The Value of Urban Locations: Location Wage Premia in Latin America and the Caribbean [*]
Preliminary and Incomplete.

Luis E. Quintero[†] and Mark Roberts[‡]

July 12, 2017

There is a large and extensive literature examining the strength of agglomeration economies and, more generally, the determinants of spatial variations in productivity for developed countries. However, but that the corresponding literature for developing countries is comparatively scant. This paper contributes to the existing gap by providing estimates for location wage premia and agglomeration economies for 16 countries in the Latin America region. While two of the countries in our sample - Brazil and Colombia - have already been considered by the literature - the remaining 14 countries have not, to our knowledge, been previously analyzed. We generate these estimates using a harmonized data set which contains information on both the nominal wages and characteristics of individual workers, and of the characteristics of the locations in which the workers live, and follow initially a standard empirical specification to all countries. By generating estimates for 16 developing countries, we provide for a significant increase of knowledge in the strength of agglomeration economies at the extensive margin. In addition to examining the strength of agglomeration economies, we also examine the roles of human capital externalities and market access in explaining sub-national productivity variations for our 16 study countries. We find that city-wide human capital externalities appear much stronger than agglomeration economies in explaining productivity variation in all of the considered countries. There is considerable heterogeneity in the estimated strength of human capital externalities across countries, which could be a reflection of country differences in educational quality[1].

1 Introduction

There exists an extensive empirical literature which documents the existence of significant agglomeration economies (see Combes and Gobillon (2015) for a detailed review of this literature). However, as has been noted by, for example, Overman and Venables (2005), Desmet and Rossi-Hansberg (2013), and Duranton (2015) this literature is largely confined to developed countries . As such, there exists an important empirical blind spot with regards to the strength of the basic

[*]Funding for this work from the World Bank is gratefully acknowledged. The views expressed here are those of the authors and do not necessarily represent the views of The Word Bank.

[†]Carey Business School. Johns Hopkins University.

[‡]Social, Urban, Rural and Resilience (SURR) Global Practice, The World Bank.

[1]We are grateful to Gilles Duranton as well as participants at the 2017 Authors' Workshop for the Flagship Study on Cities, Productivity and Growth in Latin America and the Caribbean (LAC). We thank Joao Jatene and Jane Park for excellent research assistance, and Julia Branson, Andrew Campbell-Sutton, Graeme M. Hornby, Duncan D. Hornby and Chris Hill at the GeoData Institute in the University of Southampton for excellent data work.

forces which help to govern the productivity and growth of cities in developing countries. This is at a time when there are good reasons to suspect that the strength of agglomeration economies may differ significantly for developing countries on account of, inter alia, differences in their economic structures, levels of institutional development, and infrastructure stocks.

In response to the above knowledge gap, there has been a recent effort in the literature to generate estimates of the strength of agglomeration economies for several developing countries. In particular, Duranton (2016) presents such estimates for Colombia, while Chauvin et al. (2017) do likewise for Brazil, China and India. In both cases, in order to allow for the comparability of their results, they make use of empirical specifications that have previously been applied in the literature to developed countries.

This paper contributes to the above effort to fill the knowledge gap on agglomeration economies for developing economies by providing estimates of their strength for 16 countries in the Latin America region. While, as indicated above, two of these countries - Brazil and Colombia - have already been considered by Chauvin et al. (2017) and Duranton (2016) respectively - the remaining 14 countries have not, to our knowledge, been previously analyzed in the literature. We generate these estimates using a harmonized data set which contains information on both the nominal wages and characteristics of individual workers, and of the characteristics of the locations in which the workers live. This data set has been constructed from successive rounds of household surveys extracted from the Socio-economic Database for Latin America and the Caribbean (SEDLAC). To ensure the comparability of results both across the countries we study and with those available elsewhere in the literature for other countries, we follow the two papers cited in previously by applying a standard empirical specification to all countries. By generating estimates for 16 developing countries, we provide for a considerable increase of knowledge in the strength of agglomeration economies at the extensive margin.

In addition to examining the strength of agglomeration economies, we also examine the roles of human capital externalities and market access in explaining sub-national productivity variations for our 16 study countries. By controlling for key individual worker characteristics in our estimation, we aim to minimize bias due to sorting effects, while maximizing the number of countries analyzed.

2 Empirical Strategy

In a first stage we estimate location premiums for the smallest identified administrative units (municipalities) in our data set, through a fixed effects estimation:

$$\ln(W_{i,l(i),t}) = \alpha_l L_{i,l(i),t} + \delta_t + \epsilon_{i,l(i),t} \tag{1}$$

Additionally, to control for sorting, we include a vector of worker characteristics:

$$\ln(W_{i,l(i),t}) = \alpha_l L_{i,l(i),t} + \beta \overrightarrow{Worker_{l,t}} + \delta_t + \epsilon_{i,l(i),t} \tag{2}$$

These regressions are run by country. $W_{i,l(i),t}$ is the wage of worker i in location l in year t.

In a second stage, we use the estimated location premiums as the dependent variable and analyze their determinants, while controlling for local amenities. We run the following regression:

$$\hat{\alpha}_l = \overrightarrow{A_l}\theta + \tau_l + \mu_l \tag{3}$$

Where $\hat{\alpha}_l$ are the estimated location premiums from stage 1. $\overrightarrow{A_l}$ is a vector of characteristics: population density, city level human capital (measured either by average years of education completed in the working-age population or by the share of the population with tertiary education degrees), market access (access to markets within each country excluding the local market of the municipality itself), density of local roads, and amenities (annual average air temperature, annual total precipitation, and a measure of terrain ruggedness).

Currently, we are exploring the use of IVs for key explanatory variables to control for possible endogeneity resulting from reverse causation and omitted variables. For instance, we are exploring instrumenting current population density with pre-colonial population density estimated by the number of indigenous people per sq. km before the arrival of Columbus (Maloney and Valencia Caicedo, 2016).

3 Data

For the 1st stage (equation 2) we use the Socio-Economic Database for Latin America and the Caribbean (SEDLAC). SEDLAC is a database of socio-economic statistics constructed from microdata of the Latin American and Caribbean (LAC) household surveys, developed by CEDLAS (Universidad Nacional de La Plata) and The World Bank's LAC poverty group (LCSPP). All statistics are computed from micro-data of household surveys by routines documented in Stata do files available from the authors upon request. Previously, data availability has been an important obstacle to carrying out large scale work on agglomeration economies for developing countries. In particular, surveys are not uniform across LAC countries. Comparability is, therefore, an issue of great concern. SEDLAC harmonizes the raw data to make it comparable, to the largest possible extent, across countries. This makes this dataset optimal for our endeavor. We pool cross-sections of survey data from 2000 onwards and deflate monetary values to 2000 US dollars. Due to matching issues, we cannot use SEDLAC for Brazil and instead use the Census micro-data sub-sample available in IPUMS. For Brazil, we pool data from the 2000 and 2010 Population Censuses and harmonize with SEDLAC as far as possible.

