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

Gustavo A. García
Department of Economics and *CEDE*
Universidad de los Andes
ga.garcia@uniandes.edu.co

Very preliminary draft November 2015

Abstract

This paper analyzes the relationship between agglomeration economies and wages in the context of a developing country, taking into account the market presence of the 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 the agglomeration returns. We find 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 16% 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 19% less in denser areas. Factors associated with work-spreading and des-amenities very common in big cities in developing countries could explain this lower agglomeration returns in the formal sector.

Keywords: Agglomeration gains; population density; Informality; work-spreading
JEL classification: *R12, J46, R23, J31*

1 Introduction

The pace and content of urbanization have a 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 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 has been previously demonstrated for developed countries (see for example, Ciccone and Hall (1996), Rosenthal and Strange (2008), Melo et al. (2009), Melo and Graham (2009)).

Despite 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 share of urban employment and growth, while the informal sector plays a large role in the economy and it is an important difference regarding developed economies (Schneider and Enste, 2000). According to Jütting and De Laiglesia's (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 accounted 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.

There is scarcity of the results on agglomeration economies in developing economies and those that exist are concerned with the formal sector and do not take into account the informal sector. 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. In a more recent analysis of city growth in Brazil, Da Mata et al. (2007) found that the urban elasticity, measuring urbanization economies as market potential, is 11%. In the Colombian context, the only work is by Duranton (2014), who 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.

Given the large differences in economic characteristics between the informal sector and the formal sector, such as productivity, profitability, and size, there are different point of views regarding the contribution and benefits of the informal sector 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. By contrast, Overman and Venables (2005), and Moreno-Monroy (2012) state that the informal sector also contributes 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 formal sector, but also supplies intermediate or final goods and services to this latter sector. As pointed out by Duranton (2009), between the formal and informal sector there are intense linkages, which suggests that agglomeration effects are generated within both sectors, with benefits that accrue to both. According to Overman and Venables (2005) there are two possibilities in which the existence of an informal sector can affect the benefits of agglomeration economies. On the one hand, the existence of an informal sector can drive up urban costs and crowd out the formal sector. On the other hand, the informal sector also contributes to agglomeration economies. In this sense, 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 population 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 data sources used in the analysis. Section 3 document statistically the relationship between agglomeration and wages taking into account the existence of the informal sector. Section 4 describes the applied methodology. Section 5 discusses the results, and Section 6 concludes.

2 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, 2014). 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.[1]

[1]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

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.

In addition to the GEIH we use the DANE's municipal population estimates for 2008, 2011 and 2014. These population estimates were calculated from the last census conducted in Colombia in 2005. These data offer information on population for the total of 1,119 municipalities of Colombia with a breakdown between the rural and urban area of each municipality.[2] Following Hoover (1948) and the large literature on agglomeration, we divided the total (urban and rural) municipality population by the geographic area of the municipality to calculate population density as a measurement of urban agglomeration.

We divide workers into formal and informal workers. The later 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 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 a measure of market access which is determined by the distance to Bogotá. 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]

of DANE's staff and with blinded access to sensible information.

[2]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, with a total of 1,119 municipalities in the country. Most of the this municipalities are organized around one main settlement, called as the head or urban area of the municipality and another peripheral area referred to as the rural area of the municipality (a more detailed characterization of Colombia can be found in Royuela and García (2005)).

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 population density of the municipalities. Municipalities have an average of roughly 70,000 people and range from 1,600 people to over seven million. Figure 1 shows the population density by municipality. We can note that Bogotá, Medellín, Itaguí, Cali, Bucaramanga and Barranquilla are the cities with the highest levels of urbanization, where Itaguí being the most densely populated city in Colombia with just over 12,000 people per Km2.

[Insert Figure 1 around here]

3 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 plots log wages against population 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 around 2%. Regarding, the formal and informal sectors, Figure 2 (b) and (c) show that while for the formal sector the density elasticity of wages is -3.1%, for the informal sector this elasticity is 1.2%, 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, 2008). 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 ability to absorb the knowledge, and the activities of small enterprises are more geared towards small local markets than towards generating input-output linkages (García and Moreno-Monroy, 2015).

