Agglomeration Economies in the Presence of an Informal Sector
The Colombian Case

Gustavo A. García*

March 2016

Abstract

This paper analyzes the relationship between agglomeration economies and wages in the context of a developing country, taking into account the marked presence of an informal sector. Using data from Colombia, we investigate the effect of agglomeration economies on formal and informal productivity, inquiring whether the informal sector achieves benefits from agglomeration economies and whether there are differences between the formal sector and the informal sector in agglomeration returns. We estimate an elasticity of wages with respect to employment density of around -4% for the formal sector and around 3% for the informal sector; thus there is a significantly positive effect of agglomeration on the productivity of the informal sector. The results show that informal workers in a city twice as dense have around 2% greater productivity, that imply 14% higher wages in denser areas than in less dense areas. In contrast, in the formal sector the results show that formal workers in a city twice as dense have around 3% less productivity, leading this kind of workers to earn 17% less in denser areas. Factors associated with the constraints in the creation of formal jobs and a greater labor supply of formal workers and des-amenities very common in big cities in developing countries could explain this lower agglomeration return in the formal sector.

Keywords: Agglomeration gains; employment density; informality

JEL classification: *R12, J46, R23, J31*

*Postdoctoral Research Fellow, Department of Economics and *CEDE*, Universidad de los Andes. Email: ga.garcia@uniandes.edu.co. This paper has benefited from conference and seminar at the 62nd Annual North American Meetings of the Regional Science Association International (NARSC-2015) and EAFIT University Economics Department Seminar, and especially Harris Selod and Rosella Nocolini for their helpful comments and suggestions on an early version of the paper.

1 Introduction

The pace and content of urbanization have crucial implications for developing economies. Among the key benefits of urbanization are the gains of agglomeration. The hypothesis is that there are benefits of location externalities which arise of a dense network of production and market access links that increases the productivity and decreases the unit costs of each firm in the network (Fujita et al., 1999). It is possible to think, however, that the magnitude of agglomeration economies depends on the type of workers and industries, as well as on the period and country analyzed. In this sense, it is important to understand whether agglomerations economies would produce similar benefits for developing countries as have been previously demonstrated for developed countries (see, e.g. Ciccone and Hall, 1996; Rosenthal and Strange, 2008a; Melo et al., 2009; Melo and Graham, 2009).

Despite the fact that urbanization has continued at a fast pace in developing countries, formalization seems to have stalled, or at least does not appear to be increasing as rapidly as might be expected given country growth rates. The formal sector in developing economies is only responsible for a small share of urban employment and growth, while the informal sector plays a large role in the economy and this is an important difference compared with developed economies (Schneider and Enste, 2000). According to Jütting and De Laiglesias (2009) estimates, over 55% of non-agricultural employment in developing countries is performed in activities not regulated or protected by the state (informal activities). As for the size of the informal economy measured as a percentage of GDP, Schneider et al. (2010) show that in developing countries the shadow economy accounts for around 40% of GDP. This marked presence of the informal sector can affect the extent (or quality) of the agglomeration economies and the effects of urbanization could be as likely to be found in the outcomes for the informal sector as for the formal sector.

Given the large differences in economic characteristics between the formal and informal sectors, for example in terms of productivity, profitability, and firms size, there are different points of views regarding the informal sector contributes to and benefits from agglomeration economies. For instance, Annez and Buckely (2009) argue that the informal sector is unproductive and increases the costs to the formal sector, crowding out agglomeration economies. In contrast, Overman and Venables (2005) and Moreno-Monroy (2012) state that the informal sector also contributes to and benefits from agglomeration economies via the interaction with the formal sector along the value chains, where the informal sector not only obtains inputs from, but also supplies intermediate and final goods and services to the formal sector. As pointed out by Duranton (2009), there are intense linkages between the formal and informal sector, which suggests that agglomeration effects are generated within both sectors, with benefits that accrue to both. According to Overman and Venables (2005) the existence of an informal sector can affect the benefits of agglomeration economies in two possible ways. On the one hand, the informal sector can drive up urban costs and crowd out the formal sector. On the other hand, the informal sector contributes to agglomeration economies. Therefore, the informal sector is made up of small enterprises producing on a small scale, which establishes important networks that contribute to the formation of clusters. Furthermore, as in the formal sector, the informal sector can achieve benefits from the productivity effects associated with the concentration of the activity and employment.

In this paper, we investigate the effect of agglomeration economies on formal and informal productivity, and an analysis performed of which sector, formal or informal, achieves greater benefits from the diversity of activities and those spillovers associated with urbanization economies. This analysis is carried out by using data at the worker level for Colombia throughout the period 2008-2014. The empirical analysis is based on the regressions of individual worker wage rate as measurement of labor productivity on employment density as a measurement of urban agglomeration, measuring the elasticity of wages with respect to density for the formal and informal sector, and controlling by several socioeconomic, socio-demographic and regional characteristics. These regressions comprise instrumental variables estimates to correct for the endogeneity attributable to the reverse causality between wages and agglomeration.

The purpose of this study is to provide new evidence in the context of a developing country on urbanization and its effects in developing countries, considering more closely the reality of these countries where co-exist formal and informal activities. To the best of our knowledge, this is the first paper that studies the agglomeration effects in developing countries by taking into account the presence of the informal sector.

The remainder of this paper is organized as follows: Section 2 presents the literature review. The theoretical framework is described in Section 3. Section 4 shows the empirical model used. Then, in Section 5 we present the data sources used in the analysis. Section 6 documents statistically the relationship between agglomeration and wages taking into account the existence of the informal sector. Section 7 discusses the results, and Section 8 concludes.

2 Literature review

In this section we provide a brief outline of the empirical results on the presence of agglomeration economies in some developing countries. The related literature is scarcity and in general it finds that agglomeration economies present a large effect on productivity, much higher compared to estimates for develop countries.[1]

[1] The two benchmark studies that use aggregate data for the US, Ciccone and Hall (1996) and Rosenthal and Strange (2008a) for the years 1988 and 2000 respectively, report values for the elasticity of

Developing countries are characterized by certain structural conditions such as economic and political instability, high rates of unemployment and underemployment, shallowness of markets, and low industrial and infrastructure development. These conditions can affect the extent of external economies associated with agglomeration economies. Several studies have found quantitative evidence of both localization and urban effects. In the seminal work by Henderson (1986), which analyzes the role of localization and urbanization economies in productivity in the metropolitan areas of Brazil, it was found that localization economies play an important part in this regard, while urbanization economies are present, but only weakly so. Its results show that if employment in any sector in any region were to double, productivity measured by value-added would increase by 11%. A comparable result is found by Lee and Zang (1998) in their study of manufacturing industry in South Korea. The authors found that doubling the employment of a given sector and region is associated with an increase in value-added per worker as a measurement of productivity of 7.9%. From Indian cities, the studies by Mills and Becker (1986), Becker et al. (1992), and Shukla (1996) show that equally significant increases in productivity are generated by urbanization.

Among more recent studies we have for instance to Da Mata et al. (2007) who found for the case of the Brazilian cities that the urban elasticity, measuring urbanization economies as market potential, is 11%. Similarly, Combes et al. (2013) that study the effect of density on individual wages in 87 Chinese cities and instrument density by peripherality, the historical status of the city and distance to historical cities, found that the elasticity of wages with respect to density is between 10% and 12%. For the case of India, Chauvin et al. (2014) evaluate the effect of density on individual annual earnings at the district level and also found a large elasticity around 9%-12%. In the Colombian context, the only work is by Duranton (2014), who, correcting by endogeneity with lagged population, finds that doubling the city population as measurement of urban agglomeration is associated with an increase in productivity measured by the workers wage by about 5%. Nonetheless, these evidences have an important bias, given that most of the findings are concerned with the formal sector and do not take into account the informal sector. In this sense, more research need to be pursued to gain knowledge on agglomeration effects in developing countries taking into account the presence of the informal sector and this study will hopefully shed lights on this issue.

productivity with respect to density at around 4%-5%. Combes et al. (2008) and then Combes et al. (2010) for French cities use individual data to estimate the effect of density on wages as a measure of productivity. They find an elasticity of wage with respect to density at around 3%. Using also individual data for Spain, Italia, UK and The Netherlands, de la Roca and Puga (2012), Mion and Naticchioni (2009), D'Costa and Overman (2014), and Groot et al. (2014), obtain values of 2.5%, 1%, 1.6% and 2.1%, respectively.

3 Theoretical framework

In this section we present a model on which we structure our empirical specification. The theoretical model derives in a wage model which allows testing the impact of agglomeration economies on local productivity, measured by nominal wages (Combes et al., 2008; Melo and Graham, 2009; Combes and Gobillon, 2015).