For most countries, the locations that our data set captures cover more than 80% of all population as shown in table 1. Table 2 shows summary statistics that describe the population we are using in 1st stage estimation. The estimation, as well as the presented statistics, are constrained to working-age population. As expected when working with developing countries, the percentage of workers with tertiary education is rather small (compare to 35 percent as the average of OECD countries and 43 percent for the US, as reported by the OECD for 2017). Similarly, the working

Country (ISO)	**Admin. Unit** (Division level)	**Survey years**	**1st-stage** # obs.	**2nd-stage** # locations	**% national population covered**
Guatemala	Departmento (1)	2006, 2011, 2014	58,030	22	100
Chile	Comuna (3)	2000, 2003, 2006, 2009, 2011, 2013	405,221	328	99.4
Honduras	Municipio (2)	2004 - 2011, 2013	43,261	275	97.8
Peru	Distrito (3)	2000 - 2014	459,915	1,428	96.3
Costa Rica	Distrito (3)	2002 - 2006, 2008 - 2010	117,517	401	94
El Salvador	Municipio (2)	2014	27,117	220	93.2
Nicaragua	Municipio (2)	2001, 2005	24,730	116	90.2
Ecuador	Parroquia (3)	2005 - 2012, 2014	237,801	637	88.4
Uruguay	Municipio (2)	2005	15,915	13	87.6
Bolivia	Provincia (2)	2006, 2012	14,874	73	83.9
Dominican Republic	Municipio (2)	2000 -2014	131,608	207	80.8
Brazil	Municipio (2)	2000	1,809,596	1,488	80.7
Panama	Provincia (1)	2003 - 2008, 2010 - 2013	186,956	9	80
Argentina	Aglomerado (1)	2000 - 2011, 2013	225,261	13	75.9
Mexico	Municipio (2)	2000, 2002, 2008, 2010	94,105	430	64.6
Colombia	Municipio (2)	2008, 2009, 2010	231,349	212	60.9
Total			4,083,256	5,872	77.9

Table 1: Summary of data

population is younger (compare to an average of 41 years in the US for 2014 according to the Bureau of Labor Statistics). As a robustness test, we perform estimations with a narrow sample where we constraint our sample to only male respondents that work in the private sector.

For the 2nd stage (equation 3) explanatory variables we use a LAC Geospatial Database constructed in collaboration with University of Southampton for this project. This dataset includes geo-coded data for a large number of variables, which have been constructed using consistent methods across countries. The specific variables in the vector of characteristics are: Population density (population per km^2 of admin unit), aggregate stock of human capital, market access (sum of all other locations in a country divided by Open Street map time distance), road density, terrain ruggedness index (elevation variation calculated as in Nunn and Puga (2012)), air temperature, and precipitation. Information on this dataset is found in Branson et al. (2016). Appendix A provides further details of the specific variables used in the second stage. Table 3 provides summary statistics [2] . We aim to use information for the same years that we have population data for (2000, 2010). However, in many cases, we only have data available for 2000, 2014. In this case we use the information for the latter year to match 2010 population information. The correspondent administrative unit was used to aggregate the data when necessary to the same level of the population data. Analysis aggregating estimation at the urban extents will be performed as a robustness test.

[2]Population density for Argentina seems significantly lower than for the rest of the sample. This is explained partially by the fact that we are using a different administrative level (aglomerados) for Argentina, due to data constraints.

Country	**# of observations**	**% Males**	**% Married**	**Mean age**	**Mean years of schooling**	**% Workers tertiary education**
ARG	225,261	59.5	60.6	38.5	10.7	17.4
BOL	14,874	60.2	66.6	38.2	9.5	17.1
BRA	1,809,596	57.5	56.6	33	7.8	8.3
CHL	405,221	65.1	61.9	39.7	11.2	13
COL	231,349	56.2	57.7	37.7	8.8	15.6
CRI	117,517	66.9	59.8	36.2	8.9	10.9
DOM	131,608	66	57.7	36.8	8.8	11.1
ECU	237,801	64.1	54.2	38.8	9.4	12.6
GTM	58,030	66.8	63.7	34.9	6	2.7
HND	43,261	63.1	58.9	35.9	6.6	5.3
MEX	94,105	62.4	61.9	36.4	9.4	13.6
NIC	24,730	63	60.3	35.5	6.6	8.1
PAN	186,956	64.7	61.8	38	10.7	11.2
PER	459,915	60.7	60.7	37.8	9.7	17.6
SLV	27,117	59.3	58.3	37	8.3	5.3
URY	15,915	54.2	61	39.9	10.6	13.6
All	4,083,256	60.2	58.4	35.8	8.6	11.3

Table 2: Summary Statistics 1st stage

Variables	Statistic	All	ARG	BOL	BRA	CHL	COL	CRI	DOM	ECU
Population	Mean	473.2	19	32.2	341.7	434.6	401.8	1,357.90	469	181.3
density	Median	55.7	11.8	12.3	61.4	28.7	69.8	171.1	115.9	59.6
(people/	Max	25,821	68	356	13,393	8,652	9,039	16,355	11,188	4,801
km2)	Min	0	3.9	0.6	0.2	0.1	2.6	3.3	6.5	1.1
	S.D.	1,737	20	63	1,139	1,375	1,092	2,564	1,292	426
Average	Mean	6.8	10.6	6.8	5.4	9.5	7.2	7.7	6.9	6.9
years of	Median	6.7	10.6	7	5.6	9.4	6.9	7.5	6.9	6.8
schooling	Max	14.2	11.1	11.6	9.8	13.5	10.9	14.2	11.3	11.8
(years)	Min	1.4	9.9	1.5	1.8	5.4	3.6	3.1	3	2.8
	S.D.	2	0.4	2.3	1.5	1.1	1.4	1.8	1.4	1.6
Share of	Mean	5.3	11.5	6.8	4.4	7.1	5.8	6.8	4.4	2.9
workers	Median	3.5	11.3	4.8	3.6	5.6	4.8	4	3.9	1.7
w/ tertiary	Max	55.8	14.3	24.1	24.3	37.5	23.2	55.8	18.7	22.4
education	Min	0	9.2	0	0.3	0	0	0	0	0
(%)	S.D.	5.8	1.7	6.4	3.2	5	4.3	7.6	3.3	3.6
Market	Mean	14.2	18.4	13.2	14.5	14.1	15.6	15.5	14.3	13.9
access	Median	13.6	19.6	12.7	14	13.7	14.7	15.4	14.2	13.6
index	Max	36.2	25.5	19.5	36.2	22.9	26.5	22	22.3	25
(ln)	Min	8.1	11.3	9.1	9.1	8.1	11.5	11.1	11.8	10.9
	S.D.	2.5	5.7	2.7	2.6	2.9	2.9	2.5	1.8	1.7
Road	Mean	0.2	0.1	0.1	0.1	0.1	0.1	0.7	0.2	0.1
density	Median	0.1	0.1	0.1	0.1	0.1	0.1	0.6	0.1	0.1
(km/km2)	Max	3.1	0.1	0.1	0.4	0.8	0.4	3.1	0.9	0.5
	Min	0	0	0	0	0	0	0	0	0
	S.D.	0.2	0	0	0	0.1	0.1	0.5	0.1	0.1

