concentrated in specific places of the country.

5. Results

In the previous section, there was a mention that three different datasets would be analyzed. Here the main results will discuss the case in which there are individuals present in the

---

[13] Within the same technological sector.

database for the years 2004, 2008 and 2012. At a certain point though, there will be a comparison with data for 2004-2008 and 2008-2012.

The first set of results refers to the first stage of the model (Equation 2), at the individual level (Table A.6). As mentioned in Section 3, there are two main alternatives considered here, a simple OLS estimation and a regression including individual fixed effects. Analyzing the OLS estimation version, it is noticeable that most coefficients are statistically significant for all sectors, with the relative importance of age increasing with technology and knowledge intensity (when comparing the five sets of regressions, higher coefficients for age are found for S3 and S5). For each sector, age is also growing at decreasing rates (age-squared coefficient with negative sign). In the case of education, all coefficients are positive and increase in value to higher levels of educational attainment. In fact, individuals with graduate or post-graduate degree obtain higher salaries than illiterate individuals (this difference ranges from 67 percent to 83 percent, depending on the sector). Only for service sectors, the two lowest educational levels are non-significant. The estimated coefficients for each skill seem to follow the expected signs. For instance, Cognitive skills are positively associated to wages in all sectors, while Transportation skills are positive only for S4 and S5 (services). On the opposite side, the Physical strength coefficient has a negative sign for most sectors (non-significant only for S1), what is in accordance with the literature. In the case of firm size, it is positively correlated to individual wage, meaning that bigger firms are supposedly more productive.

When individual fixed effects are considered (last five columns), education variables are omitted, as they do not offer sufficient variation over time for each individual. Another relevant result is that many variables loose significance (age and some skills), as they are also captured by fixed effects. However, some elements are still present, such as the positive and significant coefficients for Cognitive skills and increasing values with firm size. Both groups of regressions include also interactions dummies between areas REGIC, industries and years (as expressed in Equation 2). As mentioned in Section 3, the estimated coefficients of these dummies are directly used as dependent variables in the second stage, and are referred as the spatial wage.

While the first stage aimed to control for individual heterogeneity, the second stage will explore the relationship between the spatial wage and different measures of diversity,

specialization and competition, as well as urban size. Therefore, it will be possible to investigate the industrial scope of agglomeration economies. Table 3 presents a brief descriptive analysis of the main variables included in the second stage, at the area REGIC-industry level for 2012.

Table 3. Descriptive Statistics of the Main Variables for the Second Stage of the Model for 2012 (database 2004-2008-2012)

Dimensions	Variables	S1 Manufacturing low-tech		S2 Manufacturing medium-tech		S3 Manufacturing high-tech		S4 Services less-knowledge		S5 Services high-knowledge	
		mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
	Spatial wage (1st stage with OLS)	-0.154	0.426	0.438	0.837	-0.295	0.549	0.454	0.430	0.572	0.614
Urbanization	Density (1,000 inhabitants / km^2)	27.5	68.4	31.4	72.3	48.1	94.6	27.2	67.3	32.2	77.1
Diversity	Inverse of Herfindhal Index	10.76	3.88	11.41	3.73	13.00	3.29	10.49	3.87	10.62	3.94
	Shannon entropy	2.77	0.35	2.83	0.30	2.96	0.22	2.74	0.35	2.76	0.36
	Share 5 biggest industries (except the own industry)	0.58	0.11	0.57	0.10	0.53	0.08	0.60	0.11	0.59	0.11
Specialization	Location quotient	3.24	9.51	2.09	3.08	2.48	4.36	0.94	0.93	0.94	1.39
	Specialization degree	0.039	0.039	0.035	0.033	0.027	0.022	0.038	0.039	0.037	0.041
Competition	Competition degree	1.55	1.35	1.60	1.43	1.98	2.53	1.68	0.97	1.85	1.25

Obs. 1: The spatial wage presented here refers to the regression with an OLS estimation in the first stage.
Obs. 2: The unit of analysis is the combination area REGIC – industry.
<u>Source</u>: Elaboration of the authors.

The urbanization measure indicates that S3, S5 and S2, in this order, are associated to denser areas. Diversity mean increases with technology intensity for manufacturing sectors and with knowledge level for service ones. This pattern is observed even for the diversity measure related to the share of the five biggest industries because it is inversely related to diversity (a higher percentage indicates low diversity). In the case of specialization measures, notably for the specialization degree, this relationship is exactly the opposite (higher specialization is more relevant for lower technological and knowledge intensity sectors). The location quotient achieves a much lower value for service sectors, probably indicating that services are more homogeneously spread across the country. The competition degree shows a very similar pattern to diversity measures. Finally, the mean spatial wage seems higher in S5, S4 and S2, but it does not show a very clear pattern.[14] The maps below describe its regional distribution in detail.

---

[14] It is important to note that the spatial wage can present negative values, as it is a deviation against the omitted coefficient of the combination area REGIC-industry-year.

Figure 1. Regional Distribution of the Mean Logarithm of the Hourly Wage and the Spatial Wage (for OLS and FE in the first stage), by sector in 2012 (database 2004-2008-2012).

Sector 1 – Manufacturing low-tech
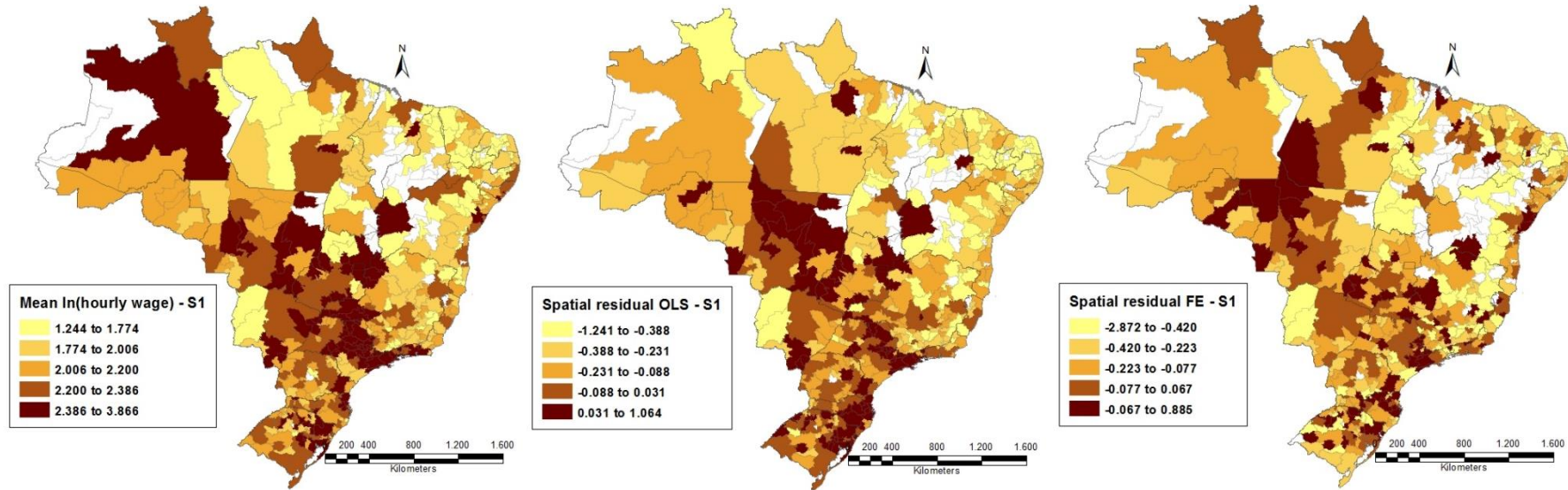
Sector 2 – Manufacturing medium-tech

19
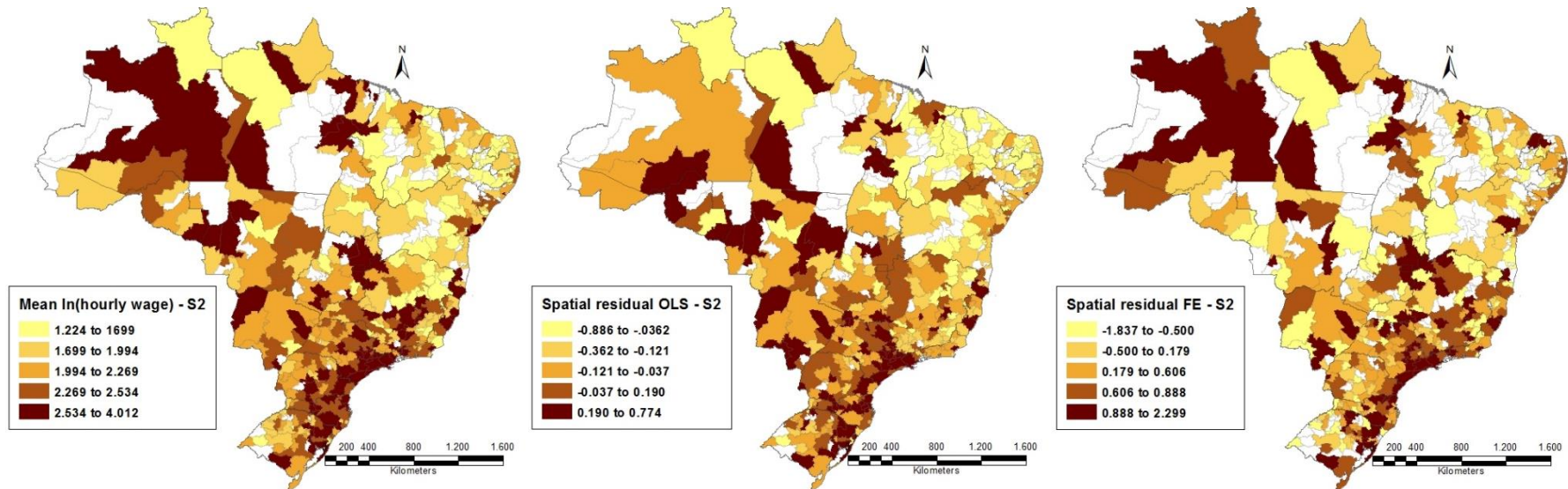
Sector 3 – Manufacturing high-tech

Mean ln(hourly wage) - S3
- 1.244 to 2.056
- 2.056 to 2.388
- 2.388 to 2.614
- 2.614 to 2.904
- 2.904 to 3.712