The profit of a competitive representative firm located in area a in year t is:

$$\pi_{a,t} = p_{a,t}Y_{a,t} - w_{a,t}L_{a,t} - r_{a,t}K_{a,t} \tag{1}$$

where $Y_{a,t}$ is the output of the firm which uses two inputs, labor $L_{a,t}$ and other factors of production $K_{a,t}$, such as land, capital or intermediate inputs. $p_{a,t}$ is the price of the good produced, $w_{a,t}$ is the wage rate on the local labor market, and $r_{a,t}$ is the unit cost of non-labor inputs.

Suppose that the production function of the firm is Cobb-Douglas and can be represented by the following equation:

$$Y_{a,t} = A_{a,t}(s_{a,t}L_{a,t})^{\alpha}K_{a,t}^{1-\alpha} \tag{2}$$

where $0 < \alpha < 1$ is a parameter, $s_{a,t}$ denotes the local labor skills, and $A_{a,t}$ is the local total factor productivity. In a competitive equilibrium, the first-order condition for the optimal use of labor is given by the following expression:

$$w_{a,t} = \alpha p_{a,t}A_{a,t}s_{a,t}^{\alpha}\left(\frac{K_{a,t}}{L_{a,t}}\right)^{1-\alpha} \tag{3}$$

Using the first-order condition for profit maximization with respect to the other factors, reordering in terms of $K_{a,t}/L_{a,t}$, and inserting it in equation (3), we obtain:

$$w_{a,t} = \alpha(1-\alpha)^{(\frac{1-\alpha}{\alpha})}\left(p_{a,t}\frac{A_{a,t}}{(r_{a,t})^{1-\alpha}}\right)^{\frac{1}{\alpha}}s_{a,t} = B_{a,t}s_{a,t} \tag{4}$$

From this last expression we note that the local average nominal wage depends positively on labor skills, $s_{a,t}$, the output price, $p_{a,t}$, and the technological efficiency of the local economy, $A_{a,t}$. We also can observe that, the local level of wages is negatively determined by the costs of the other non-labor input factors, $r_{a,t}$. The effects of agglomeration and dispersion forces work through these three factors. The output and input prices, $p_{a,t}$ and $r_{a,t}$, capture a number of agglomeration mechanisms operating through local markets, sometimes referred to as 'pecuniary externalities', while the local environmental efficiency, $A_{a,t}$, captures the effects from the pure local externalities or 'technological externalities' which are not mediated by the market.

According to Combes and Gobillon (2015), there are two types of 'technological externalities'. On the one hand, firms and consumers are grouped together in the cities where they share indivisible good such as airports, libraries, museums, universities, hospitals. In this situation all the local market participants benefit from the infrastructure via reduction of costs of access to this public goods, since this costs are spread across all the beneficiaries. This generates a first type of agglomeration economies where the local total factor productivity, $A_{a,t}$, is larger in larger cites due to the presence of local public goods, increasing the composite labor productivity effect, $B_{a,t}$, and therefore the local wages. The second type of 'technological externalities' emerges when spatial concentration induces local knowledge spillovers that make firms more productive. This type of mechanism again makes $A_{a,t}$ large in large cities.

The previous discussion has attempted to explain how city size generates agglomeration economies. However, observation also tell us that city size not only generates agglomeration economies but also dispersion forces. Typically, excess concentration in large cities can imply negative externalities due to congestion, such as longer commuting costs and scarce land for housing and plants. Congestion on local public goods implies a reduction of $A_{a,t}$, while scarce land can lead to that the costs of inputs that are not perfectly mobile, $r_{a,t}$, are higher in large cities. These space constraints work as a dispersion force that presents a negative effect on the local wage (Tabuchi, 1998).

Note that the composite labor productivity effect, $B_{a,t}$, is affect by both the pure externalities, $A_{a,t}$, and the effects related to good or input prices, $p_{a,t}$ and $r_{a,t}$. However, with this approach we can estimate the overall effect of local characteristics, but not the exact channel through which agglomeration economies work. In other words, we cannot identify price and technology effects separately, we can only estimate the combined net overall effect of all three mechanisms: $\left(p_{a,t}\frac{A_{a,t}}{(r_{a,t})^{1-\alpha}}\right)^{\frac{1}{\alpha}}$. Furthermore, note that the correlation between wage and density only show the overall impact of both agglomeration economies and dispersion effects. While the net effect of spatial concentration is identified, it is not possible identify these effects separately (Duranton and Puga, 2004; Rosenthal and Strange, 2004; Combes et al., 2008).

4 Estimation strategy

To take equation (4) to the data we formulate the following wage equation that expresses the theoretical model above into the empirical model to be estimated:

$$\ln w_{i(t)} = \alpha_0 + \beta \ln density_t + X_{i(t)}\varphi + \pi_{oi(t)} + \sigma_{ei(t)} + \eta_{ai(t)} + \delta_t + \epsilon_{i(t)} \tag{5}$$

where i identifies the worker, o refers to occupation, e refers to the economic sector,

a identifies the region and t specifies the time period. The "$i(t)$" subscripts indicate that the observations are an independent cross-sectional series where N individuals are only available in each period. The dependent variable is the logarithm of the monthly wage.

Our measurement of urban agglomeration is the logarithm of employment density of the municipalities, $\ln density$, which is defined as the number of workers per square kilometer in each municipality using an average over the period 2008-2014.[2] The basic idea behind this variable is that a high density is a potential source of increasing returns resulting from stronger knowledge and technological spillovers in areas of dense economic activity.

As mentioned in the theoretical framework, this measure of agglomeration economies can measure various factors at the same time. Firstly, this variable reflect the quality of urban life which is expressed in higher urban rents. According to Roback (1982) this aspect of the quality of urban life is capitalized as higher wages in cities. Thus, the coefficient on employment density should be positive. Secondly, it can measure the level of negative amenities or des-amenities in denser area, including congestion, pollution, and noise. Des-amenities would make working in a denser area unpleasant which could be expected to be compensated by a higher wage (Borjas, 2008; Lee, 2014), leading the coefficient of employment density to be positive. Lastly, a denser area can imply a plentiful labor supply that could decrease wages, which would make the coefficient negative. In the same sense, Rosenthal and Strange (2008a) showed that this negative sign could occur whether there is a limited amount of job creation of some type of employment, then having more workers of this type might tend to result in each worker earning lower wages. If the abovementioned three factors collide, it would be difficult to determine the sign of the coefficient on employment density, therefore this sign should be tested empirically.

It is important to note that we use the municipality as the spatial unit of analysis, and although this is not an ideal unit, it is the best available approximation of a self-contained labor market in Colombia. The municipalities are areas where a high proportion of people who live (work) in the area also work (live). As Dominicis et al. (2007) argues, if there is evidence of a concentration of residential activities, of work activities as well as of those social relationships that are created within it, this area can be considered as a self-contained labor market or a Local Labor System.[3]

The vector $X_{i(t)}$ contains the variables that measure a standard set of demographic

[2]Following Combes and Gobillon (2015) we prefer to use employment rather than population, because it better reflects the magnitude of local economy activity. In addition, the results using population are in general very similar that those obtained with employment. The results using population variable are not presented here to save space but can be provided upon request.

[3]As Openshaw and Taylor (1979) have pointed out, the municipalities or metropolitan areas are much more related to the concept of local labor markets than the usual administrative areas, so they are a good option for overcoming the Modifiable Areal Unit Problem (MAUP).

attributes such as the worker's level education, gender, age and its square, and years in the current job and its square. In addition, in our model we included sets of dummy variables to control for several sources of heterogeneity that can lead to an omitted variable bias and inconsistency of the model parameter estimates. In order to capture macro level changes in wage rates that are common to all individuals, we include time dummies, δ_t. Similarly, to control for current skills we added a set of occupation dummy variables, $\pi_{oi(t)}$. We also included a set of dummy variables for controlling by economic sector heterogeneity and regional characteristics, these are represented by $\sigma_{si(t)}$ and $\eta_{ai(t)}$.

Equation (5) can be estimated in several ways. The most straightforward one consists of splitting the sample by the formal sector and the informal sector and estimating the model for each sector. Nevertheless this means that the coefficients of individual explanatory variables are not constrained to be the same across sector, which may or may not be relevant from a theoretical point of view. This also entails a loss of precision for the estimators (Combes and Gobillon, 2015). An alternative approach consists of considering among explanatory variables some interactions between density and a sector dummy, and estimating this specification. We follow this last specification to avoid the loss of precision in the estimators that occurs when is divided the sample by sectors. Thus, our model to estimate has the following structure:

$$\ln w_{i(t)} = \alpha_0 + \beta \ln density_t + \phi \ln density_t * formal_{i(t)} + X_{i(t)}\varphi + \pi_{oi(t)} + \sigma_{si(t)} + \eta_{ai(t)} + \delta_t + \epsilon_{i(t)} \tag{6}$$

where $formal$ is a dummy variable equal to 1 if worker is formal and 0 if is informal. We define informal workers as those workers who are not covered by the health insurance and the pension system.