Variables	Statistic	GTM	HND	MEX	NIC	PAN	PER	SLV	URY
Population	Mean	282.3	89.4	1,024.40	128.6	53.5	478.5	500.9	193.1
density	Median	167.9	66	128.9	65.3	50.2	24.5	213.3	8.3
(people/	Max	1,531	835	19,743	1,584	164	25,821	6,903	2,289
km2)	Min	20.6	5.6	0.3	7.1	4.6	0	30.8	4.9
	S.D.	319	102	2,700	202	46	2,346	996	630
Average	Mean	5.2	5.2	7.8	5.1	8.9	7.4	7	9.2
years of	Median	5.1	5.2	7.9	5.2	8.9	7.2	6.7	9
schooling	Max	7.9	9.1	13.4	9.1	10.8	13.2	12.7	10.4
(years)	Min	3.8	1.4	2.4	1.5	6.7	1.6	2.8	8.7
	S.D.	0.9	1.2	1.9	1.8	1.2	2	1.6	0.5
Share of	Mean	1.4	1.1	6.9	3.3	6.6	7.3	2.4	6.2
workers	Median	1	0.6	5.3	1.9	6.8	4.3	1.4	5.8
w/ tertiary	Max	6.5	8.2	44.9	19.3	9.8	51.4	25.7	10.9
education	Min	0.5	0	0	0	3.3	0	0	1.6
(%)	S.D.	1.3	1.5	6.2	3.8	2.2	8.1	3.3	2.2
Market	Mean	13.7	13.4	15.3	13	12.9	13.3	13.3	12.6
access	Median	13.1	13.2	14.7	12.6	11.6	12.7	13.1	12.1
index	Max	18.8	19.7	24.6	18.1	22.4	28.5	18.7	21.5
(ln)	Min	10.4	11.3	9.8	9.8	9.9	8.9	9.8	9.2
	S.D.	1.9	1.4	2.9	1.7	4	2.3	1.8	3.3
Road	Mean	0.4	0.1	0.1	0.3	0.2	0.1	0.2	0.1
density	Median	0.4	0.1	0.1	0.3	0.2	0.1	0.2	0.1
(km/km2)	Max	0.7	0.4	0.6	1.1	0.4	1.5	0.6	0.2
	Min	0.1	0	0	0	0	0	0	0.1
	S.D.	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0

Table 3: Summary Statistics Adminstrative units for 2nd stage

4 Empirical results

4.1 1st stage results

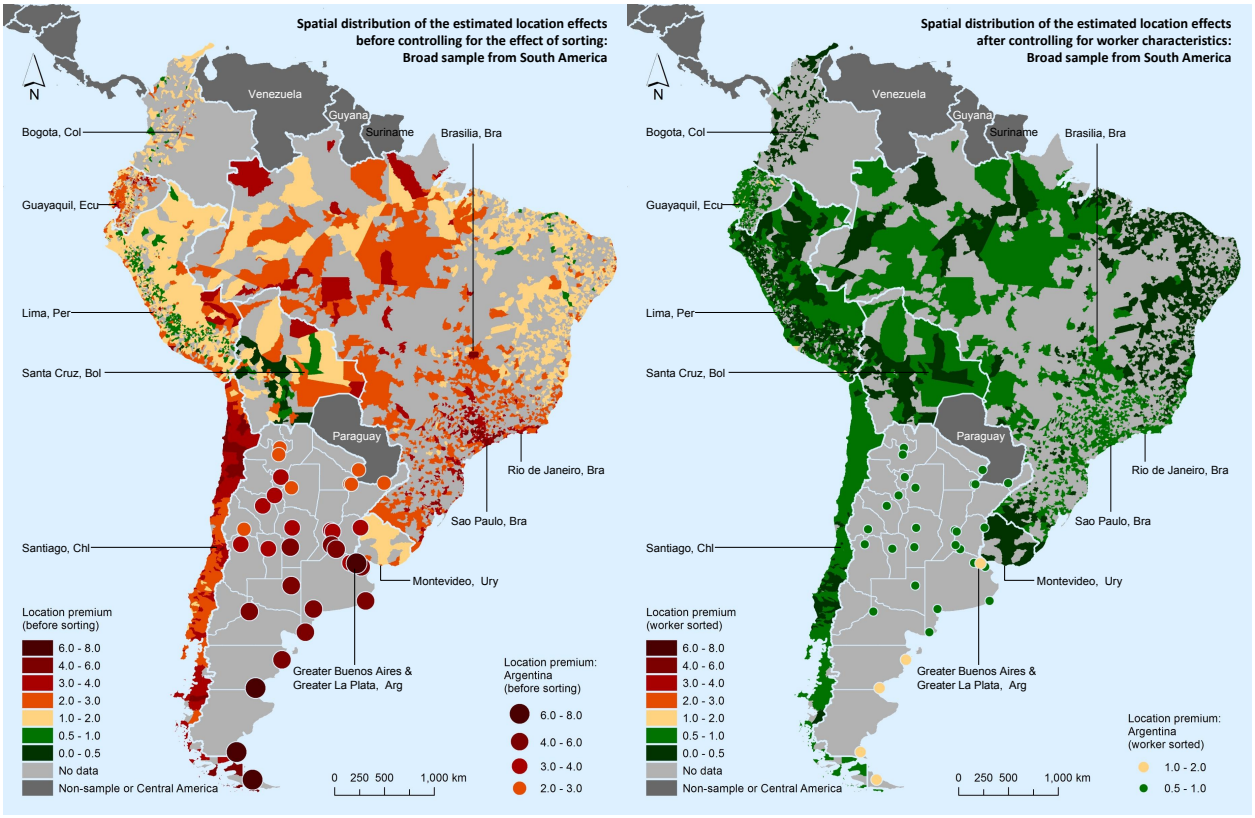
The locations obtained from the first stage can be found in figures 1 and 2 for South and Central America respectively. Individual numbers for location premia are available upon request from the authors. The figures show that there is high volatility within countries in the location effects derived without controlling for sorting ($\hat{\alpha}_i$ estimated in equation 1). These premia are equivalent to the average log wage in each city and are comparable across countries. This volatility is significantly muted once a control for sorting on worker observable characteristics is included ($\hat{\alpha}_i$ estimated in equation 2). This suggests that most of the variation in labor productivity across locations in the region is explained by sorting on observable characteristics. When age is introduced squared, the volatility is further muted, suggesting the model with age squared explains the wage variation better. The estimates suggest a non-monotonic relationship between age and income (an inverse U shape). Results for the sorting variable coefficients are shown in tables A1 and A2. The effect of sorting seems to be stronger even than in estimation carried out for developed countries. Additionally, we have performed estimation including job characteristics (not shown here) as a robustness check.

