# *Beyond Kuznets: Average urban agglomeration size and income inequality*

**David Castells-Quintana***

**Abstract:**

As countries develop the percentage of total population living in urban areas (the rate of urbanisation) tends to increase. As this happens, inequality is expected first to increase and then to decline in what is known as the Kuznets inverted-U. But the development economics literature has not paid much attention to differences in the absolute size of cities potentially affecting economy-wide inequality. If cities of different sizes experience different levels of inequality, something that the urban economics literature suggests, the size of the different cities in a country may be another relevant factor to take into account when studying the overall level of inequality. This paper studies the relationship between *average* urban agglomeration (city) size and income inequality, using panel data for as many countries around the world as possible, looking at nation-wide inequality, controlling for several determinants of inequality, and considering non-linearities in this relationship.

**Key words:**

Agglomeration, city size, inequality, development

* Department of Applied Economics. Univ Autonoma de Barcelona. 08193 Bellaterra, Barcelona, Spain. David.Castells.Quintana@uab.cat

## 1. Introduction

One major characteristic of the process of economic development is the movement of people from rural to urban areas. As a result, the percentage of population living in urban areas (the rate of urbanisation) increases. According to classical theories (i.e., Lewis 1954; Kuznets 1955), this process is related to economy-wide inequality in a non-linear way: inequality first increases, as countries urbanise, and then declines as urbanisation proceeds. This non-linear relationship between income (and urbanisation) and inequality is known as the Kuznets' inverted-U. While this relationship considers the rate of urbanisation, it does not consider the absolute size of urban areas (cities), and how this changes along the process of development. For a fixed total population, the urbanisation rate of a given country may increase as the number of cities increase, or as the existing cities increase in size. If cities of different sizes experience different levels of inequality, as the urban economics literature suggests, the size of the different cities may be another relevant factor to take into account when studying the overall level of inequality. This is an issue that to date remains understudied. This paper analyses the relationship between *average* urban agglomeration (city) size and income inequality, using panel data for as many countries around the world as possible, looking at nation-wide inequality, controlling for several determinants of inequality, and considering non-linearities in the relationship.

Income inequality within countries has increased significantly during the last decades (see for instance Milanovic 2011 and Cairo-i-Cespedes and Castells-Quintana 2016). Understanding why and how inequalities increase is important in fairness terms, but also as the association between inequality and economic performance has been shown to depend on the factors defining inequalities (i.e., World Bank 2006; Marrero and Rodriguez 2014; Castells-Quintana and Royuela 2015). As the Kuznets' hypothesis and many recent papers highlight, spatial issues, especially those associated with urban dynamics, are likely to be crucial for inequality. Most countries today are either highly urbanised or are experiencing a fast process of urbanisation, with many cities experiencing rapid growth in size. Rapid urbanisation (and fast city growth) and increasing inequalities may not only be linked but are both today major challenges for many countries around the world. Consequently, understanding the relationship between average city size and income inequality becomes crucial for policy makers concerned with urban life and sustainable and inclusive development.

In relation to existing studies, the paper is closely linked to two main strands of the literature on inequality. On the one hand, the paper relates to works in the development economics literature studying the determinants of economy-wide inequality. Papers in this literature usually consider inequality *at the country level* (i.e., Fields 1979, for Least Developed Countries; Milanovic 1994, Li et al. 1998, Gustafsson and Johansson 1999, Barro 2000, Vanhoudt 2000, and Roine et al. 2009, for world samples; Odedokun and Round 2004, for Africa; and Castells-Quintana and Larrú 2014, for Latin America). Other papers study inequality *at the regional level* (i.e., Perugini and Martino 2008; Tselios 2008, 2014; Rodríguez-Pose and Tselios 2009; Castells-Quintana et al. 2015). One key and usual issue of analysis in this literature is that of the relationship between development (and urbanisation) and income inequality in the spirit of the Kuznets' inverted-U. But no paper in the development economics literature considers the size of cities as a potential determinant of inequality. On the other hand, the paper is also linked to the urban economic literature. Recent papers in this literature study the relationship between city size and income inequality *at the city level* (Duncan and Reiss 1956; Richardson 1973; Nord 1980; Long et al. 1977; Baum-Snow and Pavan 2013; Glaeser et al., 2015; Ma and Tang 2016). While these papers focus on size, they look at city inequality and do not consider effects on the level of economy-wide inequality. To the best of my knowledge, no paper has studied the relationship between average city size and economy-wide income inequality. This paper aims at filling this gap.