Another aspect to be considered in the estimation is the endogeneity bias caused by reverse causality between wages and agglomeration. Wages can increase due to higher employment density, but it is also possible that higher wages may attract more people and firms to a given area. In order to avoid the endogeneity bias we implemented instrumental variable (IV) techniques. In the literature long-lagged values of endogenous variables have been widely used as instruments since Ciccone and Hall's (1996) pioneering work. The basic idea behind these instrument variables is that deep time lags of urban density can to some extent explain the distribution of present densities, but they do not explain the distribution of current urban productivity levels.

To construct our instruments for current density following the idea by Ciccone and Hall (1996), we use population data collected from 1912, 1918 and 1928 censuses. Although there are previous national censuses in Colombia, we preferred to take censuses from the 1900s because most of the current municipalities were created at the end of 1800's and at the beginning of 1900's. As such, we have complete information of past population for

390 municipalities.

We have to take into account that to yield unbiased estimates in the estimation of the effect of density on wages by using instrument variables, our instruments must satisfy two conditions for it to be valid, namely relevance and exogeneity. While the first condition demands that our instruments be correlated with the contemporaneous employment density, the second condition requires that our instrument be uncorrelated with the error term $\epsilon_{i(t)}$. As it has been mentioned by Combes and Gobillon (2015), it is possible to imagine a number of possible violations caused by alternative links between past population and current wages, such as local permanent characteristics that may have affected past location choices and still affect local productivity today, for instance the centrality of the location in the country, a suitable climate, or geographical features like access to the coast or presence of a large river. To minimize potential problems, we control for geographical characteristics in regressions and try to preclude such correlations and that local historical population is exogenous. The details of the test of relevance and exogeneity of the instrumental variables are presented in Section 7.

5 Data and variables

A number of literature on agglomeration economies use detailed spatial data on panel of workers or firms (see for example, Combes, et al., (2010) and Glaeser and Maré (2001)) which allows greater administrative scale analysis and to control for unobserved individual characteristics that may be correlated with locations choices. Unfortunately, this kind of data is not available in Colombia and in general in most developing countries. Instead we use a cross-section survey, the Colombian Great Integrated Household Survey (GEIH) carried out by the National Administrative Statistics Department (DANE). By using cross-section data is not possible to control for all the characteristics of the individual shaping their skills that do not change over time and the effect of which can be considered to be constant over time (Combes and Gobillon, 2015). However, there are various measures of observed skills which can be used at the cost of not controlling for unobservable individual characteristics. For instance, Duranton and Monastiriotis (2002), and Wheaton and Lewis (2002) used measures such as diplomas or years of education. Another used measure is the socio-professional category, “occupation”, which captures the exact job done by workers and part of the effects of past career, and may thus be considered as a measure that should be more correlated with current skills than education. Given that the GEIH gathers detailed information about general characteristics of populations (such as gender, age, year of education and municipality of residence), as well as about the employment conditions (whether they work, what they do, how much they earn, number of hours worked or whether they have social security for health care), we include education and occupation as measures of current skills of the workers. These kinds of measures are often recorded in labor force surveys and could allow greater comparability across developing countries.[4]

We analyzed the period between 2008 and 2014. Databases for earlier years are not comparable because several methodological changes were carried out by DANE in 2007. After excluding individuals with no labor income, those who did not report municipality of residence and eliminating the 1% of workers with the lowest and highest wages every year, we have 1,920,678 observations, with a mean of 270,000 observations per year and information for 568 municipalities.[5]

We divide workers into formal and informal workers. As mentioned, informal workers are defined to be individuals who do not have access to the social security system to receive healthcare and a retirement pension. Note that this definition has been widely used by prior research, including Jütting and De Laiglesia (2009), Perry et al. (2007), among many others. Following this definition of informality, we can observe in Table 1 that around 60% of employees in Colombia are informal workers and is a very persistent phenomenon.

[Insert Table 1 around here]

In addition to account for by the formal sector and the informal sector in the models, we also control for a standard set of demographic attributes. These include the worker's level of education, gender, age, years in the current job, occupation, economic sector and regional and geographical variables such as five regional indicators (Central, Oriental, Occident, Caribbean and Orinoco), water availability, soil erosion and altitude of municipality. We also include two measure of market access: the distance in kilometers and time distance to the capital city of the department. In Table 1 and 2 we show some descriptive statistics of these variables which were calculated using person sampling weights from GEIH to ensure that the estimates are representative.

[Insert Table 2 around here]

For all of the models, we use the log of monthly wages as the dependent variable. As mentioned, our measure of urbanization is the log employment density of the municipalities. Municipalities have an average of roughly 26,000 workers and range from 341

[4]Given the confidentiality of the data at municipal level, all the estimations in this paper were conducted following DANE's microdata-access policy, which implies working in situ under the supervision of DANE's staff and with blinded access to sensible information.

[5]Colombia covers an area of roughly 1,200,000 Km2 and is divided in 32 administrative units called departments and a Capital District that is the countrys capital, Bogotá. Departments are country subdivisions similar to US states and are granted a certain of autonomy. Each Department is composed of municipalities where there is a capital city of this department. In total Colombia has 1,119 municipalities (a more detailed characterization of Colombia can be found in Royuela and García (2015) and Nicodemo and García (2015)).

workers to over three million. Figure 1 shows the employment density by municipality. We can note that Bogotá, Medellín, Itaguí, Cali, Bucaramanga, Barranquilla and Soledad are the cities with the highest levels of urbanization, where Itagüí being the most dense city in Colombia with just over 4,500 workers per Km2.[6]

[Insert Figure 1 around here]

6 Documenting the agglomeration-wages relationship in the presence of an informal sector

We begin with an illustration that stress the paper's themes. Table 3 shows average monthly wages by formal and informal employees for the three largest municipalities and municipalities with less than 5000 inhabitants. We can observe that there is a clear relationship between wages and agglomeration. For formal employees, average wages are similar for the two groups of cities, in fact, the city wage gap decrease along time and in 2014 formal workers earn higher wages in small cities than in big cities. In contrast, informal workers earn substantially higher wages in the larger cities. Taken as a whole, Table 3 suggests that there is a positive relationship between agglomeration and wages for informal workers, but not for formal workers which seems to show a negative relationship.

[Insert Table 3 around here]

In order to confirm theses relationship between agglomeration and wages by sector, we plot log wages against employment density for 568 municipalities for the formal sector and the informal sector separately. Figure 2 (a) shows that for the total the slope of the regression line between log wages and log density which measure the density elasticity of wages is 1%. Regarding, the formal and informal sectors, Figure 2 (b) and (c) show that while for the formal sector the density elasticity of wages is -3.0%, for the informal sector this elasticity is 0.6%, confirming our previous results that there are less agglomeration returns for formal workers than for informal workers.

These results are somewhat surprising because formal workers are workers with more education than informal workers and might have a greater ability for learning from nearby human capital. Furthermore, formal workers work in medium-large enterprises and this kind of enterprises can achieve greater benefits of labor market pooling and input sharing associated with agglomeration (Rosenthal and Strange, 2008b). On the other hand, informal workers are characterized by having a limited education and working in very small enterprises (Jütting and De Laiglesia, 2009; Perry et al., 2007) which might imply less

[6] In terms of population, these same cities present the highest levels of population and Itagüí is the most densely populated cities in Colombia with 12,114 people per Km2 in 2014.

ability to absorb the knowledge, and the activities of small enterprises are more geared towards small local markets than towards generating input-output linkages (Moreno-Monroy and García, 2016).

One possible explanation for these results could be that given that there is a limited creation of formal jobs in the economy, then having more formal workers might tend to result in each worker earning lower wages. This kind of work-spreading would imply the opposite sing on employment density (Rosenthal and Strange, 2008a). The possibility that workers might concentrate in this way in equilibrium is consistent with the Harris-Todaro (1970) model which in a context of industrialization in a developing country shows that when the urban wage is fixed above the market-clearing level, there can be unemployment in equilibrium, unemployment undercover in the informal sector.

[Insert Figure 2 around here]

7 Results

7.1 Baseline estimation

In this section we present the results of the estimation of wage equations by OLS, reported in Table 4. To simplify presentation, only the coefficients on the elasticity of wages with respect to density are provided (both here and in the following table). We can observe that the specification without any other control in column 1 reports an elasticity of around 5%, which indicates that when density is twice as great, productivity is 3% higher.[7] When we add control variables to this estimation without informality effects there is a reduction on elasticity and reaches a value of 4% (see Table A1 in the Appendix). We now include the differential effect of the formal sector and the informal sector (column 2). Note that adding the formal variable the explanatory power of the regression increases substantially, which can indicate that informality account for a sizeable fraction of spatial wage disparities in Colombia.