4.2 2nd stage results

Table 4 show the results of a set of 2nd stage regressions. As we saw in the first stage, a significant amount of the spatial variation in productivity within LAC countries is accounted for by sorting. Nevertheless, even after controlling for sorting, some variation does remain and this variation is correlated with municipality characteristics. Pooled regressions, which include country fixed effects, show a strong correlation with population density (column 1 in table 4), but population density loses-out in a horse race with human capital and market access (columns 3-5 in table 4).

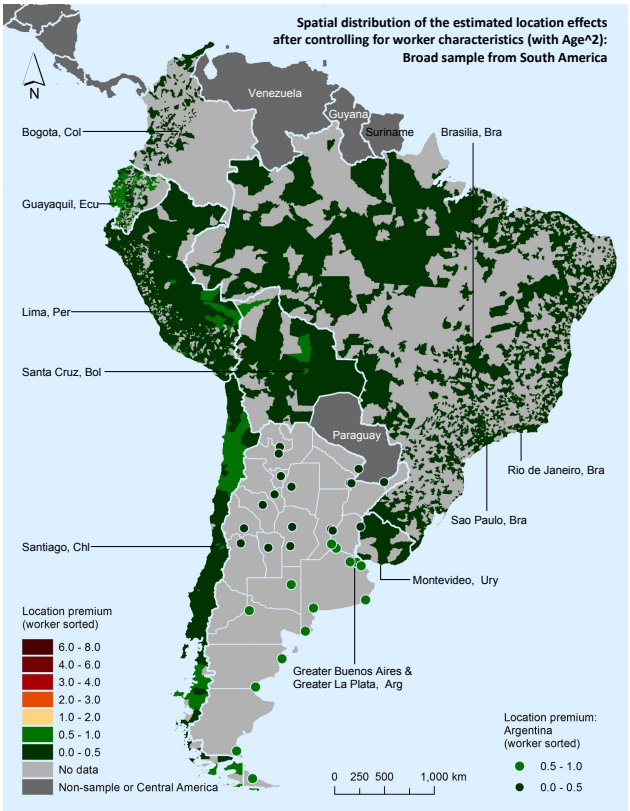
Table A3 shows the results for a narrow sample with only male workers from the private sector. The externalities of human capital are stronger in this case, reflecting the stronger market forces and agglomeration effects that are present in determining the wages in the private sector. Also, stronger evidence of agglomeration economies is found (as indicated by the coefficient on population density) for the narrow sample than for the broad sample, especially when our measure of human capital is the share of workers with tertiary education. This could indicate that the type of workers found in the private sector benefit more from agglomeration economies than those in the public sector.

Table A4 shows similar results with an alternative human capital variable, the percentage of workers with tertiary education, following Behrens et al. (2014) and Chauvin et al. (2017). In general, impact of human capital externalities is stronger when using average years of schooling. This is likely a consequence of the distribution of educational years of developing countries, where share of workers with tertiary education is much lower as discussed in section 3. Starting from a lower initial level, additional years of education, pre-tertiary levels, could make a stronger difference



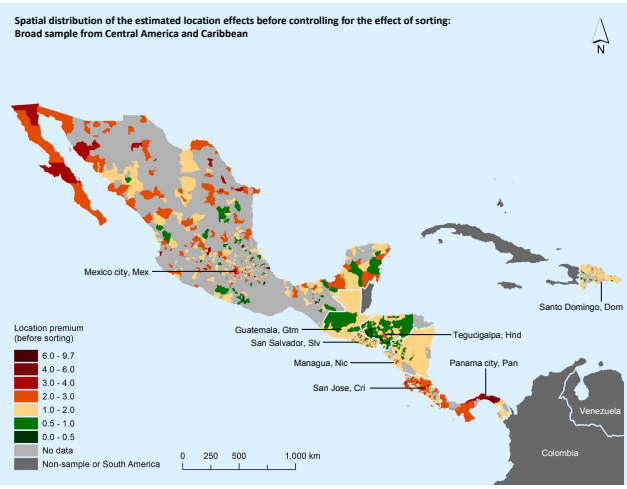
(a) Location premia after controlling for sorting.

(b) Location premia after controlling for sorting.

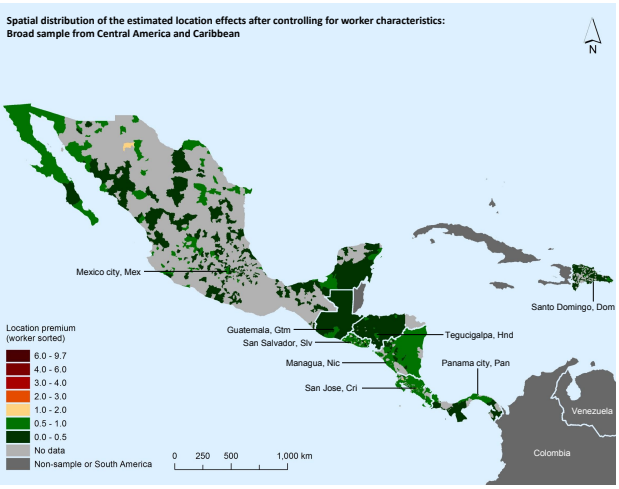


(c) Location premia after controlling for sorting with age^2

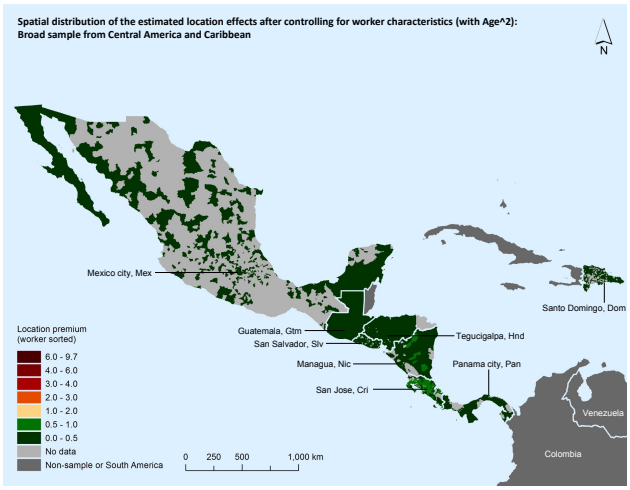
Figure 1: Location Premia in South America



(a) Location premia after controlling for sorting.



(b) Location premia after controlling for sorting.



(c) Location premia after controlling for sorting with age^2

Figure 2: Location Premia in Central America

Dependent variable: Location premium	**(1)**	**(2)**	**(3)**	**(4)**
Population density (ln)	0.049***	0.013*	0.005	0.011
	[4.638]	[1.831]	[0.557]	[1.235]
Average years of schooling (ln)		0.576***	0.574***	0.573***
		[10.319]	[9.711]	[9.688]
Market access (ln)			0.015***	0.015***
			[3.577]	[3.839]
Road density (ln)				-0.030***
				[-2.961]
Mean air temperature (ln)	0.03	0.044	0.051	0.048
	[0.695]	[1.169]	[1.624]	[1.706]
Terrain ruggedness (ln)	-0.031**	-0.024***	-0.017	-0.017
	[-2.443]	[-3.091]	[-1.580]	[-1.554]
Total precipitation (ln)	-0.028	-0.008	-0.01	-0.014
	[-0.629]	[-0.319]	[-0.447]	[-0.626]
Constant	-0.988***	-2.372***	-2.704***	-2.762***
	[-6.165]	[-14.809]	[-14.073]	[-14.290]
Observations	5,750	5,750	5,050	4,858
R-squared	0.654	0.736	0.76	0.765
Adjusted R-squared	0.653	0.735	0.759	0.764

1. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust t-statistics clustered at the country level in brackets.