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 earning shorter wages. This kind of work-spreading would imply the opposite sing on population density (Rosenthal and Strange, 2008). 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]

4 Estimation strategy

In this section we complement the results of the previous section by attempting to clarify the effects of agglomeration economies on formal and informal productivity, and if there are such effects, which sector achieves greater benefits. We begin by regressing individual worker wage rate as a measurement of labor productivity on population density. The wage equation employed for the estimate has the following structure:

$$\ln w_{i(t)} = \alpha_0 + \beta \ln density + X_{i(t)}\varphi + \pi_{oi(t)} + \sigma_{si(t)} + \eta_{ri(t)} + \delta_t + \epsilon_{i(t)} \tag{1}$$

where i identifies the worker, o refers to occupation, s refers to the economic sector, r 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. As mentioned, the dependent variable is the logarithm of the monthly wage.

Our measurement of urban agglomeration is the logarithm of the population density of the municipalities, $\ln density$, which is defined as the number of people per square kilometer in each municipality using an average of 2008, 2011 and 2014 population data. 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. 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 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 $\eta_{ri(t)}$. Altitude of each municipality also is included as an additional measure of geographical characteristics (see Tables 1 and 2 for a more detailed description of these variables).

Equation (1) 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, 2014). 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:

$$\ln w_{i(t)} = \alpha_0 + \beta \ln density + \phi \ln density * formal_{i(t)} + X_{i(t)}\varphi + \pi_{oi(t)} + \sigma_{si(t)} + \eta_{ri(t)} + \delta_t + \epsilon_{i(t)} \quad (2)$$

where $formal$ is a dummy variable equal to 1 if worker is formal and 0 if is informal, according to definition of informality presented in section 2.

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 population 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

[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).

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 population 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 (2014), 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 the next section.

5 Results

5.1 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.[4] 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

[4] We follow the formula of Combes and Gobillon (2014): $2^{\beta}-1$, where β is the elasticity of productivity with respect to density.

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 -6.8% (-12.3%+5.5%) and significant which suggests that formal workers in a city twice as dense have around 5% less productivity. In contrast, the elasticity among informal workers is 5.5% and highly significant, indicating that this kind of workers in a city twice as dense have around 4% greater productivity. This difference 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 -2.6%, while in the informal sector the elasticity presents a slight decreasing and is allocated around 5%. This suggests that in the formal sector more than half the relationship between wages and population 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 -2.6% and 5% for the formal sectors and the informal sector, respectively, are quantitatively important. Comparing a small municipality with a density of 50 persons per Km2 to Bogotá with a density around 5000 persons per Km2, these elasticities imply that in the formal sector workers in denser cities will earn 11% less than in less dense cities, whereas in the informal sector the wage difference is 26% 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.[5]

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 5.7%, and in the formal sector this coefficient increases to -1.9%. 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

[5]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.

Oronoco regions of Colombia relative to the Central region, and in the Caribbean region wages are lower than in the Central region.

Column 6 duplicates column 4 but adds log market access and its square. Market access is calculated by the Euclidian distance between the municipality under consideration and the country's capital, Bogotá. The results show that these variables are not significant but their inclusion 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 decreases to -3.4% and in the informal sector this coefficient drops to 4.1%. However, these results should be viewed with caution because our measure of market access, which is a simple Euclidian distance, could be a poor indicator of the true travel costs between municipalities, in particular in a mountainous country like Colombia.

In column 7 we re-estimate the specification of column 4 but we use the population of the urban part of a municipality instead of the total (urban and rural) municipality population to calculate the density variable. We can note that the coefficient on log density in the formal sector change marginally to -2.7%, whereas in the informal sector remains equal to 5%. It is consistent with the fact that for many cities, in particular less dense cities, the total population coincide with the urban population with which the dispersion of the dependent variable, wages, and the main explanatory variable, density, remain unchanged.

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 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 that formal and informal workers less educated obtain higher agglomeration returns, with formal workers less educated obtaining less returns in greater cities.

Nevertheless, these negative effects are 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, 2008, 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.

5.2 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 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 -4.6% and 3.2%, respectively, instead of -2.6% and 5% 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 population 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 population density to Bogotá, which presents a high population density, these elasticities suggest that formal workers in denser areas will earn 19% less than in less dense areas, and in the informal sector wages will be 16% 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) makes no change to the coefficients on city density for the formal and informal sector in column 3. Column 5 estimates lower coefficients, although, as mentioned, this specification could suffer by error measurement in the market access variable. In column 6 when we use urban population of municipality instead of the total municipality population to calculate agglomeration variable, the results remain the same as those found in column 3, in keeping with OLS results. 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.

6 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. The paper presents evidence that among formal workers, agglomeration tends to decrease wages. This may be due to the constraints in the creation of formal jobs in the economy which result in spread out wages over a large number of individuals and diminished individual wages. Among informal workers, the pattern is different, with agglomeration increasing wages. 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.