[Insert Table 4 around here]

The results of the elasticity of wages with respect to density show that in the formal sector this elasticity is -5.8% (-10.8%+5%) and significant which suggests that formal workers in a city twice as dense have around 4% less productivity. In contrast, the elasticity among informal workers is 5% and highly significant, indicating that this kind of workers in a city twice as dense have around 3.5% greater productivity. This difference

[7] We follow the formula of Combes and Gobillon (2015): $2^{\beta}-1$, where β is the elasticity of productivity with respect to density.

between formal and informal workers echoes the summary measure in Table 3 and Figure 2 and will persist throughout the paper.

Column 3 and 4 adds individual characteristics and occupation and economic sector as control variables, respectively. This divides the elasticity of the formal sector of column 2 by a factor larger than two and reaches a value around -2%, while in the informal sector the elasticity presents a slight decreasing and is allocated around 4.7%. This suggests that in the formal sector more than half the relationship between wages and employment density is explained by denser cities hosting more educated workers, which is consistent with the fact that there is a higher share of more educated formal workers in larger cities.

So far it has been found that a city density elasticity of wages of -2% and 4.7% for the formal sectors and the informal sector, respectively, are quantitatively important. Comparing a small municipality with a density of 20 workers per Km2 to Bogotá with a density around 2000 workers per Km2, these elasticities imply that in the formal sector workers in denser cities will earn 9% less than in less dense cities, whereas in the informal sector the wage difference is 24% in favor of informal workers in denser cities. The lower productivity levels of formal workers in denser cities and the wage differences across municipalities in the informal sector, are certainly important factors accounting for spatial wage disparities in Colombia.[8]

When we include geographic variables in the model (column 5) such as regional indicators, water availability, soil erosion, and altitude, the coefficient on city density in the informal sector slight increases, reaching a value of 5.5%, and in the formal sector this coefficient increases to -1.2%. We also can observe that including geographic controls does not increase substantially the explanatory power of the regression, in fact, although these results are not reported, the coefficients on several geographic controls are not statistically significant. On the other hand, we found that wages are higher in the Oriental and Orinoco regions of Colombia relative to the Central region, while there are no wages differences between the Caribbean and Occident region and the Central region.

Column 6 and 7 duplicate column 4 but adds the measures of market access and its square. The results show that these variables are not significant and their inclusion not present important effects on the coefficients of the log density of the formal and informal sector. In the formal sector the coefficient on the log density is -2.3% and in the informal sector this coefficient is 4.5%.

We now turn to analyze the possible heterogeneities in agglomeration effects. Column 8 attempts to detect non-linearities by adding the square of log density as independent variable to the specification of column 4. The findings show that the coefficients on log

[8]Table A1 in the Appendix shows that the elasticity of wages with respect to density is 4% when it is controlled by individual characteristics. This elasticity implies a wage difference of 20% between big cities and small cities.

density and the quadratic term are insignificant in the informal sector, while in the formal sector they are significant. However, when we carry out a joint test on the coefficients of the lineal terms of the log density variable, on the one hand, and a joint test on the coefficients of the quadratic terms of the log density, on the other hand, we found that the two test show that these coefficients are not joint significant, which indicate an absence of non-linearity in agglomeration effects. It is consistent with the results found by Duranton (2014) for Colombia with data between 1996 and 2012.

Lastly, in columns 9 and 10 we add interaction terms to specification of column 4. Column 9 adds the product of the worker's number of year of education by log density, and column 10 includes two products: the product of the workers number of year of education by log density, and the product among the workers number of year of education, log density and formality. The results in column 9 show that the coefficient on the interaction term is very small, negative, and highly significant, indicating that there is higher agglomeration returns for less educated workers, although this return is small. In order to determine whether this negative interaction effect differ by employment status, the results in column 10 show that the coefficients on the interaction term between education and log density is negative and significant, although very small, whereas the second interaction term among education, formality and log density is positive and significant. This suggests again that less educated workers obtain higher agglomeration returns, but once it is distinguished by formal and informal workers we find that formal workers more educated obtain higher returns in greater cities.

However, this general negative effect of lower agglomeration returns for more educated is contrary to the results from extant literature for developed countries which highlight the existence of higher returns to cities for more educated (Rosenthal and Strange, 2008a; Bacolod et al., 2009; Glaeser and Resseger, 2010). The idea is that there is a complementarity between city density and individual skills which is an important factor explaining the over-representation of more skilled workers in large cities. In large cities there are urban amenities that are more enjoyed by more educated workers. Nevertheless, although this over-representation can occur in Colombia, the large cities in this country, and in most large cities in developing countries, present important urban des-amenities, such as pollution, traffic congestion, crime, excess garbage, with which more educated worker could be more sensitive to these des-amenities affecting the benefits of agglomeration for this group.

7.2 Dealing with reverse causality between wages and agglomeration

We now turn to analyze the 2SLS estimations which use 1912, 1918 and 1928 populations as instrument for contemporaneous working populations. We began by discussing the instrument diagnostic test reported at the bottom of the Table 5.

Regarding exogeneity condition of the instruments, we inspect the Hansen's J (1982) to test the null hypothesis of exogeneity of the long-lagged instruments. The results for instruments exogeneity for all models are in agreement with previous studies using similar instruments: the null hypothesis of exogeneity is not rejected at a 5 percent level of significance, suggesting that the instruments are exogenous.

[Insert Table 5 around here]

With regard to the relevance of the instruments, the first stage regressions results indicate that the instruments for city density have considerable explanatory power. The Shea (1997) partial R-squared score values that range between 0.51 and 0.98. To further inspect the relevance of the instruments we carry out the Kleibergen-Paap test of under-identification which tests whether the model is identified, where identification requires that the excluded instruments are correlated with the endogenous regressor. When the instruments are uncorrelated with the endogenous regressor, the matrix of reduced-form coefficients is not of full rank and the model will be unidentified. Since we allow intra-group correlation, the relevant statistic in this case is the Kleibergen and Paap (2006) rank LM statistics. If we fail to reject the null hypothesis that the matrix of reduced-form coefficients is under-identified, it means that the instruments variables bias of the parameter estimates will be increased. The values presented in Table 5 for all models show that the tests reject the null hypothesis of under-identification at a 5 percent level of significance, implying that the instruments are relevant.

Nonetheless, a rejection result for the null hypothesis in the Kleibergen-Paap test should be treated with caution because weak instrument problems may still be present. Weak identification arises when the instruments are correlated with the endogenous regressor, but only weakly. As pointed out by Murray (2006) and Stock and Yogo (2005) when the instruments are poorly correlated with the endogenous regressors, the estimates from the instrumental variable model will be biased. In this case, and allowing intra-group correlation, the relevant test is the Kleibergen-Paap (2006) rank Wald F statistic. This test involved testing the significance of the excluded instruments in the structural equation, which results in the substitution the reduced-form expression for the endogenous regressor in the main equation for the model (Baum et al., 2007; Davidson and MacKinnon, 2010). The critical values for this test are from Stock and Yogo (2005). The results reveal that the Kleibergen-Paap (2006) rank Wald F statistic is higher than the Stock and Yogo (2005) critical values, suggesting that our instruments are not weak.

Consider now the estimates of the impact of agglomeration on wages. In general terms, we can note that the 2SLS coefficients on log city density for the formal and informal sector are lower than their OLS counterpart. In columns 2 and 3, where we control for individual characteristics and occupation and economic sectors variables, the 2SLS elasticity of wages with respect to density for the formal and informal sector are around -4% and 3%, respectively, instead of -2% and 4.7% for their OLS counterparts. These differences are important, more than one deviation standard, and significant, which suggests of an upward bias in the OLS coefficients of Table 4. These results confirm our previous findings that the formal sector achieves greater benefits from agglomeration economies than those obtained by the formal sector. As mentioned, this opposites sing in the relationship between wages and agglomeration in the formal sector is consistent with work-spreading in which the limited creation of formal jobs in the economy can lead to a reduction of wages when more workers enter in this sector.

On the other hand, comparing the estimates of column 1 with those of column 2 and 3, we can observe that the elasticity of wages with respect to density in the formal sector is revised downward, suggesting, as in the OLS estimates, that in this sector an important part of the relationship between wages and agglomeration is explained by the fact that in denser cities there are more educated formal workers.

These elasticities of employment density correcting for endogeneity of agglomeration show that while formal workers in a city twice as dense have around 3% less productivity, informal workers in the same city twice as dense have around 2% greater productivity. Although these values are revised downwards with respect to those obtained from the OLS estimator, they are still quantitatively important. Again comparing a city with low employment density to Bogotá, which presents a high employment density, these elasticities suggest that formal workers in denser areas will earn 17% less than in less dense areas, and in the informal sector wages will be 14% higher in denser areas than in less dense areas.