2. Country effects have been controlled in all columns.

3. In all columns, the dependent variable is the estimated location premium after controlling for the effect of sorting on worker characteristics of the broad sample.

4. Broad sample refers to all employed wage/salary workers aged 14-65.

5. Worker characteristics include age, age-squared, marital status, gender, and the years of schooling.

6. The source of the population data is the Gridded Population of the World (GPW), v4.

Table 4: 2nd stage results

than in economies where most of the population has finished secondary education. This result could also highlight the consequences of heterogeneity in the sectoral composition of developed and developing economies.

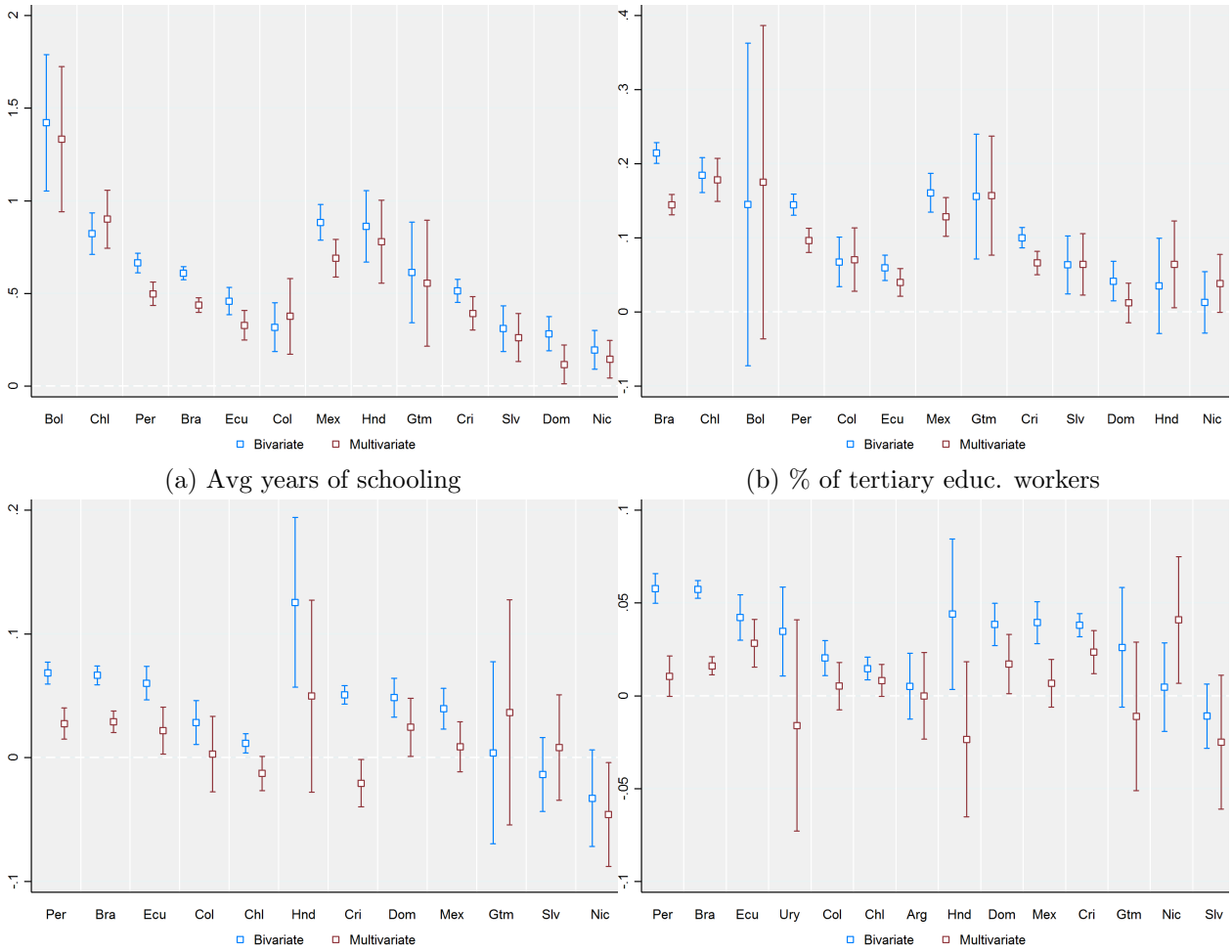
Table A5 explores non linear relationships of the variables. Of particular interest is the effect of density. The positive effect of the squared density shows a U shaped relationship, which contrasts with the inverted U shape more commonly found in the literature (Desmet and Rossi-Hansberg, 2014). Instead of the optimal density of a city (Au and Henderson, 2006) after which cities become too congested, this would suggest a minimum density threshold in the cities in the region after which agglomeration economies start emerging (or overcoming congestion costs). The difference could come from particularly high congestion costs at the current density levels in the region. This could be affected by low levels of infrastructure and institutional quality, which determine an economic context significantly different between cities in developed and developing countries. The literature has focused on studying agglomeration effects, and says little about urban congestion in developing countries, with a few exceptions (Desmet and Rossi-Hansberg, 2013; Desmet and Henderson, 2014; Duranton, 2016; Akbar and Duranton, 2017; Hanlon and Tian, 2015). Our result highlights once more the importance of studying agglomeration economies in the context of developing countries. The U shape relationship might not have been reached by developing countries cities yet, and could be lying ahead until better infrastructure eases the congestion costs and allows for agglomeration economies.

4.3 Country heterogeneity in 2nd stage

Figure 3 shows the summarized coefficients for the 2nd stage when estimated separately for each country [3]. The coefficient reported is obtained from estimating equation 3 and the 95% confidence intervals are obtained from the coefficient distribution. Bivariate refers to the estimated coefficient without other variables and multivariate refers to the estimated coefficient based on inclusion of all explanatory variables ($\overrightarrow{Worker_{l,t}}$ in equation 3).

In all countries, the effect of human capital externalities is positive on the location premia. However, the effects are very heterogenous across countries. This heterogeneity could be a reflection of country differences in educational quality. The effect of the percentage of tertiary education workers is also positive in most countries, but it is not significant for some of the countries in Central America. The effect of density loses significance in most countries after including other variables. The same happens to market access. These results are consistent with the aggregate results show in table 4. We hypothesized that the heterogeneity in these coefficients could be related to the level of development. We fit a linear regression with GDP per capita as explanatory variable. The only significant relationship found is between GDP per capita and human capital externalities as measured by the share of workers with tertiary education, shown in figure 5. As discussed before this relationship suggests an important story - as an economy develops and becomes more

[3] Argentina, Guatemala and Uruguay coefficients are not included because of extremely wide confidence intervals due to small numbers of observations (locations).