This paper also contributes to the literature on agglomeration economies related to agglomeration also encourage hard work (Rosenthal and Strange, 2008), 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 benefit of these productive effects manifested in higher wages in denser cities.

7 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.

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

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

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.

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

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. and Gobillon, L. (2014) “The Empirics of Agglomeration Economies”. PSE Working Papers n2014-31. 2014. ¡halshs-01071761¿

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.

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

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

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

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

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.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

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

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

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

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

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

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.

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
(568 municipalities)

	Year	p25	Median	p75	Mean	Std. Dev	Min	Max
	2008	9,948	18,963	37,479	66,670	339,475	1,735	7,155,052
Total population	2011	9,993	19,282	38,881	69,198	353,684	1,623	7,467,804
	2014	9,725	19,863	41,336	71,751	367,733	1,521	7,776,845
	2008	2,787	7,026	20,930	53,683	337,694	309	7,139,232
Urban population	2011	2,925	7,454	21,952	56,058	351,925	326	7,451,718
	2014	2,924	7,709	23,017	58,419	365,981	342	7,760,451
Municipal area (km2)		163.4	360.8	842.2	766.6	1,529.40	15.4	17,641.70
Altitude (m.a.s.l.)		123.0	1,026.50	1,786.50	1,138	1,369.20	2	25,221
Distance to Bogotá (km)		183.8	291.7	467.6	329	187.6	0	902.3

Source: DANE

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)

	Only pop. density (1)	Only pop. density (2)	Indiv. charac. (3)	Sector occup. (4)	Geog var. (5)	Market access (6)	Mun. pop. (7)	Non lineal (8)	Educ. effects 1 (9)	Educ. effects 2 (10)
Log pop density	0.048***	0.055***	0.048***	0.052***	0.057***	0.041***	0.050***	0.070	0.069***	0.070***
	(0.0112)	(0.009)	(0.0099)	(0.0098)	(0.0086)	(0.0076)	(0.0087)	(0.0789)	(0.0102)	(0.0101)
Formal x Log pop den		-0.123***	-0.076***	-0.077***	-0.076***	-0.075***	-0.077***	-0.127***	-0.076***	-0.079***
		(0.0037)	(0.0026)	(0.0025)	(0.0024)	(0.0025)	(0.0025)	(0.0116)	(0.0025)	(0.0032)
Log pop density^2								0.002		
								(0.0048)		
Formal x Log pop den^2								-0.006***		
								(0.0013)		
Log dist to Bogotá						0.010				
						(0.0384)				
Log dist to Bogotá^2						-0.007				
						(0.0054)				
Educ x Log pop den									-0.002***	-0.002***
									(0.0002)	(0.0002)
Educ x Formal x Log pop den										0.0003**
										(0.0001)
Observations	1,920,678	1,920,678	1,914,957	1,913,815	1,913,815	1,913,815	1,913,815	1,913,815	1,913,815	1,913,815
Municipalities	568	568	568	568	568	568	568	568	568	568
R2	0.017	0.248	0.443	0.468	0.475	0.472	0.469	0.470	0.469	0.469

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 to Bogotá as a measure of market access. Column 7 replicates column 4 using urban municipal population as a measure of agglomeration.