Spatial residual OLS - S3
- -1.912 to -0.750
- -0.750 to -0.551
- -0.551 to -0.383
- -0.383 to -0.221
- -0.221 to 0.464

Spatial residual FE - S3
- -3.125 to -0.666
- -0.666 to -0.385
- -0.385 to -0.210
- -0.210 to 0.038
- 0.038 to 1.031
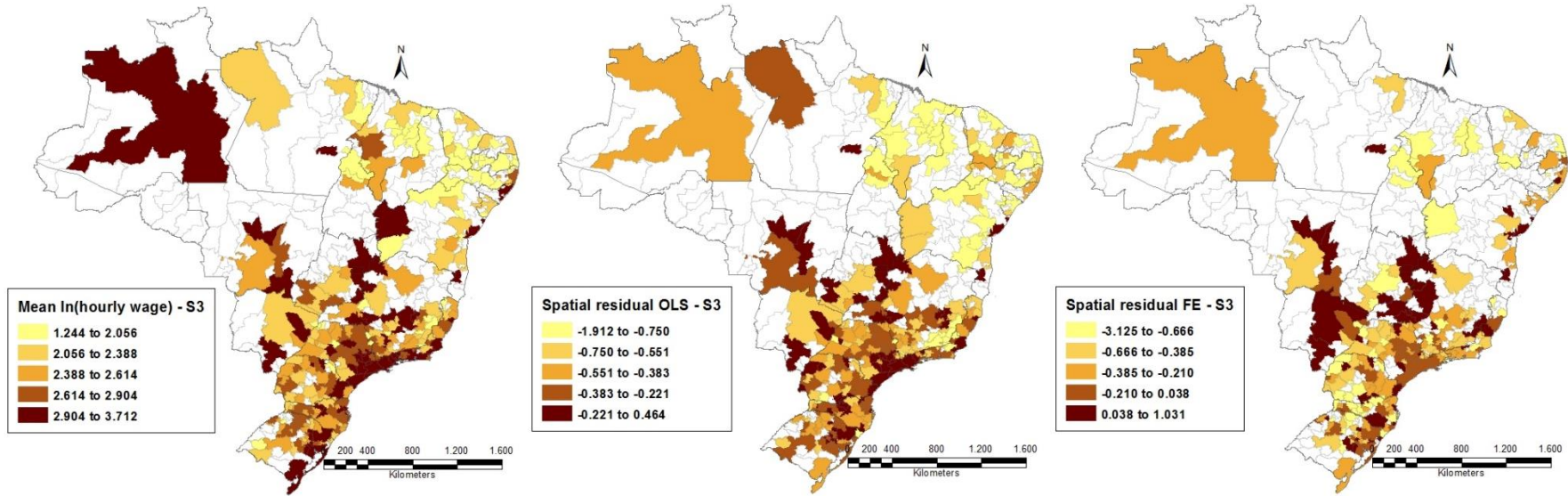
Sector 4 – Services less-knowledge

Mean ln(hourly wage) - S4
- 1.244 to 1.697
- 1.697 to 1.936
- 1.936 to 2.086
- 2.086 to 2.248
- 2.248 to 2.896

Spatial residual OLS - S4
- -0.614 to 0.269
- 0.269 to 0.431
- 0.431 to 0.556
- 0.556 to 0.688
- 0.688 to 1.351

Spatial residual FE - S4
- -2.223 to 0.170
- 0.170 to 0.356
- 0.356 to 0.511
- 0.511 to 0.665
- 0.665 to 1.549

20
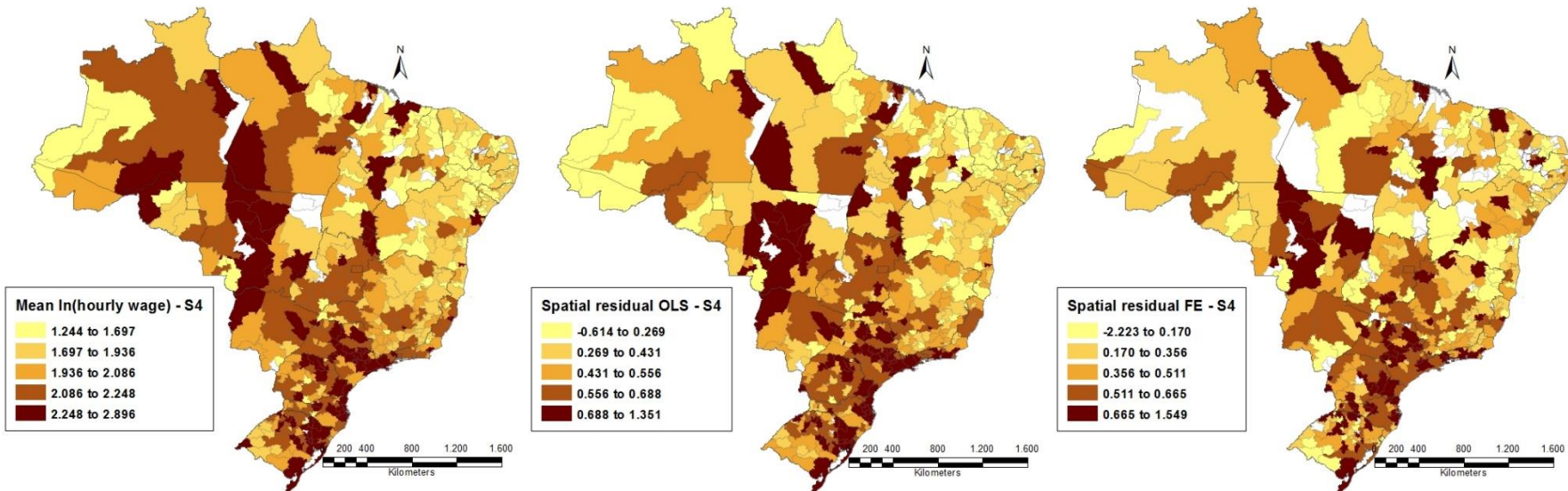
Sector 5 – Services high-knowledge
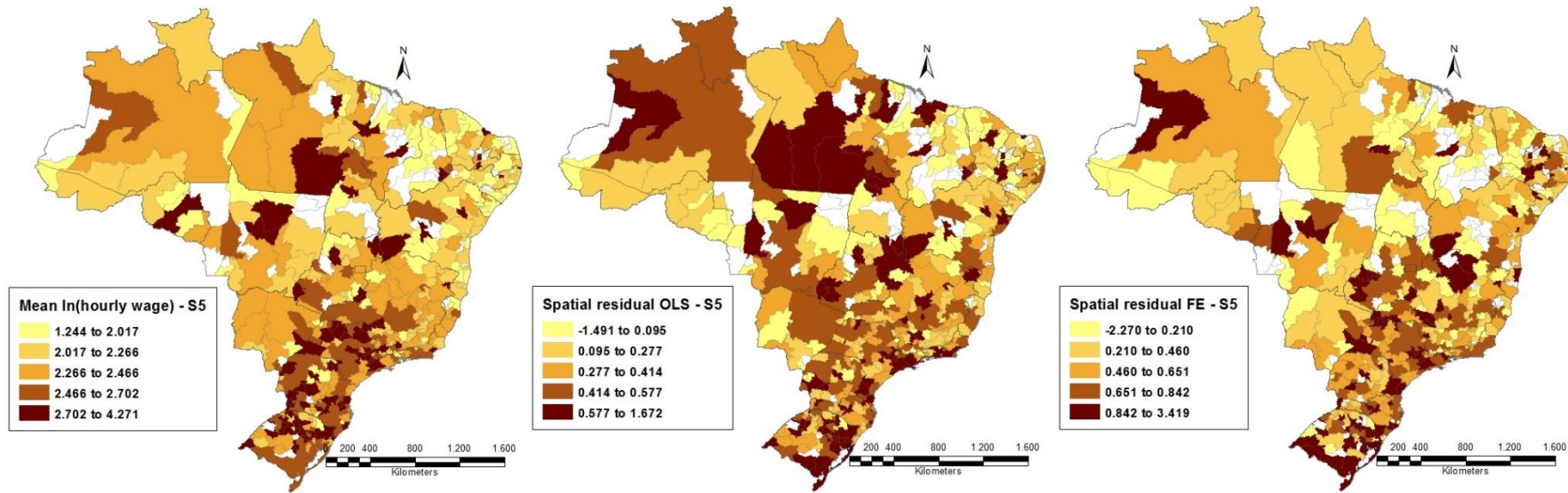
Obs. 1: All maps refer to 2012, and in the case of spatial wages, they are calculated as a simple average of each industry in the area REGIC inside the sector. In the case of the mean logarithm of the hourly wage, it is calculated as a simple average of the individual wages in the sector-area REGIC.
Obs. 2: The classes of each map are based on the division of quintiles of each variable.
Obs. 3: Areas REGIC in white do not present any data in the sample and sector.
Source: Elaboration of the authors.