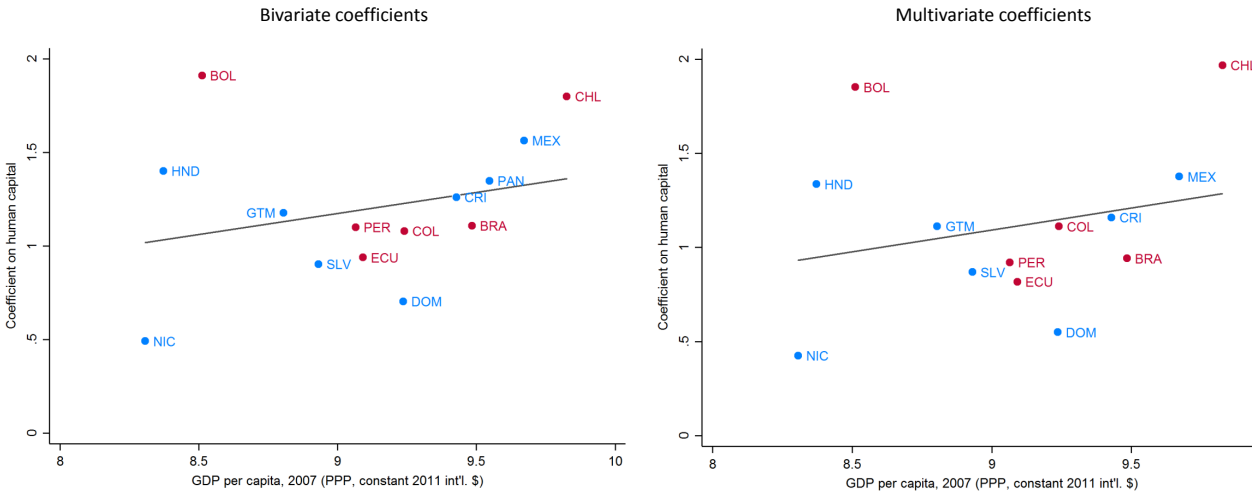
(a) Avg years of schooling

(b) % of tertiary educ. workers

(c) Density

(d) Mkt access

Figure 3: Heterogeneity in coefficients from 2^{nd} stage. Dependent variable is the estimated location premium after controlling for worker characteristics. Bivariate refers to the estimated coefficient and 95% confidence interval without other variables. Multivariate refers to the estimated coefficient and 95% confidence interval based on inclusion of all explanatory variables in equation 3. Sorted by South America and Central America plus Dominican Republic, and descending order of the estimated coefficients based on the bivariate model.



Sorting	Regression equation	t-stat (slope)	R-squared
None	Y= - 0.855 + 0.226X	0.99	0.075

Sorting	Regression equation	t-stat (slope)	R-squared
None	Y= - 0.999 + 0.232X	0.85	0.062

Figure 4: Estimated coefficients on human capital externality (average years of schooling) and GDP per capita. Dependent variable is the estimated location premium after controlling for worker characteristics. Bivariate refers to the estimated coefficient and 95% confidence interval without other variables. Multivariate refers to the estimated coefficient and 95% confidence interval based on inclusion of all explanatory variables in equation 3.

sophisticated, it becomes more important to focus on tertiary education. Table A4 shows full results for the 2^{nd} stage with this alternative measurement of human capital externalities.

5 Conclusions

This paper contributes to the existing gap by providing estimates for location wage premia and agglomeration economies for 16 countries in the Latin America region (covering approximately 80% of the considered countries' population). While two of the countries in our sample - Brazil and Colombia - have already been considered by Chauvin et al. (2017) and Duranton (2016) respectively - the remaining 14 countries have not, to our knowledge, been previously analyzed in the literature. We generate these estimates using a harmonized data set which contains information on both the nominal wages and characteristics of individual workers, and of the characteristics of the locations in which the workers live.

Within country location wage premia variation is largely explained by sorting on demographic

worker characteristics, mainly education. Once this sorting is taken into account, most of the variation of wages disappears. The model that introduces the relationship of age as non monotonic further mutes the variation observed in the raw wage data. This suggests that the usual higher productivities observed in Latin American cities are not necessarily produced by concentration economies and other productivity enhancing processes that happen in large cities, but by the mere location of better prepared workers in cities. Nevertheless, some variation does remain and this variation is correlated with municipality characteristics. Location productivity shows a strong correlation with human capital and market access (measured through a gravity equation). Density's effect disappears when the latter are introduced, which suggests that the effects observed actually come from human capital externalities and the more centralized locations. One possible alternative explanation is that density is indeed relevant but that its effect is not linear. In contrast with Desmet and Rossi-Hansberg (2014) and Au and Henderson (2006), we find a U shape relationship between location wage premia and density. This suggests a minimum density threshold in the cities in the region after which agglomeration economies start emerging (or overcoming congestion costs).

Heterogeneity across countries is investigated. It remains true, across most countries, that the effect of density, when measured linearly, loses in a horse race with market access and human capital externalities. In particular, the association of productivity and human capital at the municipality level is heterogeneous, which could be a reflection of country differences in educational quality.

This paper responds to an interest in studying the agglomeration economies of developing countries empirically. More than simply extending the state of knowledge on the extensive margin, we believe the dynamics could be significantly different in developing countries, where, for example, congestions costs could play a stronger role due to poor infrastructure and institutions. We find indeed that, some relationships are different in developing countries, such as weaker agglomeration economies, a U shaped relationship between density and city labor productivity, and a very strong role of sorting of workers with more education into larger cities. This could explain a higher concentration of high skilled industries in fewer cities in the developed world. Policy-wise, however, it is not clear what the welfare impact is of this strong sorting.

Current work on the paper includes exploring an IV approach to address the endogeneity of some variables, for example density, which we are instrumenting with pre-colonial population density as in Maloney and Valencia Caicedo (2016). Future work also includes performing the analysis at the urban extent level. In contrast with considering each administrative unit separately, grouping locations that share a labor market would more accurately capture city size and the extent of agglomeration economies.

References

Akbar, Prottoy and Gilles Duranton, "Measuring the Cost of Congestion in Highly Congested City: Bogotá," 2017.

Au, Chun-Chung and J Vernon Henderson, "Are Chinese cities too small?," *The Review of Economic Studies*, 2006, *73* (3), 549–576.

Behrens, Kristian, Gilles Duranton, and Frédéric Robert-Nicoud, "Productive cities: Sorting, selection, and agglomeration," *Journal of Political Economy*, 2014, *122* (3), 507–553.

Branson, Julia, Andrew Campbell-Sutton, Graeme Hornby, Duncan Hornby, and Chris Hill, "A geospatial database for Latin America and the Caribbean," Technical Report, University of Southampton 2016.