The remainder of the paper is organised as follows. Section 2 sets an empirical model. In section 3 the data used is described along some basic stylised facts. Section 4 discusses estimations and

results, while in section 5 some robustness checks are performed. Finally, section 6 concludes and derives policy implications from the results.

## 2. Deriving and empirical model:

Differently to empirical studies addressing the city size-inequality relationship, this paper considers economy-wide inequality. In this line, we perform our analysis at country level. We estimate a *cross-country* regression using panel data for inequality at country level:

$$inequality_{it} = \alpha_1 income_{it-1} + \alpha_2 income_{it-1}^2 + \beta(AveAggSize_{it-1}) + \psi X_{i\,t-1} + \varepsilon_{it} \qquad (1)$$

where $inequality_{it}$ is income inequality in country $i$ in time $t$, $income$ is income per capita (in logs), $X$ potential factors influencing income inequality, and $\varepsilon_{it}$ a country-time specific shock. Income per capita is considered in linear and quadratic form to capture the Kuznets' inverted-U. The key independent variable is *Ave Agg Size*, average (urban) agglomeration size, for each considered country and year. Urban agglomeration size, rather than city size, is considered, as the literature has shown that for both income and income inequality what matters is the size of the urban agglomeration rather than that of the city (although in the paper we may indistinctly refer to agglomeration or city size).

As with income per capita, we consider a linear as well as a quadratic term for our average agglomeration size. According to the city size-inequality literature, city size may influence income inequality in two different ways. One the one hand higher productivity from larger size is expected, due to agglomeration economies. This may benefit more the high-skill workers, and companies may be able to pay higher returns to abilities and efforts. But, on the other hand, larger cities also provide more opportunities, which may more strongly benefit low-income workers, reducing income inequality.[2] However, these two effects may have different weights for different city sizes, which would make the relationship between (city) size and (city) inequality non-linear; negative when small cities grow, but positive when large cities grow. In this line, we could also expect a non-linear relationship between average agglomeration size and the overall level of inequality; negative for low levels of average agglomeration size, but positive for high levels.

## 3. Data and Stylised facts:

### *Data*

To study the relationship between average agglomeration size and income inequality the performed analysis relies on panel data for as many countries as possible depending on data availability between 1960 and 2010. Data for income inequality for several countries and for a long time span is scarce, which has conditioned the analysis of the evolution and the determinants of inequality. In order to overcome this limitation, data from the Standardised World Income Inequality Database (SWIID), version 5.0, (Solt 2014) is used. SWIID uses a custom missing-data multiple-imputation algorithm to standardise observations. The database combines data from several sources, including the UN-WIID Database, the OECD Income Distribution Database, Eurostat, the World Top Incomes Database, the University of Texas Inequality Project, and the Luxemburg Income Study data. The SWIID data has been homogenized to maximise the comparability of available income inequality data across countries and over time. However, following Solt (2009; 2014), multiple-imputations are performed when using the data in order to take into account uncertainty from SWIID estimates.

[2] However, as the Todaro Paradox (Todaro 1969, 1976) states, higher opportunities in large cities attract more people into these cities, which may outweigh the benefits of job creation and lead to higher unemployment rates.