The set of maps presented above allows different analysis. First, for each sector, the comparison of the first map on the left with the two others[15] shows that when individual characteristics are taken into account, at least part of the regional distribution changes significantly. This is an evidence that controlling for individual characteristics in this matter is essential for the analysis in the regional context. For instance, for Sectors 1 to 4, comparing the left and the center maps, the Northern region of the country presents darker areas for the individual wage than for the spatial wage. For all sectors, the South-Southeast of the country remains relevant in terms of concentrating higher levels of the spatial wage (after controlling for individual characteristics). In relation to Sector 3, there is a significant share of areas without information (44%), due to a high regional concentration.

The inclusion of individual fixed effects in the first stage significantly changes the regional distribution of the spatial wage. As previously mentioned, in this case the spatial wage is calculated only for individuals who moved between industries (inside the technological sector) and/or area REGIC over time (between 2004, 2008 and 2012). All maps are based in 2012 data, being associated to the last place in which workers are observed. The fact that just a selected part of the sample of moving workers is used to build the third column of maps explains why certain counterintuitive patterns emerge (such as darker areas in the countryside of the Northeast, far from the coast) for most sectors.

Even if there is indication of a possible spatial pattern in the data, a few reasons excluded the possibility of estimating a spatial model. Firstly, the dependent variable in the second stage is not available for all areas REGIC. Moreover, by taking into account areas REGIC, a great share of possible neighborhood effects will be lost. This is so because these areas are defined by daily commuting patterns, already measuring most of the relevant relations in the local labor market. If spatial effects are included here, they would possibly be capturing spurious relations, missing in the specified model. These spatial effects could also capture part of the effects of local industrial composition. The main conclusion in this respect is that the spatial scale considered here minimizes the need for spatial dependence analysis, which would be more justified in a smaller scale.

---

[15] The map in the center refers to the spatial wage for the model with pooled OLS estimation in the first stage, while the map in the right is related to the spatial wage obtained from the estimation with individual fixed effects in the first stage.

The dependent variable of the second stage is similar to the spatial wage depicted in the middle and right maps (except that the maps are showing the average of the spatial wage over industries). Following Equation 3, a set of controls in the industry and area levels are included as explanatory factors, in addition to different combinations of measures for urbanization degree, specialization, diversity and competition. These different combinations, named *Compositions*, aim to identify a robust pattern in the results for the analysis of the industrial scope of agglomeration economies even when the explanatory variables are changed.

Table 4. Second-stage Regressions for the Spatial Wage with and without Individual Fixed Effects in the First Stage.

		2nd stage with OLS (1st stage with OLS)					2nd stage with OLS (1st stage with individual Fixed Effects)				
		S1. Manufacturing low-tech	S2. Manufacturing medium-tech	S3. Manufacturing high-tech	S4. Services less-knowledge	S5. Services high-knowledge	S1. Manufacturing low-tech	S2. Manufacturing medium-tech	S3. Manufacturing high-tech	S4. Services less-knowledge	S5. Services high-knowledge
Composition 1		S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
Urbanization	ln(density)	0.0631***	0.0460***	0.0749***	0.0511***	0.0636***	0.0384***	0.191***	0.0318**	0.0711***	0.0600***
Diversity	ln(herfindhal)	0.0665***	0.108***	0.118***	0.0435**	0.0376*	0.0415*	0.233***	0.00680	0.0192	0.127***
Specialization	ln(location quotient)	-0.00430	-0.0126	-0.0216	-0.00410	0.0135	-0.0186*	0.0649**	-0.0293	0.0282*	0.0225
Competition	ln(competition degree)	-0.0223**	-0.0656***	-0.0334*	-0.0478***	0.00284	-0.0388***	-0.122***	-0.0132	-0.0428**	-0.0155
	R^2	0.378	0.390	0.280	0.323	0.442	0.115	0.494	0.062	0.146	0.322
Composition 2		S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
Urbanization	ln(density)	0.0788***	0.0564***	0.0901***	0.0577***	0.0678***	0.0565***	0.179***	0.0465***	0.0774***	0.0667***
Diversity	ln(shannon entropy)	0.345***	0.391***	0.400***	0.230***	0.0650	0.200**	0.630***	0.0254	0.109	0.272***
Specialization	ln(specialization degree)	0.0855***	0.0536***	0.0629***	0.0499***	0.0110	0.0628***	0.00789	0.0428*	0.0361***	0.0212
Competition	ln(competition degree)	-0.0151**	-0.0552***	-0.0134	-0.0463***	-0.00834	-0.0175*	-0.190***	0.0156	-0.0671***	-0.0346*
	R^2	0.393	0.393	0.284	0.327	0.441	0.121	0.490	0.063	0.147	0.320
Composition 3		S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
Urbanization	ln(density)	0.0792***	0.0566***	0.0940***	0.0558***	0.0600***	0.0555***	0.164***	0.0452***	0.0750***	0.0556***
Diversity	ln(share 5 biggest sectors)	-0.222***	-0.285***	-0.165**	-0.206***	-0.190***	-0.150**	-0.749***	-0.0462	-0.124*	-0.405***
Specialization	ln(specialization degree)	0.0675***	0.0428***	0.0512***	0.0430***	0.0167	0.0533***	0.00718	0.0433*	0.0346***	0.0229
Competition	ln(competition degree)	-0.00751	-0.0516***	-0.0130	-0.0440***	-0.00544	-0.0118	-0.178***	0.0165	-0.0653***	-0.0261
	R^2	0.392	0.393	0.281	0.327	0.443	0.121	0.495	0.063	0.147	0.323
Composition 4		S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
Urbanization	ln(density)	0.0786***	0.0581***	0.0926***	0.0587***	0.0654***	0.0554***	0.179***	0.0462***	0.0761***	0.0637***
Diversity	ln(herfindhal)	0.159***	0.158***	0.150***	0.0960***	0.0558**	0.105***	0.283***	0.0170	0.0677**	0.160***
Specialization	ln(specialization degree)	0.0907***	0.0564***	0.0641***	0.0507***	0.0168	0.0684***	0.0178	0.0438*	0.0411***	0.0310*
Competition	ln(competition degree)	-0.0123*	-0.0528***	-0.0118	-0.0454***	-0.00913	-0.0153*	-0.186***	0.0159	-0.0660***	-0.0358*
	R^2	0.398	0.396	0.287	0.328	0.442	0.123	0.492	0.063	0.148	0.322
Observations	N (REGIC * sector * year)	4912	3217	2301	5836	5208	4522	2895	2094	5378	4884

*: α = 0.10; **: α = 0.05; ***: α = 0.01;

Obs.: additional controls are ln(area in squared kilometers), ln(distance to the equator), ln(altitude), year dummies (2004 omitted), sector dummies, macro-region dummies (North omitted), constant.

Source: Elaboration of the authors.

Focusing initially in the five first columns of Table 4, which present the results for the second stage with OLS estimation in the first stage, there is a very robust indication for the urbanization measure. Its coefficients are positive and statistically significant for all sectors, ranging from 0.0511 to 0.0940 in different *Compositions*. The effects are ordered from the highest to the lowest in all cases as S3, S1, S5, S4 and S2. This can be considered as an evidence that high-tech and low-tech manufacturing benefit more from urban scale. Another robust result is found for diversity measures: there is a positive relationship for all sectors between more diverse areas and higher productivity (measured by spatial wages). The highest coefficients are observed for manufacturing sectors, especially S3 (except in *Composition* 4, in which S1 and S2 present higher values).

The specialization measure considered in the first specification (location quotient), is non-significant for all sectors (and negative for most of them). However, all other *Compositions* include the specialization degree for this dimension, with positive coefficients (which are non-significant only for S5). Therefore, the location quotient seems unsuitable for the problem at hand, while the specialization degree is more suitable to capture the importance of this dimension for local productivity. Finally, the estimated coefficients of competition degree are negative in all cases. High significance is found only for S2 and S4, and S1 is slightly significant in most specifications. Therefore, spatial wages in S3 and S5 seem to be unaffected by local competition, while sectors with smaller technological or knowledge intensity are negatively affected.