Chauvin, Juan Pablo, Edward Glaeser, Yueran Ma, and Kristina Tobio, "What is different about urbanization in rich and poor countries? Cities in Brazil, China, India and the United States," *Journal of Urban Economics*, 2017, *98*, 17–49.

Combes, Pierre-Philippe and Laurent Gobillon, "The Empirics of Agglomeration Economies. Pages pp. 247–348 of: Duranton, Gilles, Henderson, Vernon, & Strange, William (eds), Handbook of Urban and Regional Economics, vol. 5," 2015.

Desmet, Klaus and Esteban Rossi-Hansberg, "Urban accounting and welfare," *The American Economic Review*, 2013, *103* (6), 2296–2327.

_ and _ , "Analyzing urban systems: have megacities become too large?," 2014.

_ and J Vernon Henderson, "The geography of development within countries," 2014.

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

_ , "Agglomeration effects in Colombia," *Journal of Regional Science*, 2016, *56* (2), 210–238.

Hanlon, W Walker and Yuan Tian, "Killer cities: Past and present," *The American Economic Review*, 2015, *105* (5), 570–575.

Maloney, William F and Felipe Valencia Caicedo, "The persistence of (subnational) fortune," *The Economic Journal*, 2016.

Nunn, Nathan and Diego Puga, "Ruggedness: The blessing of bad geography in Africa," *Review of Economics and Statistics*, 2012, *94* (1), 20–36.

Overman, Henry G and Anthony J Venables, *Cities in the developing world*, Centre for Economic Performance, London School of Economics and Political Science, 2005.

A 2nd stage explanatory variables

Gridded Population of the World (GPW) global 1 km population count data for 2000 and 2010 data was used to determine population in admin units specified in table 1. UN adjusted measurements were used, which adjust raster cell values so that when summed to national level they are consistent with UN population estimates.

A small number of study areas, whether admin units or urban extents, contain NULL values. These occur due to the presence of very small or remote, sparsely or entirely unpopulated study areas. In GIS data terms, the NULL values occur because study areas do not intersect any data cells in the population base data. Robustness test were carried out with other population data sources, such as Worldpop and Landscan. Similar results were obtained. GPW 1 km scale is not a significant difficulty in our case, as less than 1 percent of all admin units are smaller than than 2 km^2 account for 118 of 14,439 units.

Average years of schooling were averaged by admin unit using the household surveys from SEDLAC (and IPUMS for Brazil). Minors are excluded from the calculation, to prevent confounding between a large share of young population and low educational averages. Robustness tests were carried out using instead percentage of workers with tertiary education, following Behrens et al. (2014) and Chauvin et al. (2017).

B First Stage Results

	Argentina				**Bolivia**			
	broad		narrow		broad		narrow	
marriage	0.118***	0.0793***	0.112***	0.0855***	0.148***	0.105***	0.230***	0.219***
education	0.0855***	0.0834***	0.0763***	0.0755***	0.0905***	0.0886***	0.0814***	0.0811***
male	0.111***	0.119***			0.154***	0.163***		
Age	0.0108***	0.0442***	0.0107***	0.0471***	0.00541***	0.0354***	-0.000706	0.0149
Age^2		-0.000414***		-0.000485***		-0.000378***		-0.000209
Observations	245,948	245,948	97,011	97,011	14,874	14,874	6,516	6,516
R-squared	0.760	0.762	0.769	0.770	0.461	0.463	0.467	0.468
	Honduras				**Mexico**			
marriage	0.128***	0.0712***	0.106***	0.0886***	0.158***	0.0925***	0.202***	0.176***
education	0.106***	0.104***	0.0907***	0.0905***	0.0948***	0.0912***	0.0848***	0.0839***
male	-0.0140	0.00811			0.151***	0.172***		
Age	0.0105***	0.0502***	0.00812***	0.0354***	0.0119***	0.0592***	0.0109***	0.0483***
Age^2		-0.000515***		-0.000372***		-0.000613***		-0.000506***
Observations	38,269	38,269	17,685	17,685	94,105	94,105	41,579	41,579
R-squared	0.351	0.355	0.332	0.333	0.666	0.671	0.701	0.702
	Chile				**Colombia**			
marriage	0.138***	0.124***	0.161***	0.146***	0.105***	0.0737***	0.145***	0.126***
education	0.103***	0.102***	0.101***	0.101***	0.103***	0.101***	0.0827***	0.0823***
male	0.178***	0.182***			0.219***	0.228***		
Age	0.00973***	0.0255***	0.0108***	0.0386***	0.0122***	0.0463***	0.0102***	0.0492***
Age^2		-0.000194***		-0.000369***		-0.000437***		-0.000525***
Observations	405,058	405,058	178,321	178,321	231,349	231,349	104,122	104,122
R-squared	0.726	0.726	0.744	0.745	0.552	0.556	0.586	0.588
	Nicaragua				**Panama**			
marriage	0.180***	0.146***	0.176***	0.168***	0.122***	0.101***	0.143***	0.131***
education	0.0651***	0.0635***	0.0704***	0.0702***	0.104***	0.102***	0.0851***	0.0849***
male	0.0376***	0.0519***			0.215***	0.221***		
Age	0.00861***	0.0353***	0.0123***	0.0247***	0.0106***	0.0322***	0.00792***	0.0317***
Age^2		-0.000349***		-0.000169		-0.000273***		-0.000321***
Observations	24,730	24,730	7,486	7,486	205,122	205,122	95,382	95,382
R-squared	0.295	0.298	0.304	0.304	0.728	0.729	0.715	0.715

Table A1: 1st stage results

	Costa Rica				**Dominican Republic**			
	broad		narrow		broad		narrow	
marriage	0.153***	0.125***	0.164***	0.145***	0.145***	0.106***	0.158***	0.138***
education	0.0918***	0.0902***	0.0766***	0.0761***	0.0673***	0.0657***	0.0530***	0.0528***
male	0.141***	0.150***			0.247***	0.260***		
Age	0.00734***	0.0307***	0.00563***	0.0327***	0.0125***	0.0486***	0.0126***	0.0509***
Age^2		-0.000303***		-0.000366***		-0.000462***		-0.000519***
Observations	129,202	129,202	64,076	64,076	131,608	131,608	65,046	65,046
R-squared	0.772	0.773	0.769	0.770	0.527	0.531	0.557	0.559
	Peru				**El Salvador**			
marriage	0.108***	0.0658***	0.144***	0.125***	0.0777***	0.0424***	0.122***	0.0961***
education	0.0720***	0.0692***	0.0613***	0.0606***	0.0735***	0.0718***	0.0648***	0.0645***
male	0.250***	0.261***			-0.0467***	-0.0292***		
Age	0.00736***	0.0360***	0.00833***	0.0347***	0.00790***	0.0441***	0.00392***	0.0551***
Age^2		-0.000363***		-0.000354***		-0.000459***		-0.000695***
Observations	459,915	459,915	195,175	195,175	27,117	27,117	11,978	11,978
R-squared	0.443	0.446	0.497	0.498	0.490	0.494	0.448	0.452
	Ecuador				**Guatemala**			
marriage	0.102***	0.0814***	0.118***	0.106***	0.116***	0.0400***	0.0777***	0.0558***
education	0.0719***	0.0707***	0.0556***	0.0553***	0.102***	0.0991***	0.0934***	0.0933***
male	0.197***	0.204***			0.0388***	0.0642***		
Age	0.00819***	0.0277***	0.00787***	0.0294***	0.00842***	0.0530***	0.00764***	0.0351***
Age^2		-0.000244***		-0.000290***		-0.000581***		-0.000373***
Observations	237,801	237,801	109,634	109,634	58,030	58,030	28,424	28,424
R-squared	0.634	0.635	0.656	0.657	0.346	0.354	0.337	0.338
	Uruguay				**Brazil**			
marriage	0.181***	0.144***	0.220***	0.193***	0.231***	0.151***		
education	0.112***	0.109***	0.118***	0.117***	0.118***	0.114***		
male	0.189***	0.198***			-0.403***	-0.419***		
Age	0.0178***	0.0531***	0.0207***	0.0587***	0.0225***	0.0860***		
Age^2		-0.000434***		-0.000501***		-0.000868***		
Observations	15,915	15,915	6,025	6,025	1,809,596	1,809,596		
R-squared	0.682	0.685	0.674	0.675	0.829	0.835		