Regarding the following columns of Table 5, we can note that including geographic controls (column 4) and market access variables make no change to the coefficients on city density for the formal and informal sector in column 3. In column 7 appears again with the same sign and significance of the coefficients for the lineal and quadratic terms, which supports the evidence above of absence of non-linearities in agglomeration in Colombia.

Finally, columns 8 and 9 of Table 5 show that the sing and significance of the coefficients on city density and interaction terms are similar to its corresponding OLS counterpart, in fact, the coefficients on the interaction terms between education and density, and among education, formality and density are equal to the OLS coefficients. These results confirm the evidence found above in which less educated workers obtain higher agglomeration returns and in particular those that are in the informal sector. Once again, the urban des-amenities in large cities in developing countries could be an important factor affecting the returns of the agglomeration for more educated workers.

In order to assess the robustness of the results, Table 6 shows the results when we experiment with the instrumentation strategy and the estimation technique. Table 6 uses the same specification as column 3 of Table 5. Column 1 uses only 1912 population as instrument for contemporaneous employment, while columns 2 and 3 use only as instrument 1918 and 1928 population, respectively. The coefficients on employment density for the formal and informal sector in columns 1 and 2 are exactly the same: -0.042 and 0.027, respectively. In column 3 the coefficient on employment density for both sector increases slightly to -0.040 for the formal sector and 0.029 for the informal sector. Similar estimates to our baseline model are obtained again in column 4 using only 1912 and 1928 population as instrument.

Given the importance of Bogotá, the capital of Colombia, we ask whether the results would change if we exclude this city from analysis.[9] Column 5 in Table 6 shows the results of regression excluding observations corresponding to Bogotá. Accordingly, the elasticities are lower than when the capital city is included, but the difference is no sizable: for the formal sector the elasticity is equal to -0.044, while in the informal sector is 0.025. Finally column 6 and 7 use 1912, 1918 and 1928 population as in Table 5 but estimate the regression with the generalized method of moments (GMM) and with limited-information maximum likelihood (LIML) instead 2sls. As a result in the case of GMM estimates, the coefficients on employment density increase and reach values of -0.035 for the formal sector and 0.035 for the informal sector. Using LIML the results are again similar as in the first columns. In general, it appears that the instrumented coefficients on employment density by sector are generally statistically similar to our baseline 2sls of 0.039 in the formal sector and 0.030 in the informal sector. These results are robust to the choice of specification, instruments and estimation technique.

[Insert Table 6 around here]

8 Conclusions

This paper carries out a first evidence about the relationship between agglomeration economies and wages in a developing country, Colombia, account for the presence of a large informal sector. Among the main results we have found that there is an elasticity

[9] According to Arango and Bonilla (2015) and García (2014), Bogotá accounts for about 17% of total population of the country, 25% of total employment, and 25% of GDP.

of wages with respect to employment density of around -4% for the formal sector and around 3% for the informal sector. These coefficients are very robust to the inclusion of a wide variety of controls and the use of different specification and instrumental variable strategies.

These elasticities suggest that among formal workers, agglomeration tends to decrease wages, while among informal workers, the pattern is different, with agglomeration increasing wages. In quantitative terms these elasticities indicate that, moving from a city with a density around 20 workers per Km2 to Bogota with a employment density of around 2000 workers per Km2 is associated with about 17% lower wages in the formal sector and 14% higher wages in the informal sector. This sizable lose from urbanization for the formal workers may be due to the constraints in the creation of formal jobs at urban level which result in a greater labor supply of this type of work, leading to lower wages in denser areas for formal workers. The paper is, therefore, one of very few to have provided empirical evidence in supporting that there are positive agglomeration returns in the informal sector, and these are higher than those achieve in the formal sector.

We also found novel evidence of higher agglomeration returns for less educated worker. This result is contrary to the results from extant literature for developed countries that highlight the existence of higher agglomeration returns for more educated people. One possible explanation for this result could be the existence of important des-amenities in denser cities in developing countries which are not compensated by wages, in particular for more educated workers who are more sensitives to these negative amenities that make working unpleasant and therefore affect the benefits of agglomeration. However, it is important to note that further studies are necessary in order to understand better this inference.

This paper also contributes to the literature on agglomeration economies related to agglomeration also encourage hard work (Rosenthal and Strange, 2008a), in this case, informal work. According to literature on agglomeration, cities are productive places because they allow for pooling of labor, sharing of intermediate inputs, and knowledge spillover, and informal workers also achieve benefits of these productive effects manifested in higher wages in denser cities.

9 References

Annez, P. and Buckely, R. (2009) Urbanization and Growth: Setting the Context. In M. Spence, P. Annez, and R. Buckley, (ed) *Urbanization and Growth*, the Commission Growth and Development, Washington, D.C.

Arango, L. and and Bonilla, G. (2015). "Human Capital Agglomeration and Social Returns to Education in Colombia". *Borradores de Economía* 883, Banco de la República, Bogotá-Colombia

Bacolod, M., Blum B., and Strange, W. (2009) "Skills in the City". *Journal of Urban Economics* 65(2):136153

Baum, C., Schaffer, M. and Stillman, S. (2007) "Enhanced Routines for Instrumental Variables/Generalized Method of Moments Estimation and Testing". *Stata Journal*, 7:465-506.

Becker, C., Williamson, J. and Mills, E. (1992) *Indian Urbanization and Economic Growth since 1960.* Baltimore, MD: Johns Hopkins University Press

Borjas, George J. 2008. Labor Economics. Boston: McGraw-Hill.

Ciccone, A. and Hall, R. E. (1996) "Productivity and the Density of Economic Activity". *American Economic Review*, 86: 54-70

Chauvin, J.P., Glaeser, E., and Tobio, K. (2014). Urban Economics in the US and India. Harvard University.

Combes, PP., Duranton, G., and Gobillon, L. (2008). "Spatial Wage Disparities: Sorting Matters!" *Journal of Urban Economics*, 63(2):723742

Combes, PP., Duranton, G., Gobillon, L. and Roux, S. (2010) Estimating Agglomeration Economies with History, Geology, and Worker Effects. In E.L. Glaeser (ed), *Agglomeration Economies*, University of Chicago Press: Chicago.

Combes, PP, Démurger, S., and Li, S. (2013). "Urbanisation and Migration Externalities in China". *Discussion Paper* 9352, Centre for Economic Policy Research.

Combes, PP and Gobillon, L. (2015). The Empirics of Aglomeration. In: G. Duranton, V. Henderson, J. Strange, (ed), *Handbook of Regional and Urban Economics*, vol. 5.

DCosta, S. and Overman, H. (2014). "The Urban Wage Growth Premium: Sorting or Learning?". *Regional Science and Urban Economics*, 48:168179.

Da Mata, D., Deichmann, U., Henderson, J., Lall, S. and Wang, H. (2007) "Determinants of City Growth in Brazil". *Journal of Urban Economics*, 62(2):25272

Davidson, R. and MacKinnon, J. (2010) "Wild Bootstrap Tests for IV Regression". *Journal of Business and Economic Statistics*, 28(1):128-144.

de la Roca, J. and Puga, D., (2012). "Learning by working in big cities". Discussion Paper 9243, Centre for Economic Policy Research.

Dominicis, L., Arbia, G. and de Groot, H. (2007) “The Spatial Distribution of Economic Activities in Italy”. Tinbergen Institute Discussion Paper. 2007-094/3.

Duranton, G. (2009) Are Cities Engines of Growth and Prosperity for Developing Countries. In M. Spence, P. Annez, and R. Buckley, (ed) *Urbanization and Growth*, the Commission Growth and Development, Washington, D.C.

Duranton, G. (2014) “Agglomeration Effects in Colombia”. Samuel Zell and Robert Lurie Real Estate Center Working Paper # 758, The Wharton School, University of Pennsylvania

Duranton, G. and Monastiriotis, V. (2002) “Mind the Gaps: The Evolution of Regional Earnings Inequalities in the UK 1982-1997”. *Journal of Regional Science*, 42(2):219256

Duranton, G. and Puga, D. (2004). Micro-foundations of Urban Agglomeration Economies. In: Henderson, J.V., Thisse, J.F. (ed), *Handbook of Regional and Urban Economics*, vol. 4. North-Holland, Amsterdam

Fujita, M., Krugman, P. and Venables, A. J. (1999) *The Spatial Economy*, MIT Press, Cambridge, MA

García, G. (2014). “Labor Informality: Choice or Sign of Segmentation? A Quantile Regression Approach at the Regional Level for Colombia”. *Archivos de Economía* 412, DNP, Bogotá-Colombia

Glaeser, E. and Maré, D (2001) “Cities and Skills”. *Journal of Labor Economics*, 19(2):316-342

Glaeser, E. and Resseger, M. (2010) “The Complementarity Between Cities and Skills”. *Journal of Regional Science*, 50(1):221244

Groot, S.P.T., de Groot, H.L.F., and Smit, M.J. (2014). “Regional Wage Differences in the Netherlands: Microevidence on agglomeration externalities”. *Journal of Regional Science*. 54(3)503523

Hansen, L. (1982) “Large Sample Properties of Generalized Method of Moments Estimators”. *Econometrica*, 50: 1029-1054

Harris, J. and Todaro, M. (1970) “Migration, Unemployment, and Development: A Two Sector Analysis”. *American Economic Review*, 107: 126142

Henderson, J. V. (1986) “Efficiency of Resource Usage and City Size”. *Journal of Urban Economics*, 19:4770

Hoover, E. (1948) The Location of Economic Activity. New York: McGraw-Hill

Jütting, J. and De Laiglesia, J. (2009) “Is Informal Normal? Towards More and Better Jobs in Developing Countries”. OECD Development Centre, 163 pages

Kleibergen, F. Paap, R. (2006) “Generalized Reduced Rank Tests using the Singular Value Decomposition”. *Journal of Econometrics*, 133(1): 97-126.