It is possible to compare the results presented in Table 4 with the theoretical frameworks summarized in Table 1. The MAR approach is associated to (+) specialization, (-) diversity and (-) competition, while Jacobs' is related to (-) specialization, (+) diversity and (+) competition, and Porter's is synthetized by (+) specialization, (-) diversity and (+) competition. Then, there is no clear pattern to indicate the more appropriate theoretical framework for each sector. In fact, for S5, it seems that only diversity is significant (and positive), suggesting at least partially that Jacobs' perspective is more adequate. For S1, S2 and S4, even if diversity is positive and significant, the combination of positive specialization and negative competition coefficients indicate that Marshall could be more adequate to explain their patterns. In the case of S3, there are elements from both Marshall and Jacobs' perspectives. The result to be highlighted here is that the most adequate industrial mix to foster productivity can vary for each sector, being associated to their technological intensity.

Moving for the five last columns of Table 4, they depict the results for the spatial wage regression when individual fixed effects are included in the first stage. Therefore, only movers (between different areas REGIC and/or industries inside a technological sector) measure the spatial wage. Even if many of the coefficients change significantly from the previous analysis, in the case of density they are still positive and significant for all sectors. It is noteworthy that for S2 there is an important increase in the elasticity size (which was around 0.05 and went up to something around 0.16 and 0.19), but for the other sectors these elasticities assume a very similar value to the previous case. In the literature, most studies find that when controlling for sorting (by including individual fixed effects), the estimated coefficient for urbanization decreases significantly (Combes, Duranton and Gobillon, 2008). Following Groot et al. (2014), there is no consensus whether the inclusion of individual fixed effects is the best strategy, because it may create a selection bias by the fact that the spatial wage will be measured only by migrants (who should be more associated with risk-taking, entrepreneurship, adaptation, among other characteristics, which can be positively related to productivity).

Table 5. Second-stage Regressions for the Spatial Wage with and without Individual Fixed Effects on the First Stage, with Instrumental Variables for ln(density).

		2nd stage with IV (1st stage with OLS)					2nd stage with IV (1st stage with individual Fixed Effects)				
		S1. Manufacturing low-tech	S2. Manufacturing medium-tech	S3. Manufacturing high-tech	S4. Services less-knowledge	S5. Services high-knowledge	S1. Manufacturing low-tech	S2. Manufacturing medium-tech	S3. Manufacturing high-tech	S4. Services less-knowledge	S5. Services high-knowledge
Composition 1		S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
Urbanization	ln(density)	0.0632***	0.0380***	0.0693***	0.0423***	0.0698***	0.0377***	0.184***	0.0582***	0.0568***	0.0561***
Diversity	ln(herfindhal)	0.0662***	0.122***	0.126***	0.0567***	0.0284	0.0425	0.245***	-0.0319	0.0406	0.133***
Specialization	ln(location quotient)	-0.00423	-0.0163	-0.0235*	-0.00537	0.0156	-0.0189*	0.0615**	-0.0207	0.0260*	0.0211
Competition	ln(competition degree)	-0.0222**	-0.0698***	-0.0353*	-0.0540***	0.00851	-0.0392***	-0.125***	-0.00487	-0.0530***	-0.0191
R²		0.378	0.389	0.280	0.322	0.441	0.115	0.494	0.059	0.145	0.322
Composition 2		S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
Urbanization	ln(density)	0.0971***	0.0574***	0.0900***	0.0597***	0.0800***	0.0681***	0.173***	0.0856***	0.0713***	0.0724***
Diversity	ln(shannon entropy)	0.272***	0.386***	0.400***	0.224***	0.0263	0.152*	0.655***	-0.131	0.130	0.254**
Specialization	ln(specialization degree)	0.0906***	0.0539***	0.0629***	0.0504***	0.0149	0.0660***	0.00629	0.0604**	0.0344**	0.0230
Competition	ln(competition degree)	-0.0135**	-0.0552***	-0.0134	-0.0453***	-0.00136	-0.0166*	-0.190***	0.0140	-0.0703***	-0.0313*
R²		0.391	0.393	0.284	0.327	0.441	0.120	0.490	0.058	0.147	0.320
Composition 3		S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
Urbanization	ln(density)	0.0942***	0.0511***	0.0932***	0.0488***	0.0694***	0.0650***	0.147***	0.0847***	0.0620***	0.0524***
Diversity	ln(share 5 biggest sectors)	-0.179***	-0.303***	-0.167**	-0.228***	-0.162**	-0.122*	-0.798***	0.0355	-0.165**	-0.414***
Specialization	ln(specialization degree)	0.0745***	0.0405***	0.0508***	0.0407***	0.0203*	0.0579***	0.000175	0.0655***	0.0301**	0.0217
Competition	ln(competition degree)	-0.00778	-0.0516***	-0.0130	-0.0472***	-0.000994	-0.0121	-0.178***	0.0143	-0.0713***	-0.0277
R²		0.391	0.393	0.281	0.327	0.442	0.121	0.495	0.058	0.147	0.323
Composition 4		S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
Urbanization	ln(density)	0.0956***	0.0592***	0.0945***	0.0590***	0.0776***	0.0663***	0.175***	0.0838***	0.0697***	0.0693***
Diversity	ln(herfindhal)	0.140***	0.157***	0.148***	0.0956***	0.0437	0.0927***	0.288***	-0.0112	0.0741**	0.154***
Specialization	ln(specialization degree)	0.0975***	0.0569***	0.0651***	0.0509***	0.0216*	0.0728***	0.0158	0.0654***	0.0387***	0.0332*
Competition	ln(competition degree)	-0.0108*	-0.0528***	-0.0118	-0.0452***	-0.00213	-0.0145*	-0.186***	0.0148	-0.0695***	-0.0325*
R²		0.397	0.396	0.287	0.328	0.441	0.122	0.492	0.058	0.147	0.322
Observations	N (REGIC * sector * year)	4912	3217	2301	5836	5208	4522	2895	2094	5378	4884

*: α = 0.10; **: α = 0.05; ***: α = 0.01;

Obs. 1: additional controls are ln(area in squared kilometers), ln(distance to the equator), ln(altitude), year dummies (2004 omitted), sector dummies, macro-region dummies (North omitted), constant.

Obs. 2: instruments for ln(density) are ln(density in 1940), ln(distance to the coast), dummy of sugarcane production in colonial times, dummy of gold exploring in colonial times, dummy of coffee production in colonial and imperial times.

<u>Source</u>: Elaboration of the authors.

In relation to the other variables capturing industrial composition, signs are similar, but there are important changes in significance. Diversity measures are still positive, but non-significant for S3 and less significant for S4 in all specifications. The specialization degree becomes non-significant or slightly significant for S2, S3 and S5 in the different specifications. In the case of competition degree, elasticities are still negative but lose significance for S1, S3 and S5. Then, S4 gets closer to Marshall's explanation, while S1 and S5 spatial wages will be positively associated to diversity (Jacobs perspective). Finally, the explanatory power of each set of models varies significantly. When the first stage is based on an OLS, second stage variables are more relevant to explain the spatial wage, with a R^2 ranging from 0.28 to 0.44. On the other hand, the last five columns show a more heterogeneous pattern for this measure of fit: while the R^2 of S2 ranges around 0.49, the one for S3 is much lower (around 0.06). The fit of the model is also relatively high for S5, an indication that for the remaining sectors (S1, S3 and S4), after controlling for individual heterogeneity, spatial wage variation is less associated to the local industrial composition.

The second issue explored here involves a potential endogeneity of the urbanization measure (density). As discussed in Section 3, the inclusion of individual fixed effects aims to control for a sorting process, meaning that more productive individuals will migrate to bigger cities where they can find higher salaries, reinforcing city size and local productivity. However, it is also important to tackle the potential endogeneity of the quantity of labor (Combes et al., 2011). Table 5 presents a similar set of results as Table 4, but with instrumental variables estimations. Once again, *Composition* 1 seems less adequate, especially for the regressions with OLS in the first stage. The main conclusions seem to hold, especially for density coefficients, which are still positive and significant. Focusing on the first five columns once again (OLS in the first stage), diversity seems to be positively associated to the spatial wage, even if for S5 in some specifications this coefficient is non-significant.

There is no clear order of effects among sectors, except for a highlight to the fact that S3 presents the highest elasticities. For competition, the coefficient of S5 is non-significant and the one for S1 becomes less significant. These results also indicate that specialization is relatively more important for S1 and then S3, while the elasticities of competition show a stronger negative effect for S2 and S4. Therefore, the inclusion of instrumental variables does not change in a relevant way the main conclusions draw from Table 4, what is an evidence of robust results.

Another robustness check involves comparing these models for different time-periods. Instead of analyzing all years simultaneously (2004-2008-2012), it is possible to break them down into two groups: individuals who are observed in 2004-2008 and individuals present in 2008-2012. It is important to notice that the sample will not necessarily be compose by the same individuals as in the previous case, because now the requirement is that the individual is observed in the database only in two instead of three years.

Brazil has seen a significant formalization process over the last decade, leading to an increase in the original database of this study (see Section 4). I t is not possible to analyze the whole labor market at once due to the lack of identified information at the individual level for the informal sector. However, it is necessary to investigate whether and how this formalization may have changed productivity determinants over the country. The aforementioned comparison of two periods gives elements to address this issue at least partially, and at least indirectly, controls for the regional heterogeneity in the formalization process identified in Section 4. Table A.7 in Annex 7 provides the main results for the models with IV in the second and OLS in the first stage. This set of results can be compared to the first five columns of Table 5.