Table A2: 1st stage results (continued)

C Alternative Specifications

Dependent variable: Location premium	**(1)**	**(2)**	**(3)**	**(4)**
Population density (ln)	0.057***	0.018*	0.005	0.013
	[6.108]	[2.025]	[0.532]	[1.362]
Average years of schooling (ln)		0.628***	0.620***	0.622***
		[9.259]	[8.496]	[8.846]
Market access (ln)			0.020***	0.019***
			[5.314]	[5.769]
Road density (ln)				-0.036***
				[-3.784]
Mean air temperature (ln)	0.042	0.058	0.068	0.067
	[0.781]	[1.158]	[1.503]	[1.660]
Terrain ruggedness (ln)	-0.036**	-0.028***	-0.022*	-0.021*
	[-2.624]	[-3.562]	[-2.029]	[-1.912]
Total precipitation (ln)	-0.037	-0.015	-0.018	-0.022
	[-0.763]	[-0.537]	[-0.764]	[-0.928]
Constant	-0.917***	-2.425***	-2.822***	-2.904***
	[-5.518]	[-11.656]	[-13.936]	[-14.286]
Observations	5,748	5,748	5,049	4,858
R-squared	0.564	0.651	0.681	0.687
Adjusted R-squared	0.562	0.65	0.679	0.685

1. *** p < 0.01, ** p < 0.05, * p < 0.1. Robust t-statistics clustered at the country level in brackets.

2. Country effects have been controlled in all columns.

3. In all columns, the dependent variable is the estimated location premium after controlling for the effect of sorting on worker characteristics of the broad sample.

4. Broad sample refers to all employed wage/salary workers aged 14-65.

5. Worker characteristics include age, age-squared, marital status, gender, and the years of schooling.

6. The source of the population data is the Gridded Population of the World (GPW), v4.

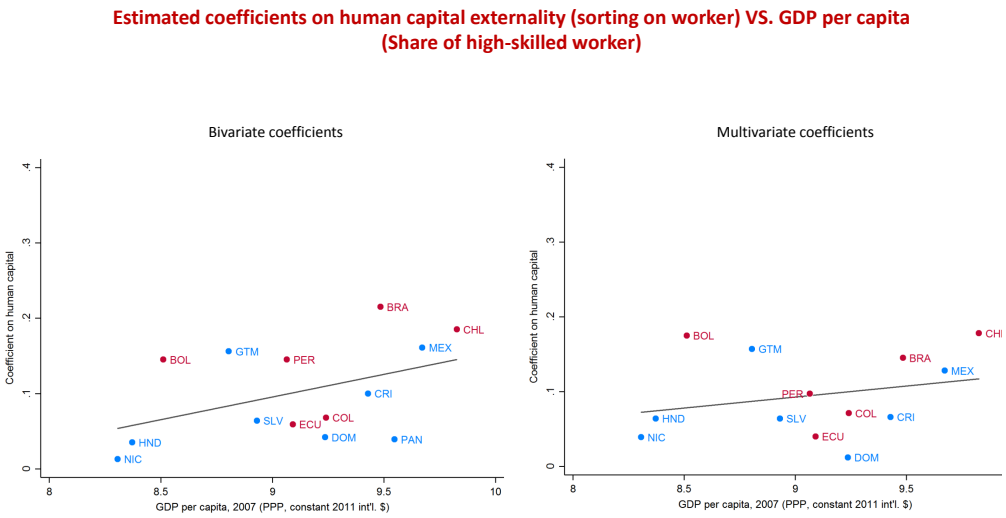
Table A3: 2nd stage regression - location premia from narrow sample

Dependent variable: Location premium	**(1)**	**(2)**	**(3)**
Population density (ln)	0.021*	0.001	0.007
	[2.104]	[0.121]	[0.788]
% of workers with tertiary education	0.022***	0.021***	0.021***
	[6.248]	[6.590]	[6.365]
Market access (ln)		0.026***	0.026***
		[4.301]	[4.452]
Road density (ln)			-0.028**
			[-2.499]
Constant	-1.296***	-1.821***	-1.874***
	[-7.351]	[-7.121]	[-7.335]
Observations	5,750	5,050	4,858
R-squared	0.701	0.727	0.732
Adjusted R-squared	0.7	0.726	0.731

All notes apply as in table A3 .

Variables representing geographic characteristics of each location, i.e., average annual air temperature, terrain ruggedness, and annual total precipitation, have been added in all specifications.

Table A4: Alternative Human Capital Measure.



Sorting	Regression equation	t-stat (slope)	R-squared
Worker	Y= - 0.448 + 0.060X	1.73	0.199

Sorting	Regression equation	t-stat (slope)	R-squared
Worker	Y= - 0.172 + 0.029X	0.87	0.065

Figure 5: Development and the effect of % of tertiary educ. workers

Dependent variable: Location premium	**(1)**	**(2)**	**(3)**
Population density (ln)	-0.024	0.012	-0.024
	[-1.317]	[1.297]	[-1.331]
Population density^2 (ln)	0.004**		0.004**
	[2.242]		[2.248]
Market access (ln)	0.013***	0.004	0.015
	[2.997]	[0.106]	[0.488]
Market access^2 (ln)		0	0
		[0.341]	[-0.082]
Average years of schooling (ln)	0.565***	0.573***	0.565***
	[9.668]	[9.647]	[9.698]
Road density (ln)	-0.030**	-0.030***	-0.030***
	[-2.909]	[-2.994]	[-2.958]
Constant	-2.669***	-2.680***	-2.685***
	[-13.847]	[-7.109]	[-7.591]
Observations	4,858	4,858	4,858
R-squared	0.766	0.765	0.766
Adjusted R-squared	0.765	0.764	0.765

Same notes apply as in table A4.

Table A5: 2nd stage with non-linear relationships