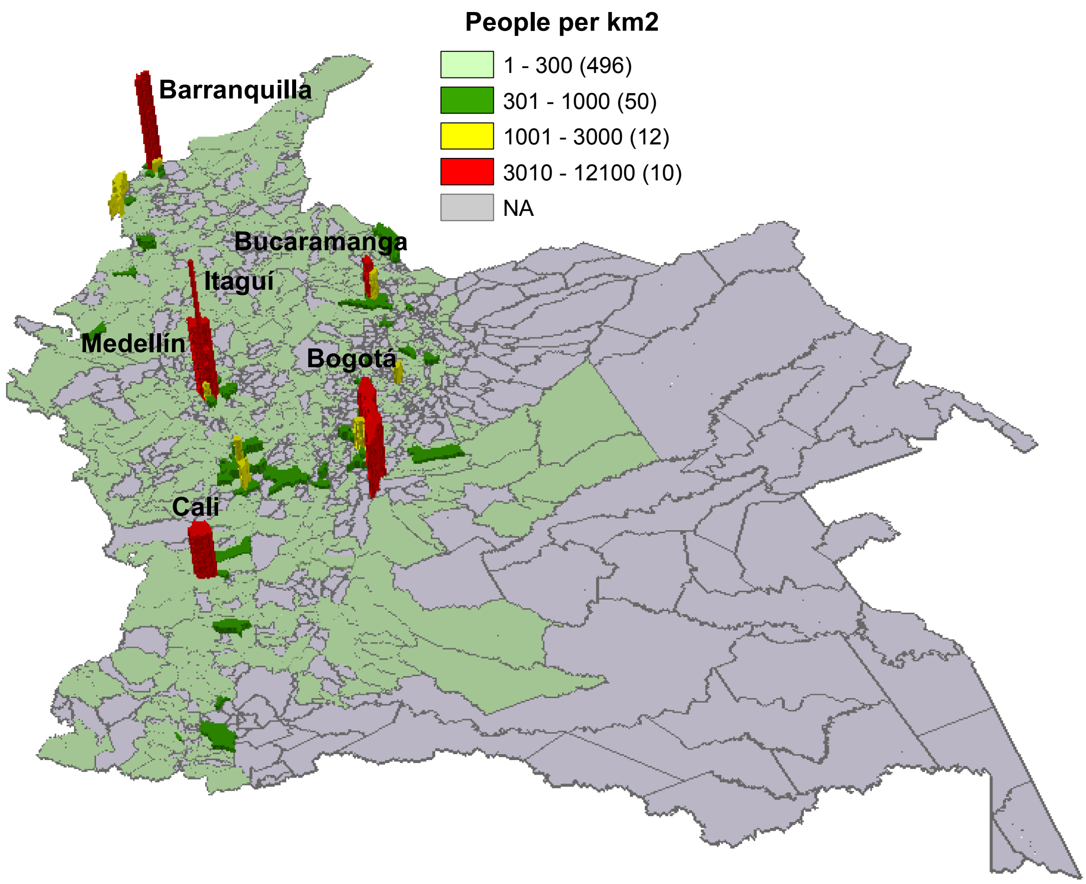
Table 5. Agglomeration effects and informality (2SLS)

	Only pop. density (1)	Indiv. charac. (2)	Sector occup. (3)	Geog var. (4)	Market access (5)	Mun. pop. (6)	Non lineal (7)	Educ. effects 1 (8)	Educ. effects 2 (9)
Log pop density	0.039***	0.030**	0.034***	0.038***	0.023**	0.033***	0.043	0.055***	0.055***
	(0.0110)	(0.0128)	(0.0128)	(0.0093)	(0.0114)	(0.0124)	(0.1076)	(0.0135)	(0.0133)
Formal x Log pop den	-0.124***	-0.078***	-0.079***	-0.078***	-0.077***	-0.079***	-0.153***	-0.078***	-0.081***
	(0.0043)	(0.0031)	(0.0030)	(0.0030)	(0.0030)	(0.0030)	(0.0146)	(0.0030)	(0.0037)
Log pop density2							0.002		
							(0.0071)		
Formal x Log pop den2							-0.009***		
							(0.0019)		
Log dist to Bogotá					-0.001				
					(0.037)				
Log dist to Bogotá^2					-0.006				
					(0.0052)				
Educ x Log pop den								-0.002***	-0.002***
								(0.0002)	(0.0002)
Educ x Formal x Log pop den									0.0003**
									(0.0001)
Observations	1,820,006	1,814,561	1,813,482	1,813,482	1,813,482	1,813,482	1,813,482	1,813,482	1,813,482
Municipalities	390	390	390	390	390	390	390	390	390
R2	0.246	0.443	0.468	0.475	0.471	0.468	0.469	0.469	0.470
Instruments exogeneity									
Hansen J statistic	1.111	1.77	1.829	5.15	4.776	1.638	8.515	1.849	1.884
Chi-sq P-val	0.574	0.413	0.401	0.076	0.092	0.441	0.074	0.397	0.400
Instruments relevance									
1. First-stage statistics									
Shea partial R2									
Log pop density	0.813	0.982	0.98	0.796	0.982	0.776	0.514	0.87	0.874
Formal x Log pop den	0.985	0.814	0.813	0.982	0.824	0.980	0.784	0.982	0.984
Log pop density2							0.509		
Formal x Log pop den2							0.757		
Educ x Log pop den								0.947	0.952
Educ x Formal x Log pop den									0.984
2. Under-identification test									
Kleibergen-Paap rk LM stat	15.28	15.38	15.61	9.416	14.13	15.82	12.36	15.61	15.68
Chi-sq P-val	0.002	0.001	0.001	0.024	0.003	0.001	0.030	0.001	0.001
3. Weak identification test									
Kleibergen-Paap rk Wald F stat	59.06	60.07	61.48	59.66	63.39	60.84	14.95	49.16	41.29

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 population to calculate the log population density variable in all columns. The square of these instruments are 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.