In fact, significance does not seem to be affected in a relevant way by the split of the database, meaning that the relationships explored here are present in the whole period. The main change refers to the size of the main elasticities, which decrease from 2004-2008 to 2008-2012. This is especially true in the case of the urbanization measure, with the highest decrements happening for S2, S4 and S1. A possible explanation for this result is the fact that this formalization process was stronger in more remote areas, places in which initial informality was higher. It is expected that the urban size will be smaller for these new areas with incoming workers, reducing density elasticities in relation to the spatial wage. For the other variables there is not such a clear pattern, with increasing or decreasing coefficients depending on the sector. However, the fact that signs, significance and relative size do not change significantly compose another indication of robustness of the results.

In summary, the main results found here seem to be reasonably robust to different specifications and time-spans. The most suitable industrial mix for each sector can vary, meaning that not necessarily only one theoretical framework can explain the economy as a whole. There seems to be a heterogeneous effect for different sectors, a fact that must be taken into account when proposing public policies aiming to foster productivity. Moreover,

private sector agents should balance the most relevant factors for their industry when choosing their location.

6. Conclusions

This study aimed to contribute to the literature on static agglomeration effects, controlling for individual skills and comparing different estimation strategies. Focusing on the industrial scope of agglomeration economies in Brazil, we provide a large set of results covering different concerns usually present in this literature.

The analysis was based on identified registration data covering all formal firms and their employees in Brazil in three distinct years (2004, 2008 and 2012). After all proper cleaning procedures, a sample was selected, with information in almost the whole country. Then, the two-stage estimation procedure allowed the following variations: a first stage for the logarithm of the hourly wage in the individual level with and without individual fixed effects; and a second stage for the spatial wage (obtained as the estimated interaction dummies area-sector-year in the first stage) with and without instrumental variables for employment density. In addition, the analysis was divided in two time-spans, 2004-2008 and 2008-2012 in order to assess whether the formalization process seen in Brazil in this period could have affected the results.

With a separate analysis for each of the five broad sectors (S1 – Manufacturing low-tech, S2 – Manufacturing medium-tech, S3 – Manufacturing high-tech, S4 – Services less-knowledge, and S5 – Services high-knowledge), the main conclusion is that the most adequate industrial mix to foster productivity can vary for each sector, being associated to their technological intensity. An unambiguous and universally valid theoretical framework can apparently not explain the results found for the different sectors.

Synthetizing the theoretical alternatives, the MAR approach is associated to (+) specialization, (-) diversity and (-) competition, Jacobs' approach is related to (-) specialization, (+) diversity and (+) competition, and Porter's approach is synthetized by (+) specialization, (-) diversity and (+) competition. In the simplest estimation (OLS in the first and second stages), for S5 it seems that only diversity is significant (and positive), suggesting that Jacobs' perspective is more adequate. For S1, S2 and S4, even if diversity is positive and significant, the combination of positive specialization and negative competition coefficients

indicates that the MAR framework could be more adequate to explain these patterns. In the case of S3, there are elements from both Marshall's and Jacobs' perspectives. These results seem to be robust to different specifications and estimation strategies.

Another important aspect to be highlighted is that when splitting the database into two time-spans, even if the main conclusions remain unchanged, the size of the elasticities for urbanization economies is much smaller in the second period, for all sectors. This can be an evidence of the fact that the recent formalization process was more concentrated in remote areas of the country (with higher initial informality). Then, the urban size in these new areas is expected to be smaller, reducing density elasticities in relation to the spatial wage for the newcomers in the formal labor market.

Finally, the urbanization economies coefficient (the logarithm of employment density) is positive and significant for all sectors, ranging from 0.0511 to 0.0940 in different specifications, under the simplest estimation (OLS in the first and the second stages). Ordering the effects between the sectors from the highest to the lowest, we find the following sequence: S3, S1, S5, S4 and S2. This can be considered as an evidence that high-tech and low-tech manufacturing sectors benefit more from urban scale in Brazil, followed by services associated with higher knowledge.

Consequently, different city sizes can be more appropriate for distinct industries. The local industry mix can also vary for each type of activity, meaning that sometimes diversity is more relevant, while in another occasion specialization can be more important, as well as competition. This implies that both public policy and private sector actions must consider this heterogeneity when searching for the most appropriate incentives and locational strategies to increase productivity in a certain region or specific production unit.

Annex 1 – Aggregation of Manufacturing and Service Industries

Two-digit industries were aggregated in three manufacturing and two service sectors, based on technological and knowledge intensity.[16]

Table A.1. Aggregations of Manufacturing and Service Industries According to Technological Knowledge Intensity

Aggregation	Industry code (CNAE 2 digits)	Industry name
Low technology manufacturing	15	Manufacture of food products and beverages
	16	Manufacture of tobacco products
	17	Manufacture of textiles
	18	Manufacture of wearing apparel
	19	Manufacture of leather and related products
	20	Manufacture of wood products
	22	Printing and reproduction of recorded media
	36	Manufacture of furniture, other manufacturing
Medium technology manufacturing	21	Manufacture of paper, cellulose and paper products
	23	Manufacture of coke, refined petroleum products, nuclear fuel and ethanol
	25	Manufacture of rubber and plastic products
	26	Manufacture of non-metallic mineral products
	27	Manufacture of basic metals
	28	Manufacture of fabricated metal products, except machinery and equipment
Medium-high and high technology manufacturing	24	Manufacture of chemical products
	29	Manufacture of machinery and equipment
	30	Manufacture of office machinery and computer equipment
	31	Manufacture of electrical equipment and materials, and machinery
	32	Manufacture of electronic material and communication equipment and products
	33	Manufacture of medical, precision and optical instrum., industrial automation equip. and chronometers
	34	Manufacture and assembly of motor vehicles, trailers and semi-trailers
	35	Manufacture of other transport equipment
Less knowledge-intensive services	55	Accommodation and food service activities
	60	Land transport
	63	Support activities for transportation and travel agencies
	70	Real estate activities
	71	Rental and leasing activities
	91	Activities of membership organisations
	92	Cultural, recreational and sport activities
	93	Social services
	95	Domestic services
	99	Activities of extraterritorial organisations and bodies
Knowledge-intensive services	61	Water transport
	62	Air transport
	64	Postal and courier activities and telecommunication
	65	Financial services
	66	Insurance and pension plans
	67	Auxiliary activities to financial services, insurance and pension plans
	72	Computer programming, consultancy and related activities
	73	Research and development
	74	Services mainly to companies
	80	Education
	85	Human health and social services
	90	Urban cleaning, sewage and related activities

Source: Elaboration of the authors.

---

[16] This classification is inspired in the one proposed by Eurostat: http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:High-tech (last access in 15/01/2915).

Annex 2 – Definition of variables and data sources

Table A.2. Definition and Description of the Main Variables Considered in the Model

Variables	Definition	Level	Data source
Hourly wage	Monthly wage received in December divided by 4.3 times the number of weely hours in the contract.	Individual	RAIS microdata
Age	Age at the end of the year.	Individual	RAIS microdata
Education attainment	Illiterate, incomplete primary school, complete primary school to incomplete high school, complete high school to incomplete college, college degree or more.	Individual	RAIS microdata
Skills - factors	See Annex 4.	Individual	RAIS microdata and Maciente (2012)
Firm size	Size of the firm in which the individual is working: up to 4 employees, 5 to 9, 10 to 19, 20 to 49, 50 to 99, 100 to 249, 250 to 499, 500 to 999, 1,000 employees or more.	Individual	RAIS microdata
Labor density in the formal sector	Total employment divided by the area (in km2).	REGIC	RAIS - aggregated data
Area	Area in km2.	REGIC	IPEADATA
Specialization	See Annex 3.		RAIS - aggregated
Diversity	See Annex 3.		RAIS - aggregated
Competition	See Annex 3.		RAIS - aggregated
Altitude	Average altitude of the REGIC obtained from the weighted mean of the altitude of each municipality (weight = area).	REGIC	Adapted from Naritomi et al. (2012)
Distance to the Equator line	Distance to the Equator measured as the absolute value of the latitude coordinate - obtained as a weighted average of the distance to the Equator line of the centroids of all municipalities that compose the REGIC (weight = area).	REGIC	Adapted from Naritomi et al. (2012)

Instruments	Definition	Level	Data source
Population density in 1940	Population in 1940 for MCAs 1940-2000 is redistributed for MCAs 2000-2010 based on the populational share of the latter on the former in 2000, aggregated by REGIC and divided by the area in km2.	REGIC	IPEADATA
Distance to the coast	Distance to the coast (in 100 km) - obtained as a weighted average of the distance to the sea of the centroids of all municipalities that compose the REGIC (weight = area).	REGIC	Adapted from Naritomi et al. (2012)
Sugar	Proximity to the sugarcane boom (until 1760), calculated as the weighted average of the municipal index for all municipalities pertaining to the REGIC, ranging from 0 (more than 200 km) to 1 according to the proximity to the nearest municipality in sugarcane areas (Naritomi et al., 2012) - weight = area.	REGIC	Adapted from Naritomi et al. (2012)
Gold	Proximity to the gold boom, calculated as the weighted average of the municipal index for all municipalities pertaining to the REGIC, ranging from 0 (more than 200 km) to 1 according to the proximity to the nearest municipality in gold exploration areas (Naritomi et al., 2012) - weight =	REGIC	Adapted from Naritomi et al. (2012)
Coffee	Proximity to the coffee boom until 1886, calculated as the weighted average of the municipal index for all municipalities pertaining to the REGIC, ranging from 0 (more than 200 km) to 1 according to the proximity to the nearest municipality in coffee areas (Naritomi et al., 2012)	REGIC	Adapted from Naritomi et al. (2012)

Source: Elaboration of the authors.