Lee Y.S. (2014). “Competition, Wage, and Agglomeration Diseconomy”. *International Regional Science Review*. doi:10.1177/0160017614542338

Lee, Y.J. and Zang, H. (1998) "Urbanization and Regional Productivity in Korean Manufacturing". *Urban Studies*, 35:208599

Melo, P. and Graham, D. (2009) "Agglomeration Economies and Labour Productivity: Evidence from Longitudinal Worker Data for GB's Travel-to-Work Areas". SERC Discussion Paper 31. SERC

Melo, P., Graham, D. and Noland, R. (2009) "A Meta-Analysis of Estimates of Urban Agglomeration Economies". *Regional Science and Urban Economics*, 39: 332-342

Mills, E. and Becker, C. (1986) *Studies in Indian Urban Development.* New York: Oxford University Press

Mion, G. and Naticchioni, P. (2009). "The Spatial Sorting and Matching of Skills and Firms". *Canadian Journal of Economics* 42(1):2855.

Moreno-Monroy, A. (2012) "Informality in Space: Understanding Agglomeration Economies During Economic Development". *Urban Studies*, 49(10): 2019-2030

Moreno-Monroy, A. and García, G. (2015) "Intra-metropolitan Clustering of Formal and Informal Manufacturing Activity: Evidence from Cali, Colombia". *Journal of Economic and Social Geographic*, Forthcoming.

Murray, M. (2006) "Avoiding Invalid Instruments and Coping with Weak Instruments". *The Journal of Economic Perspectives*, 20(4):111-132

Nicodemo, C. and García, G. (2015). "Job Search Channels, Neighborhood Effects, and Wages Inequality in Developing Countries: The Colombian Case". *Developing Economies*, 53(2):75-99

Openshaw, S. and Taylor, P. (1979) A Million or so Correlation Coefficients: Three Experiments on the Modifiable Area Unit Problem, in N. Wrigley (ed), *Statistical Applications in the Spatial Sciences*, London: Pion

Overman, H. and Venables, A. (2005) "Cities in the Developing World". *CEP Discussion Paper*, No 695, London

Perry, G., Maloney, W., Arias, O., Fajnzylber, P., Mason, A. and Saavedra-Chanduvi, J. (2007) Informality: Exit and Exclusion, The World Bank

Roback, J. (1982). "Wages, Rents, and the Quality of Life". *The Journal of Political Economy* 90:125778

Rosenthal, S. and Strange, W. (2004). Evidence on the Nature and Sources of Agglomeration Economies. In: Henderson, V., Thisse, J.F. (ed), *Handbook of Regional and Urban Economics*, vol. 4. North-Holland, Amsterdam

Rosenthal, S. and Strange, W. (2008a) "Agglomeration and Hours Worked". *The Review of Economics and Statistics*, 90(1): 105118

Rosenthal, S. and Strange, W. (2008b) "The Attenuation of Human Capital Spillovers". *Journal of Urban Economics*, 64(2):373389.

Royuela, V. and García, G. A. (2015) “Economic and Social Convergence in Colombia”. *Regional Studies*, 49(2):219-239

Shea, J. (1997) “Instrument Relevance in Multivariate Linear Models: A Simple Measure”. *The Review of Economics and Statistics*, 79(2) 348-352

Schneider F. and Enste, D. (2000) “Shadow Economies: Size, Causes and Consequences”. *Journal of Economic Literature*, 38: 77114

Schneider, F., Buehn, A. and Montenegro, C.E. (2010). “New Estimates for the Shadow Economies all Over the Word”. *International Economic Journal*, 24(4): 443-461.

Shukla, V (1996) *Urbanization and Economic Growth.* Delhi: New York: Oxford University Press

Stock, J. and Yogo, M. (2005) Testing for Weak Instruments in Linear IV Regression. In D. Stock, K. A. A. J. H. (ed) *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg.* Cambridge: Cambridge University Press.

Tabuchi, T. (1998). “Urban Agglomeration and Dispersion: A Synthesis of Alonso and Krugman”. *Journal of Urban Economics*, 44(3):333-351

Wheaton, W. and Lewis, M. (2002) “Urban Wages and Labor Market Agglomeration”. *Journal of Urban Economics*, 51(3):542562

Tables and Figures

Table 1. Summary statistics at individual level

	2008		2011		2014		All years	
	Mean	**Std. Dev**	**Mean**	**Std. Dev**	**Mean**	**Std. Dev**	**Mean**	**Std. Dev**
Monthly wage $	287.9	273.19	366.67	336.9	339.75	293.13	354.98	324.27
Male	0.57	0.49	0.56	0.50	0.55	0.50	0.56	0.50
Age	37.99	12.81	38.31	13.18	38.54	13.43	38.29	13.15
Years of education	9.53	4.56	8.93	5.26	9.96	4.52	9.56	4.65
Education by levels:								
Primary school	0.26	0.44	0.26	0.44	0.23	0.42	0.25	0.43
Middle school	0.19	0.39	0.16	0.37	0.18	0.38	0.18	0.38
High school	0.33	0.47	0.36	0.48	0.34	0.47	0.34	0.47
Technical or technological	0.08	0.27	0.08	0.27	0.14	0.34	0.10	0.30
University	0.13	0.33	0.13	0.34	0.12	0.32	0.12	0.32
Years in the current job	6.63	8.78	6.31	8.26	6.04	8.06	6.30	8.79
Informality	0.63	0.48	0.63	0.48	0.58	0.49	0.61	0.48
Region:								
Central	0.51	0.50	0.51	0.50	0.53	0.50	0.52	0.50
Oriental	0.11	0.31	0.11	0.31	0.11	0.31	0.11	0.31
Occident	0.17	0.38	0.17	0.37	0.17	0.37	0.17	0.37
Caribbean	0.18	0.38	0.18	0.38	0.17	0.37	0.18	0.38
Orinoco	0.03	0.16	0.02	0.16	0.03	0.16	0.02	0.16
Occupation:								
Professional	0.05	0.22	0.06	0.23	0.07	0.25	0.06	0.23
Managers	0.11	0.31	0.11	0.31	0.10	0.30	0.11	0.31
White collar	0.07	0.26	0.06	0.24	0.06	0.24	0.07	0.25
Low white collar	0.09	0.28	0.09	0.28	0.09	0.29	0.09	0.28
Sales employees	0.01	0.12	0.01	0.12	0.01	0.12	0.01	0.12
Blue collar	0.20	0.40	0.19	0.39	0.18	0.38	0.19	0.39
Low blue collar	0.06	0.24	0.07	0.25	0.07	0.26	0.07	0.25
Skilled service workers	0.08	0.27	0.09	0.28	0.10	0.30	0.09	0.28
Unskilled service workers	0.27	0.45	0.27	0.45	0.27	0.44	0.28	0.45
Agricultural workers	0.04	0.20	0.04	0.20	0.04	0.19	0.04	0.20
Sector:								
Agriculture	0.05	0.22	0.05	0.22	0.04	0.20	0.05	0.22
Industry	0.16	0.37	0.15	0.36	0.14	0.35	0.15	0.36
Building	0.06	0.23	0.07	0.25	0.07	0.25	0.07	0.25
Commerce and hotel	0.29	0.45	0.29	0.46	0.30	0.45	0.29	0.45
Transport and tel	0.10	0.30	0.10	0.30	0.10	0.30	0.09	0.29
Financial	0.09	0.29	0.10	0.29	0.10	0.31	0.10	0.30
Adm. Pub	0.09	0.29	0.09	0.28	0.09	0.28	0.09	0.28
Service	0.15	0.36	0.15	0.35	0.15	0.36	0.15	0.36

Note: All data are weighted using person sampling weights from GEIH to be representative.