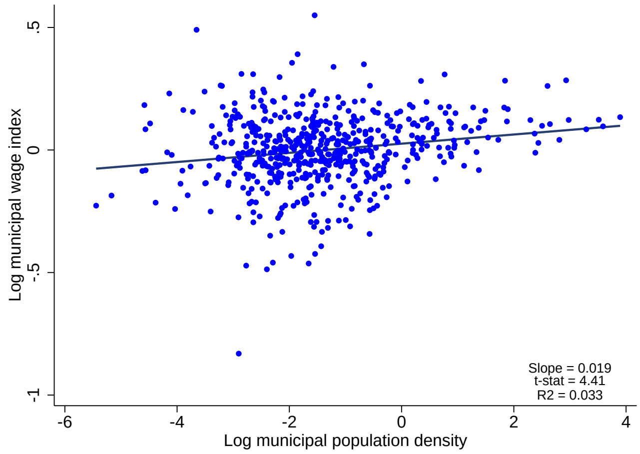
Figure 1. Population density by municipality in Colombia, 2014



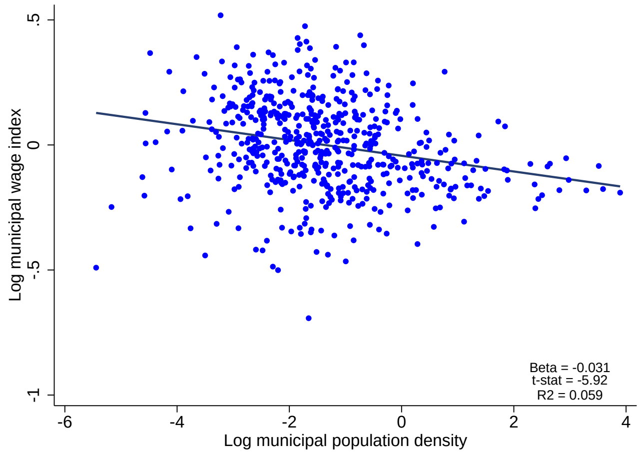
Source: DANE

Figure 2. Population density and wages in Colombia

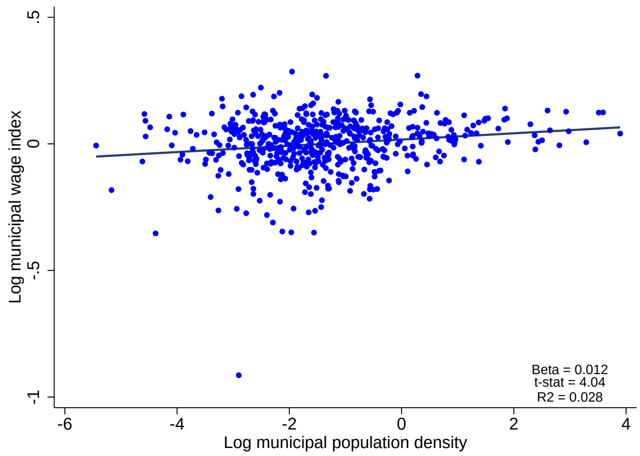
a) Total



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 2014 total population. There are 568 municipalities. All variables are centered around their mean.

8 Appendix

Table A1. Agglomeration effects, baseline model without informality (OLS)

	Only pop. density (1)	Indiv. charac. (2)	Sector occup. (3)	Geog var. (4)	Market access (5)	Mun. pop. (6)	Non lineal (7)	Educ. effects 1 (8)	Educ. effects 2 (9)
Log pop density	0.048***	0.040***	0.043***	0.043***	0.030***	0.041***	0.036	0.070***	0.069***
	(0.0112)	(0.0115)	(0.0111)	(0.0094)	(0.0081)	(0.0100)	(0.0952)	(0.0112)	(0.0088)
Log pop density^2							-0.0001		
							(0.0059)		
Log dist to Bogotá					0.054				
					(0.0387)				
Log dist to Bogotá^2					-0.013**				
					(0.0054)				
Educ x Log pop den								-0.003***	-0.003***
								(0.0002)	(0.0002)
Observations	1,920,678	1,914,957	1,913,815	1,913,815	1,913,815	1,913,815	1,913,815	1,913,815	1,913,815
Municipalities	568	568	568	568	568	568	568	568	568
R2	0.017	0.373	0.406	0.415	0.412	0.406	0.406	0.409	0.42

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, 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 to Bogotá as a measure of market access. Column 6 replicates urban column 3 using municipal population as a measure of agglomeration. Column 9 replicates column 8 but adds geographical characteristics as controls.