Annex 3 – Indicators of the Industrial Scope of Agglomeration Economies

Table A.3. Indicators and Measures of the Industrial Scope of Agglomeration Economies

Indicator	Measure	Formula	Interpretation	Source
Specialization	Location quotient	$LQ_{industry,region} = \dfrac{E_{industry,region}/E_{region}}{E_{industry,country}/E_{country}}$	LQ > 1: region is more specialized in industry i.	Glaeser et al. (1992)
	Specialization degree	$SD_{region} = \sum_{ind}\left(\dfrac{E_{industry,region}}{E_{region}} - \dfrac{E_{industry}}{E_{country}}\right)^2$	SD close to 0: industrial composition in the region is similar to the national one; SD close to 2: region is completely specialized.	Henderson (2003)
Competition	Competition degree	$C_{industry,region} = \dfrac{F_{industry,region}/E_{industry,region}}{F_{industry,country}/E_{industry,country}}$	C > 1: higher competition in ind in the region than in the country.	Glaeser et al. (1992)
Diversity	Inverse Herfindhal index	$IHI_{region} = \left(\dfrac{E^2_{region}}{\sum_{ind}E^2_{industry,region}}\right)$	Higher IHI means more diversity.	Combes et al. (2008, 2011)
	Non-diversity	% of the largest five industries other than the industry in question on city's employment.	The higher the %, the lower the diversity	Glaeser et al. (1992)
	Shannon entropy	$S_{region} = -\sum_{industry}\left(\dfrac{E_{industry,region}}{E_{region}}\ln\dfrac{E_{industry,region}}{E_{region}}\right)$	High value means high diversification (accounts for the size distribution of sectors)	Groot et al. (2014)
Urbanization	Density	$Den_{region} = \ln\left(\dfrac{E_{region}}{A_{region}}\right)$	Higher density is related to greater urbanisation.	Groot et al. (2014), Combes et al. (2008, 2011)

Obs.: Region is area REGIC in the empirical analysis.
Source: Elaboration of the authors.

Maciente (2012) has developed a matching among the occupational classification in Brazil and the profile of skills based on ONET (Occupational Information Network). Based on that, the author finds a weighting system for each ability required on the job, based on the worker qualification and the job complexity. Then, with a factor analysis it is possible to calculate the skill requirement of each occupation, with the classification presented in Table A.3.

Table A.4. Description of the Factors Defined by Maciente (2012) Aggregating Skill Requirements of each Occupation in Brazil

Factor	Description
1	Cognitive skills
2	Maintenance and operation skills
3	Assistance skills
4	Management skills
5	Design and engineering skills
6	Transportation skills
7	Artistic skills
8	Accuracy and automation skills
9	Supervised work skills
10	Teaching and social science skills
11	Physical strength
12	Telecommunication skills
13	Independence skills
14	Natural science skills
15	Attention skills
16	On-the-job experience
17	Conflict management skills
18	Team-work skills
19	Sales skills
20	Monitoring and compliance skills
21	Clerical skills

Source: Maciente (2012).

Annex 5 – Comparison of the Full Sample and the Individuals Who Provide Variation for the Spatial Wage Calculation When Individual Fixed Effects are Included

Table A.5. Comparison of Selected Descriptive Statistics of the Full Sample and the Individuals with Panel Variation, 2012 (for the database 2004-2008-2012)

	S1 Manufacturing low-tech		S2 Manufacturing medium-tech		S3 Manufacturing high-tech		S4 Services less-knowledge		S5 Services high-knowledge	
	full sample	panel	full sample	panel	full sample	panel	full sample	panel	full sample	panel
ln(real hourly wage)	2.36	2.34	2.67	2.65	3.08	3.01	2.24	2.20	2.60	2.61
age	40.26	37.98	41.03	39.12	40.86	38.50	43.36	41.43	41.38	39.55
Education attainment										
Illiterate	0.5%	0.3%	0.3%	0.1%	0.1%	0.0%	0.4%	0.2%	0.2%	0.1%
Incomplete primary school	27.8%	20.8%	21.1%	13.4%	10.9%	5.4%	31.3%	19.9%	11.7%	8.1%
Comp. primary school - incompl. middle school	26.0%	23.7%	23.7%	20.6%	18.4%	12.3%	30.7%	28.9%	18.1%	14.5%
Complete middle school - incomplete college	39.2%	45.1%	46.2%	53.8%	50.2%	56.3%	33.0%	45.0%	41.3%	44.0%
College degree or more	6.6%	10.1%	8.8%	12.1%	20.5%	26.0%	4.7%	6.0%	28.8%	33.3%
Firm size										
Up to 4 employees	2.7%	3.1%	2.0%	2.5%	0.8%	1.2%	11.2%	8.6%	2.5%	2.5%
5 to 9	4.0%	4.9%	3.7%	4.2%	1.4%	1.7%	15.8%	10.8%	3.8%	4.2%
10 to 19	6.6%	7.4%	7.6%	7.8%	3.6%	4.2%	10.9%	10.9%	5.7%	6.1%
20 to 49	10.6%	12.0%	12.4%	13.2%	6.8%	8.7%	14.8%	16.7%	9.6%	9.8%
50 to 99	9.4%	10.5%	12.1%	13.9%	8.2%	10.2%	9.6%	12.0%	7.5%	7.9%
100 to 249	14.8%	16.1%	15.7%	19.1%	14.9%	16.2%	11.1%	14.3%	11.3%	12.4%
250 to 499	14.2%	15.2%	15.3%	15.6%	14.9%	15.2%	8.8%	10.4%	11.8%	12.4%
500 to 999	12.7%	13.9%	13.8%	11.7%	13.6%	14.9%	8.4%	8.1%	12.8%	12.7%
1,000 or more employees	25.0%	17.0%	17.3%	12.0%	35.8%	27.6%	9.6%	8.3%	35.1%	32.1%
Macro-region										
North	1.7%	2.5%	1.3%	1.8%	1.6%	2.4%	2.1%	2.7%	2.6%	2.4%
Northeast	14.7%	13.7%	7.0%	7.5%	2.8%	2.8%	13.9%	13.9%	13.2%	11.1%
Southeast	49.6%	41.0%	70.2%	68.7%	76.0%	74.4%	66.8%	62.5%	67.7%	68.5%
South	29.6%	36.1%	20.0%	19.9%	18.9%	19.4%	12.8%	15.2%	11.1%	13.0%
Center-West	4.4%	6.7%	1.5%	2.1%	0.7%	1.0%	4.4%	5.7%	5.4%	5.0%
Number of observations	45,902	6,722	34,487	4,645	37,840	6,179	72,014	13,890	58,350	18,032

Source: Elaboration of the authors.

Annex 6 – Results for the First-Stage Regressions with and without Individual Fixed Effects