Table 2. Summary statistics at municipal level

		Municip.	p25	Median	p75	Mean	Std. Dev	Min	Max
Number of workers		568	5,200	8,879	13,905	26,574	156,488	341	3,399,830
Municipal area (Km2)		568	161.3	358.6	842.2	765.9	1,529.6	15.4	17,641.70
Altitude (m.a.s.l.)		568	123.0	1,026.5	1,786.5	1,138	1,369.2	2	25,221
Distance to the capital	**in Km**	537	45.5	87.1	137.0	104.2	81.4	0	452
city of the department	**in minutes**	537	56	102	165	124.4	99.9	0	789

Note: Number of workers using an average over the period 2008-2014. The distance to the capital city of the department in kilometers and time is calculated by Google map.

Table 3. Average wages between formal and informal employees in select municipalities

Sector	Municipalities	Monthly wages ($)			
		2008	2011	2014	All years
Formal	Bogotá, Medellín, Cali	433.9	555.1	490.9	492.0
	Less than 5,000 inhabitants	426.4	549.6	506.2	499.1
Informal	Bogotá, Medellín, Cali	253.1	303.8	269.5	272.4
	Less than 5,000 inhabitants	146.2	177.9	182.5	168.4

Note: All data are weighted using person sampling weights from GEIH to be representative. All differences in means between groups of municipalities for the formal sector and the informal sector are significant at 1%.

Table 4. Agglomeration effects, baseline model with informality (OLS)

Dependent variable: log monthly wage

	Only pop. density (1)	Only pop. density (2)	Indiv. charac. (3)	Sector occup. (4)	Geog var. (5)	Market access 1 (6)	Market access 2 (7)	Non lineal (8)	Educ. effects 1 (9)	Educ. effects 2 (10)
Log employment density	0.046***	0.050***	0.045***	0.049***	0.055***	0.045***	0.045***	0.098	0.065***	0.066***
	(0.0103)	(0.0088)	(0.0094)	(0.0092)	(0.0073)	(0.0101)	(0.0100)	(0.0804)	(0.0096)	(0.0095)
Formal x Log emp den		-0.108***	-0.067***	-0.068***	-0.067***	-0.068***	-0.068***	-0.108***	-0.067***	-0.070***
		(0.0031)	(0.0022)	(0.0021)	(0.0021)	(0.0021)	(0.0021)	(0.0107)	(0.0021)	(0.0028)
Log employment density^2								0.004		
								(0.0043)		
Formal x Log emp den^2								-0.004***		
								(0.0011)		
Log dist Km to capital city						-0.015				
						(0.0487)				
Log dist Km to capital city^2						-0.001				
						(0.0081)				
Log dist time to capital city							0.006			
							(0.0470)			
Log dist time to capital city^2							-0.005			
							(0.0078)			
Educ x Log emp den									-0.002***	-0.002***
									(0.0002)	(0.0002)
Educ x Formal x Log emp den										0.0003***
										(0.0001)
Observations	1,920,678	1,920,678	1,914,957	1,913,815	1,901,513	1,905,428	1,905,428	1,913,815	1,913,815	1,913,815
Municipalities	568	568	568	548	537	537	568	568	568	568
R2	0.017	0.249	0.444	0.469	0.476	0.469	0.469	0.470	0.470	0.470

Note: Robust standard errors clustered at the municipality level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

All models include year dummy variables. In columns 3 to 10 individual characteristics included are: education indicators (primary, basic school, high school, technical or technological education, and university), gender, age and its squared, years in the current job and its squared. In columns 4 to 10 all models include occupation (10) and economic sector (8) indicators. Geographical characteristics in column 5 include five regional indicators (Central, Oriental, Occident, Caribbean and Orinoco), a water availability index, a soil erosion index and the log of altitude and its square. Column 6 uses the distance in Km to the capital city of the department as a measure of market access. Column 7 uses time distance to the capital city of the department as a measure of access.

Table 5. Agglomeration effects and informality (2SLS)
Dependent variable: log monthly wage

	Only pop. density (1)	Indiv. charac. (2)	Sector occup. (3)	Geog var. (4)	Market access 1 (5)	Market access 2 (6)	Non lineal (7)	Educ. effects 1 (8)	Educ. effects 2 (9)
Log employment density	0.034***	0.026**	0.030***	0.036***	0.028**	0.028**	0.024	0.049***	0.050***
	(0.0104)	(0.0120)	(0.0120)	(0.0100)	(0.0125)	(0.0125)	(0.1095)	(0.0126)	(0.0124)
Formal x Log emp den	-0.109***	-0.069***	-0.069***	-0.068***	-0.069***	-0.069***	-0.132***	-0.049***	-0.050***
	(0.0035)	(0.0026)	(0.0025)	(0.0025)	(0.0025)	(0.0025)	(0.0135)	(0.0025)	(0.0032)
Log employment density^2							0.001		
							(0.0062)		
Formal x Log emp den^2							-0.007***		
							(0.0015)		
Log dist. Km to capital city					-0.011				
					(0.0603)				
Log dist. Km to capital city^2					-0.008				
					(0.0105)				
Log dist. time to capital city					0.036				
					(0.0632)				
Log dist. time to capital city^2					-0.012				
					(0.0109)				
Educ x Log emp den								-0.002***	-0.002***
								(0.0002)	(0.0002)
Educ x Formal x Log emp den									0.0003**
									(0.0001)
Observations	1,820,006	1,814,561	1,813,482	1,805,276	1,809,712	1,809,712	1,813,482	1,813,482	1,813,482
Municipalities	390	390	390	385	380	380	390	390	390
R2	0.247	0.443	0.468	0.475	0.469	0.469	0.469	0.469	0.469
Instruments exogeneity									
Hansen J statistic	1.168	1.855	1.926	5.150	0.787	0.839	8.819	1.955	1.955
Chi-sq P-val	0.558	0.396	0.382	0.076	0.675	0.657	0.066	0.376	0.376
Instruments relevance									
1. First-stage statistics									
Shea partial R2									
Log employment density	0.803	0.804	0.802	0.778	0.983	0.819	0.505	0.864	0.869
Formal x Log emp den	0.987	0.984	0.982	0.985	0.819	0.983	0.780	0.984	0.986
Log pop density2							0.506		
Formal x Log pop den2							0.756		
Educ x Log pop den								0.954	0.959
Educ x Formal x Log pop den									0.987
2. Under-identification test									
Kleibergen-Paap rk LM stat	15.08	15.19	15.44	13.06	14.07	14.05	12.50	15.43	15.46
Chi-sq P-val	0.002	0.002	0.001	0.004	0.003	0.003	0.029	0.001	0.001
3. Weak identification test									
Kleibergen-Paap rk Wald F stat	49.49	50.34	51.39	59.86	48.99	48.82	7.81	41.07	34.27

Note: Robust standard errors clustered at the municipality level in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$
This table replicates Table 4 using 1912, 1918 and 1928 populations as instrument for contemporaneous working population to calculate the log density variable in all columns. The square of these instruments are employment used in column 7. In columns 8 and 9 we use the average of population in 1912, 1918 and 1928 in the calculation of the product of education and log density and/or formality variables.

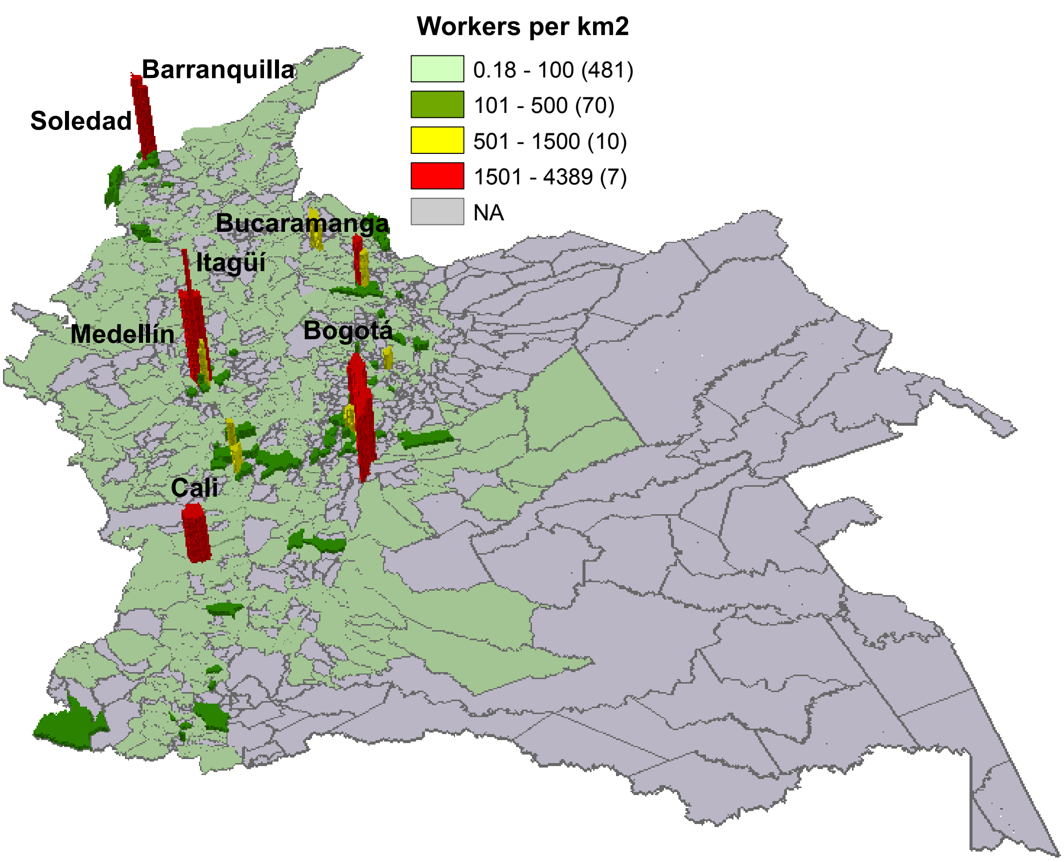
Table 6. Agglomeration effects and informality, robustness check