Table A.6

	First stage with OLS					First stage with Fixed Effects				
	S1. Manufacturing low-tech	S2. Manufacturing medium-tech	S3. Manufacturing high-tech	S4. Services less-knowledge	S5. Services high-knowledge	S1. Manufacturing low-tech	S2. Manufacturing medium-tech	S3. Manufacturing high-tech	S4. Services less-knowledge	S5. Services high-knowledge
age	0.0497***	0.058***	0.0704***	0.035***	0.0534***	0.134	0.2843	0.1249	0.1063	0.1224
age2	-0.0005***	-0.0006***	-0.0007***	-0.0004***	-0.0005***	-0.0008***	-0.0009***	-0.0011***	-0.0006***	-0.0009***
illiterate (omitted)										
primary school - incomplete	0.111***	0.1241***	0.0837*	0.0081	-0.0209					
secondary school incomplete	0.1865***	0.2104***	0.1786***	0.0124	0.021					
tertiary education - incomplete	0.2959***	0.3505***	0.2968***	0.0862***	0.15***					
graduate + pos-graduate	0.8391***	0.8103***	0.706***	0.6789***	0.6981***					
Factor 1 - Cognitive skills	0.2168***	0.2496***	0.2737***	0.1742***	0.3007***	0.0992***	0.0946***	0.0957***	0.0829***	0.1639***
Factor 2 - Maintenance and operation skills	0.0231***	0.0086***	-0.0058**	-0.0012	0.048***	-0.0041	0.0012	0.0016	-0.0016	-0.008***
Factor 3 - Assistance skills	-0.0249***	0.0048	-0.0211***	0.0031	0.0053*	-0.006	-0.0017	-0.0147***	0.0216***	0.01***
Factor 4 - Management skills	0.0827***	0.0255***	0.0424***	0.0181***	0.0319***	0.0274***	0.003	0.0185***	-0.0028	0.0083***
Factor 5 - Design and engineering skills	0.0118***	0.0041	0.0008	-0.0088***	-0.0381***	-0.0005	-0.016***	-0.0089***	-0.0061*	-0.0152***
Factor 6 - Transportation skills	-0.0174***	-0.0387***	-0.0135***	0.0626***	0.0173***	-0.0088***	-0.0155***	-0.0046	0.0387***	0.0083***
Factor 7 - Artistic skills	0.024***	-0.0068*	0.0127***	0.0459***	0.0112***	-0.0018	-0.0006	-0.0076**	0.0164***	0.0017
Factor 8 - Accuracy and automation skills	0.0204***	0.0122***	0.03***	0.01***	0.0329***	-0.0025	0.0042	0.0007	-0.0049***	0.0156***
Factor 9 - Supervised work skills	0.0383***	0.0447***	-0.001	0.0814***	0.039***	0.0206***	0.0187***	0.0019	0.0464***	0.0227***
Factor 10 - Teaching and social science skills	0.0299***	-0.024***	0.0205***	0.0264***	-0.0018	-0.0034	-0.0221***	-0.0022	0.0086**	-0.0006
Factor 11 - Physical strength	-0.0027	-0.026***	-0.0426***	-0.0385***	-0.0748***	0.0113***	-0.0067**	-0.0142***	-0.0249***	-0.0327***
Factor 12 - Telecommunication skills	0.0475***	0.0207***	0.0042*	0.0584***	0.0255***	0.0169***	0.0053	-0.0024	0.0138***	0.015***
Factor 13 - Independence skills	0.0049**	-0.0008	-0.011***	-0.0187***	0.0056***	-0.0051*	-0.0036	-0.0128***	0.0028*	0.0055***
Factor 14 - Natural science skills	-0.0344***	-0.0527***	-0.032***	-0.0356***	-0.0259***	-0.0231***	-0.0205***	-0.007**	-0.0145***	-0.0061**
Factor 15 - Attention skills	-0.0241***	-0.0275***	-0.0121***	0.0514***	0.0092***	-0.0088***	-0.0076***	-0.0116***	0.0167***	0.0081***
Factor 16 - On-the-job experience	-0.0295***	0.0082***	0.017***	0.031***	-0.0032	-0.0013	0.0062*	0.0047	0.015***	-0.0014
Factor 17 - Conflict management skills	0.0493***	0.0447***	0.0257***	-0.0734***	-0.003	0.0287***	0.0244***	0.0159***	-0.0249***	0.0007
Factor 18 - Team-work skills	-0.0285***	-0.0002	0.0456***	-0.0276***	-0.0279***	-0.0179***	-0.0055	0.0146***	-0.0181***	-0.0068**
Factor 19 - Sales skills	0.0125***	0.0128***	0.0189***	-0.0133***	-0.0056***	0.003	0.0043	-0.0053	0.0011	-0.0044**
Factor 20 - Monitoring and compliance skills	-0.0334***	-0.0195***	-0.0312***	-0.0473***	-0.0543***	-0.0067***	-0.0032	-0.0154***	-0.0171***	-0.018***
Factor 21 - Clerical skills	0.0353***	-0.0164***	0.0207***	0.0083***	0.0455***	0.0011	-0.0123***	0.0071**	-0.005**	0.0088***
firm size (up to 4 employees) - omitted										
firm size 5 to 9	0.0808***	0.0944***	0.1537***	0.0715***	0.1336***	0.0416***	0.0396***	0.0601***	0.0226***	0.1126***
firm size 10 to 19	0.1679***	0.1672***	0.2214***	0.0834***	0.2306***	0.0928***	0.091***	0.1171***	0.0457***	0.1704***
firm size 20 to 49	0.2505***	0.2172***	0.3192***	0.1572***	0.3335***	0.1471***	0.1374***	0.1776***	0.0873***	0.2351***
firm size 50 to 99	0.339***	0.3362***	0.405***	0.2415***	0.4089***	0.1938***	0.1988***	0.2389***	0.1308***	0.2724***
firm size 100 to 249	0.447***	0.4933***	0.5409***	0.3108***	0.48***	0.2403***	0.2772***	0.3192***	0.1816***	0.3146***
firm size 250 to 499	0.5272***	0.6069***	0.6128***	0.3593***	0.4877***	0.2858***	0.3429***	0.3652***	0.2221***	0.3361***
firm size 500 to 999	0.5282***	0.6885***	0.695***	0.346***	0.4889***	0.3113***	0.3808***	0.426***	0.244***	0.3405***
firm size 1000 or more employees	0.587***	0.8077***	0.8325***	0.3989***	0.4366***	0.3613***	0.4729***	0.488***	0.2349***	0.3185***
Observations (individuals * years)	137,706	103,461	113,520	216,042	175,050	137,706	103,461	113,520	216,042	175,050

*: α = 0.10; **: α = 0.05; ***: α = 0.01;

Obs.: Additional controls are dummies for the iteration of area REGIC, year and industry, and a constant term.

<u>Source</u>: Elaboration of the authors.

Annex 7 – Comparison between 2004-2008 and 2008-2012

Table A.7. Second-Stage Regression Results for the Spatial Wage, with Instrumental Variables for ln(density) and OLS Estimation in the First Stage, 2004-2008 and 2008-2012

		2nd stage with IV (1st stage with OLS) - 2004-2008					2nd stage with IV (1st stage with OLS) - 2008-2012				
		S1. Manufacturing low-tech	S2. Manufacturing medium-tech	S3. Manufacturing high-tech	S4. Services less-knowledge	S5. Services high-knowledge	S1. Manufacturing low-tech	S2. Manufacturing medium-tech	S3. Manufacturing high-tech	S4. Services less-knowledge	S5. Services high-knowledge
Composition 1		S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
Urbanization	ln(density)	0.0445***	0.0586***	0.0685***	0.0403***	0.0678***	0.0314***	0.0136	0.0550***	0.0198*	0.0687***
Diversity	ln(herfindhal)	0.0888***	0.0832***	0.105**	0.0613***	0.0169	0.0881***	0.126***	0.152***	0.108***	0.0104
Specialization	ln(location quotient)	-0.0147*	0.00724	-0.0168	-0.000659	-0.0118	-0.0126*	-0.00663	0.00640	-0.00402	0.00930
Competition	ln(competition degree)	-0.0310***	-0.00330	-0.0457**	-0.0488***	-0.0321*	-0.0283***	-0.0445***	-0.0103	-0.0497***	0.0200
	R^2	0.369	0.360	0.243	0.322	0.500	0.401	0.344	0.244	0.321	0.456
Composition 2		S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
Urbanization	ln(density)	0.0788***	0.0814***	0.0906***	0.0602***	0.0808***	0.0617***	0.0370***	0.0835***	0.0410***	0.0806***
Diversity	ln(shannon entropy)	0.305***	0.302***	0.337**	0.228***	-0.00537	0.281***	0.248***	0.572***	0.291***	-0.0224
Specialization	ln(specialization degree)	0.0853***	0.0797***	0.0647***	0.0594***	0.0205	0.0688***	0.0391***	0.0965***	0.0479***	0.0147
Competition	ln(competition degree)	-0.0107*	-0.0119	-0.0301***	-0.0437***	-0.0208	-0.0119**	-0.0392***	-0.0161*	-0.0416***	0.0152
	R^2	0.383	0.369	0.249	0.330	0.500	0.413	0.343	0.257	0.327	0.454
Composition 3		S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
Urbanization	ln(density)	0.0736***	0.0680***	0.0891***	0.0423***	0.0701***	0.0501***	0.0203*	0.0824***	0.0226*	0.0665***
Diversity	ln(share 5 biggest sectors)	-0.227***	-0.340***	-0.184*	-0.306***	-0.140*	-0.278***	-0.358***	-0.279***	-0.329***	-0.166**
Specialization	ln(specialization degree)	0.0654***	0.0723***	0.0521***	0.0513***	0.0280**	0.0519***	0.0336***	0.0757***	0.0336***	0.0245**
Competition	ln(competition degree)	-0.00452	-0.00736	-0.0295***	-0.0456***	-0.0214	-0.00484	-0.0341***	-0.0159*	-0.0449***	0.0138
	R^2	0.383	0.376	0.248	0.331	0.501	0.415	0.349	0.252	0.324	0.457
Composition 4		S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
Urbanization	ln(density)	0.0777***	0.0813***	0.0935***	0.0588***	0.0775***	0.0580***	0.0312***	0.0857***	0.0371***	0.0770***
Diversity	ln(herfindhal)	0.158***	0.157***	0.141***	0.112***	0.0386	0.142***	0.164***	0.212***	0.150***	0.0259
Specialization	ln(specialization degree)	0.0933***	0.0882***	0.0683***	0.0624***	0.0296**	0.0740***	0.0518***	0.0956***	0.0535***	0.0222*
Competition	ln(competition degree)	-0.00774	-0.00929	-0.0279***	-0.0429***	-0.0220*	-0.0103*	-0.0358***	-0.0139*	-0.0409***	0.0141
	R^2	0.391	0.376	0.253	0.333	0.500	0.418	0.353	0.262	0.331	0.455
Obervations	N (REGIC * sector * year)	3930	2620	1878	4673	4308	4195	2834	2234	5184	4910

*: α = 0.10; **: α = 0.05; ***: α = 0.01;

Obs. 1: additional controls are ln(area in squared kilometers), ln(distance to the equator), ln(altitude), year dummies (2004 omitted), sector dummies, macro-region dummies (North omitted), constant.