Dependent variable: log monthly wage

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Log employment density	0.027**	0.027**	0.029**	0.030**	0.025**	0.035***	0.030**
	(0.0128)	(0.0131)	(0.0125)	(0.0126)	(0.0118)	(0.0099)	(0.0120)
Formal x Log emp den	-0.069***	-0.069***	-0.069***	-0.069***	-0.069***	-0.070***	-0.069***
	(0.0025)	(0.0025)	(0.0024)	(0.0025)	(0.0025)	(0.0022)	(0.0025)
Instruments:							
1912 population	Y	N	N	Y	Y	Y	Y
1918 population	N	Y	N	N	Y	Y	Y
1928 population	N	N	Y	Y	Y	Y	Y
Observations	1,813,482	1,820,611	1,828,924	1,813,482	1,702,185	1,813,482	1,813,482
Municipalities	390	408	423	390	389	390	390
R2	0.468	0.467	0.468	0.468	0.470	0.469	0.46(
Instruments exogeneity							
Hansen J statistic	.	.	.	1.363	2.506	1.926	1.926
Chi-sq P-val	.	.	.	0.243	0.286	0.382	0.382
Instruments relevance							
1. First-stage statistics							
Shea partial R2							
Log employment density	0.750	0.981	0.795	0.983	0.695	0.982	0.802
Formal x Log emp den	0.980	0.765	0.983	0.796	0.431	0.982	0.802
2. Under-identification test							
Kleibergen-Paap rk LM stat	14.49	15.29	15.84	15.43	14.44	15.44	15.44
Chi-sq P-val	0.001	0.001	0.001	0.001	0.002	0.001	0.001
3. Weak identification test							
Kleibergen-Paap rk Wald F stat	59.96	58.29	74.97	48.34	56.29	51.39	51.39

Note: Robust standard errors clustered at the municipality level in parentheses. *** p<0.01, ** p<0.05, * p<0.1
All columns include year indicators, individual characteristics, occupation and economic sector indicators. 2SLS in columns 1-5, GMM in column 6, and LIML in column 7. Column 5 excludes observations corresponding to Bogotá.

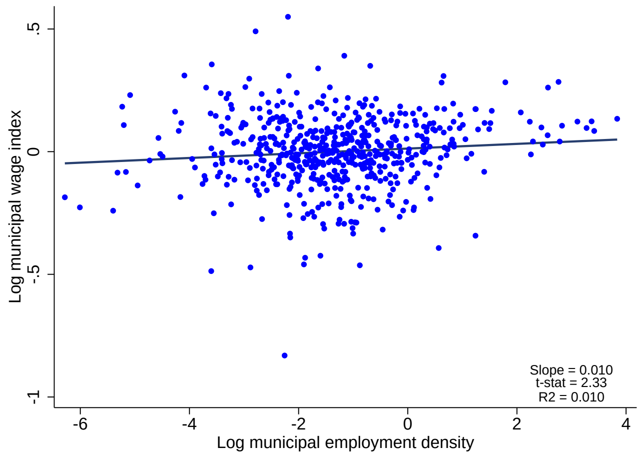
Figure 1. Employment density by municipality in Colombia



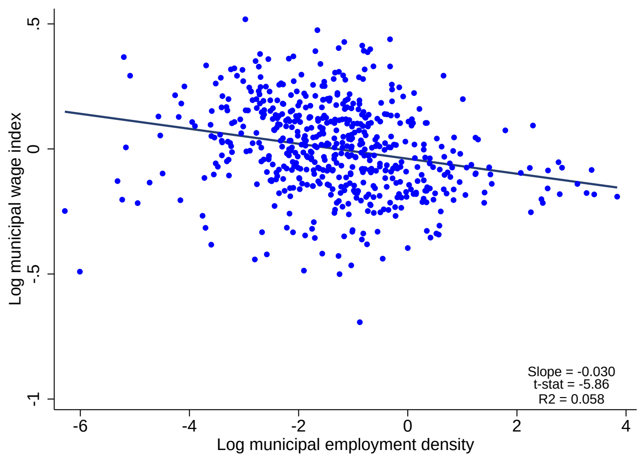
Note: Average employment density between 2008 and 2014.

Figure 2. Employment density and wages in Colombia

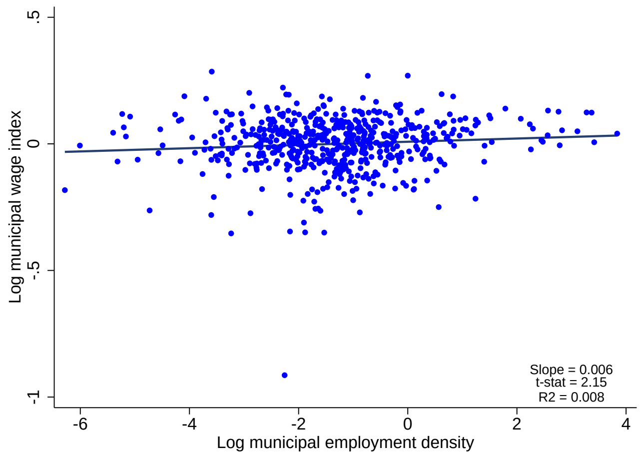
a) Total



b) Formal



c) Informal



Note: The vertical axis represents log municipal wages computed using 2008-2014 wage data after controlling for years effects, individual characteristics, occupation and economic sector. The horizontal axis represents log of average employment density between 2008 and 2014. There are 568 municipalities. All variables are centered around their mean.

10 Appendix

Table A1. Agglomeration effects, baseline model without informality (OLS)
Dependent variable: log monthly wage

	Only pop. density (1)	Indiv. charac. (2)	Sector occup. (3)	Geog var. (4)	Market access 1 (5)	Market access 2 (6)	Non lineal (7)	Educ. effects 1 (8)	Educ. effects 2 (9)
Log employment density	0.046***	0.039***	0.042***	0.043***	0.037***	0.037***	0.075	0.067***	0.061***
	(0.0104)	(0.0106)	(0.0102)	(0.0082)	(0.0113)	(0.0113)	(0.0921)	(0.0103)	(0.0094)
Log employment density^2							0.002		
							(0.0050)		
Log dist. Km to capital city					0.020				
					(0.0560)				
Log dist. Km to capital city^2					-0.008				
					(0.0093)				
Log dist. time to capital city					0.042				
					(0.0528)				
Log dist. time to capital city^2					-0.011				
					(0.0087)				
Educ x Log emp density								-0.003***	-0.003***
								(0.0002)	(0.0002)
Observations	1,920,678	1,914,957	1,913,815	1,901,513	1,905,428	1,913,815	1,913,815	1,913,815	1,893,885
Municipalities	568	568	568	548	537	537	568	568	521
R2	0.017	0.374	0.406	0.416	0.407	0.407	0.406	0.409	0.420

Note: Robust standard errors clustered at the municipality level in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$
All models include year dummy variables. In columns 2 to 9 individual characteristics included are: education indicators (primary, basic school, high school, technical or technological education, and university), gender, age and its squared, years in the current job and its squared. In columns 3 to 9 all models include occupation (10) and economic sector (8) indicators. Geographical characteristics in column 4 include five regional indicators (Central, Oriental, Occident, Caribbean and Orinoco), a water availability index, a soil erosion index and the log of altitude and its square. Column 5 uses the distance in Km to the capital city of the department as a measure of market access. Column 6 uses time distance to the capital city of the department as a measure of market access. Column 9 replicates column 8 but adds geographical characteristics and the time distance to the capital city of the department as controls.