Obs. 2: instruments for ln(density) are ln(density in 1940), ln(distance to the coast), dummy of sugarcane production in colonial times, dummy of gold exploring in colonial times, dummy of coffee production in colonial and imperial times.

Source: Elaboration of the authors.

References

Amaral, P.V.; Lemos, M.; Simões, R.; Chein, F. (2010) Regional imbalances and market potential in Brazil. *Spatial Economic Analysis*, Vol. 5, Issue 4, pp. 463-482.

Bell, B.; Nickell, S.; Quintini, G. (2002) Wage equations, wage curves and all that. *Labour Economics*, Vol. 9, pp. 341-360.

Beaudry, C.; Schiffauerova, A. (2009), "Who's Right, Marshall or Jacobs? The Localization versus Urbanization Debate," *Research Policy*, Vol. 38, 318–337.

Boschma, R.A. (2005), Proximity and innovation: a critical assessment. *Regional Studies*, Vol. 39, No. 1, pp. 61-74.

Capello, R. (2014), Proximity and regional innovation processes: is there space for new reflections? In: Torre, A.; Wallet, F. (Eds.) *Regional Development and Proximity Relations*. Cheltenham, UK: Edward Elgar, pp. 163-194.

Card, D. (1995) The wage curve: a review. *Journal of Economic Literature*, Vol. 33, No. 2, pp. 785-799.

Ciccone, A.; Hall, R.E. (1996) Productivity and the density of economic activity. *The American Economic Review*, Vol. 86, No. 1, pp. 54-70.

Combes, P.P.; Duranton, G.; Gobillon, L. (2008) Spatial wage disparities: sorting matters! *Journal of Urban Economics*, 63, pp. 723-742.

_____ (2011) The identification of agglomeration economies. *Journal of Economic Geography*. Vol. 11, pp. 253-266.

Combes, P.P.; Duranton, G.; Gobillon, L.; Roux, S. (2010) Chapter 1. Estimating agglomeration economies with history, geology, and worker effects. In: Glaeser, E.L. *Agglomeration Economies*. The University of Chicago Press, pp. 15-66.

Combes, P.P.; Duranton, G.; Overman, H.G. (2005) Agglomeration and the adjustment of the spatial economy. *Papers in Regional Science*, Vol. 84, Issue 3, pp. 311-349.

Combes, P.P.; Mayer, T.; Thisse, J.F. (2008) *Economic Geography*. Princeton University Press, Princeton, USA.

De Groot, H.L.F; Poot, J.; Smit, M.J. (2009), Agglomeration externalities, innovation and regional growth: theoretical perspectives and meta-analysis. In: Capello, R.; Nijkamp, P. (Eds.) *Handbook of Regional Growth and Development Theories*. Cheltenham, UK: Edward Elgar, pp. 256-281.

Duranton, G.; Puga, D. (2004), Micro-foundations of urban agglomeration economies. In: Henderson, J.V.; Thisse, J.F. (Eds.) *Handbook of Regional and Urban Economics*. Vol. 4. Amsterdam: North-Holland, pp. 2063-2117.

_____. (2014). The growth of cities. In: Aghion, P.; Durlauf, S.N. (Eds.) *Handbook of Economic Growth*, Elsevier, 1st edition, Vol. 2, pp. 781-85.

Fally, T.; Paillacar, R.; Terra, C. (2010) Economic geography and wages in Brazil: evidence from micro-data. *Journal of Development Economics*, 91, pp. 155-168.

Fontes, G.G.; Simões, R.F.; Oliveira, A.M.H.C. (2010) Urban attributes and wage disparities in Brazil: a multilevel hierarchical model. *Regional Studies*, Vol. 44, No. 5, pp. 595-607.

Freguglia, R.S.; Menezes-Filho, N.A. (2012) Inter-regional wage differentials with wage heterogeneity: evidence from Brazil. *The Annals of Regional Science*, Vol. 49, Issue 1, pp. 17-34.

Fujita, M.; Thisse, J.F. (2012), *Economics of Agglomeration: Cities, industrial location and regional growth*. Cambridge: Cambridge University Press, 2nd edition.

Glaeser, E.L. (1998) Are cities dying? The Journal of Economic Perspectives, Vol. 12, No. 2, pp. 139-160.

_____. (2011), *Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier*. London, UK: Macmillan.

Glaeser, E.L.; Gottlieb, J.D. (2009) The wealth of cities: agglomeration economies and spatial equilibrium in the United States. *NBER Working Paper* 14806.

Glaeser, E.L.; Kallal, H.D.; Scheinkman, J.A.; Shleifer, A. (1992), Growth in cities. *Journal of Political Economy*, Vol. 100, pp. 1126-1152.

Glaeser, E.L.; Maré, D.C. (2001) Cities and skills. *Journal of Labour Economics*, 19 (2), pp. 316-342.

Graham, D.J.; Melo, P.S. Juwattanakulpaisarn, P.; Noland, R.B. (2010) Testing for causality between productivity and agglomeration economies. *Journal of Regional Science*, Vol. 50, No. 5, pp. 935-951.

Gordon, I.A.; McCann, P. (2000), Industrial clusters: complexes, agglomerations and/or social networks? *Urban Studies*, Vol. 37, No. 3, pp. 513-532.

Groot, S.P.T.; De Groot, H.L.F.; Smit, M.J. (2014) Regional wage differences in the Netherlands: micro evidence on agglomeration externalities. *Journal of Regional Science*, Vol. 54, Issue 3, pp. 503-523.

Henderson, J.V. (1986) Efficiency of resource usage and city size. *Journal of Urban Economics*, Vol. 19, pp. 47-70.

_____. (2010), Cities and development. *Journal of Regional Science*, Vol. 50, No. 1, pp. 514-540.

Heuermann, D.; Halfdanarson, B.; Suedekum, J. (2010) Human capital externalities and the urban wage premium: two literatures and their interrelations. *Urban Studies*, Vol. 47, No. 4,pp. 749-767.

IBGE – Instituto Brasileiro de Geografia e Estatística (2013) *Divisão Urbano Regional*.

Jacobs, J. (1969) *The Economy of Cities*. New York, Vintage.

Lindley, J.; Machin, S. (2014) Spatial changes in labour market inequality. *Journal of Urban Economics*, 79, pp. 121-138.

Maciente, A.N. (2013) The determinants of agglomeration in Brazil: input-output, labor and knowledge externalities. *PhD Dissertation*, University of Illinois at Urbana-Champaign.

Marshall, A. (1890) *The Principles of Economics*. London, Macmillan.

Melo, P.C.; Graham, D.J.; Noland, R.B. (2009) A meta-analysis of estimates of urban agglomeration economies. Regional Science and Urban Economics, Vol. 39, pp. 332-342.

Mion, G.; Naticchioni, P. (2009) The spatial sorting and matching of skills and firms. *Canadian Journal of Economics*, Vol. 42, No. 1, pp. 28-55.

Moretti, E. (2011) Local labor markets. In: Ashenfelter, O.; Card, D. (Eds.) Handbook of Labor Economics, Vol. 4B, Amsterdam: North-Holland, pp. 1237–1313.

Moulton, B.R. (1990) An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The Review of Economics and Statistics*. Vol. 72, No. 2, pp. 334-338.

Naritomi, J.; Soares, R.R.; Assunção, J.J. (2012) Institutional development and colonial heritage within Brazil. The Journal of Economic History. Vol. 72, Issue 2, pp. 393-422.

Pessoa, A. (2014) Agglomeration and regional growth policy: externalities versus comparative advantages. *The Annals of Regional Science*. Vol. 53, No. 1, pp. 1-27.

Porter, M.E. (1990), *The Competitive Advantage of Nations*, Houndmills: Macmillan.

Puga, D. (2010) The magnitude and causes of agglomeration economies. *Journal of Regional Science*, Vol. 50, N. 1, pp. 203-219.

Rosenthal, S.S.; Strange, W.C. (2004) Chapter 49 – Evidence on the nature and sources of agglomeration economies. In Henderson, J.V.; Thisse, J.F. (Eds.) *Handbook of Regional and Urban Economics. Volume 4*. Elsevier, pp. 2119-2171.

Silva, M.V.B.; Silveira-Neto, R.M. (2009) 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*, Vol. 13, No. 2, pp. 299-331.

Simões, R.F.; Freitas, E.E. (2014) Urban attributes and regional differences in productivity: evidence from the external economics of Brazilian micro-regions from 2000 – 2010. *International Journal of Economics*, Vol. 01, No. 2, pp. 30-44.

Starrett, D. (1978), Market allocations of location choice in a model with free mobility. *Journal of Economic Theory*, Vol. 17, pp. 21-37.

UNFPA (2007), *State of World Population 2007: Unleashing the Potential of Urban Growth*, New York: UNFPA.

Van Oort, F.G.; Lambooy, J.G. (2014) Cities, knowledge and innovation. In: Fischer, M.M.; Nijkamp, P. (Eds.) *Handbook of Regional Science*, Springer-Verlag Berlin Heidelberg, pp. 475-488.