To construct the key explanatory variable, *Ave Agg Size*, data from the World Urbanisation Prospects - WUP - (UN 2014) is used.[3] The WUP gives data on agglomeration size, in terms of population, for agglomerations of more than 300 thousand inhabitants from 1950 onwards for as many countries in the world as possible (up to 199 countries, including more than 1690 urban agglomerations worldwide).[4] To construct the variable we consider all agglomerations above 300 thousand inhabitants and calculate country-year means.[5]

For income per capita (in logs) and its square, to capture Kuznets' inverted-U, data from the Penn World Tables (PWT) is used. As controls ($X$ in equation 1) several variables that the literature has found to potentially influence inequality at country level are considered, including economic growth (*ecogrowth*), investment shares (*ki*), government spending (*kg*), and educational levels (average years of *schooling*). As robustness additional variables that may be relevant to explain inequality are considered, including the percentage of urban population, fertility rates, coal rents, exports and the size of the agricultural sector (these last three as percentage of GDP). Other variables that may be correlated with average agglomeration size, like the population of the largest city, the percentage of total population living in urban agglomerations of more than one million inhabitants, and the percentage of urban population living in the largest city, are also considered, so $\beta$, our key coefficient, does not capture other relationship different than that of our key variable with inequality. All of these variables come from different sources, including the World Bank and the PWT. Annex A lists all variables' names, definitions and sources. Annex B shows descriptive statistics for main variables, while Annex C shows correlations among them.

### *Stylised facts*

Looking at the data some clear facts emerge. The first of these facts is the rapid pace of urbanisation. The percentage of the world population living in urban areas has increased from around 30 in 1950 to around 54 in 2015, and is expected to reach 66 by 2050 (according to WUP 2014 estimates). A second fact relates to the increase in the number of urban agglomerations. Considering urban agglomerations of more than 300 thousands inhabitants, the number of urban agglomerations around the world has increased from 304 in 1950 to 1729 in 2015 (and is projected to get to 2225 in 2030). The number of urban agglomerations with more than 1 million inhabitants has also gone up dramatically, from 77 in 1950 to 501 in 2015. And the number of agglomerations with more than 10 million inhabitants has gone from 2 in 1950 (Tokyo and New York) to 29 in 2015. A third fact relates to the average agglomeration size, which also shows a rapid increase, either looking at agglomerations across the world or looking at the average agglomeration size within countries. Figure 1 shows this increasing trend in average agglomeration size, while Figure 2 maps values for countries around the world in 2015. The mean across countries in average agglomeration size has increased from 253 thousand inhabitants in 1950 to 1.268 millions in 2015. Two single-agglomeration countries, Honk-Kong and Singapore, display the highest values. Among the top 20 countries only 3 are developed (Japan, Portugal and Greece), the rest are developing countries. In terms of population, these two facts, a higher number of urban agglomerations and a higher average agglomeration size, translates into more and more people living in large cities. While in 1950 around 300 million people in the world lived in urban agglomerations of more than 300 thousand inhabitants, this figure exceeds the 2.2 billions in 2015, which is almost a third of the total

[3] Frick and Rodriguez-Pose (2016) also use WUP data to calculate values of average city size for countries around the world and to analyse their effect on national economic growth.

[4] As many author have already highlighted, working with data on city size and urbanization rates poses the challenge of the definition of what constitutes a city, which may vary across countries (two recent papers working with city-size data across countries are Frick and Rodriguez-Pose 2016 and Gonzalez-Navarro and Turner 2016). WUP data takes this into account and aims at smoothing these differences as much as possible to ease comparability across countries.

[5] The focus on agglomerations above 300 million inhabitants lies on three main reasons: i) data availability, ii) the fact that agglomeration economies and congestion costs have been shown to be significant only in sufficiently large cities, and iii) the fact that according to Zipf´s law (rank times population size tends to be the same for all cities in a given country) information on cities above 300 thousand inhabitants should be enough to delineate the size of all cities.

world population, and 57% of the world urban population. And among all urban agglomerations, the cities of more than 10 million inhabitants concentrate alone more than 12 per cent of the world urban population.

Figure 1: increasing trend in average agglomeration size

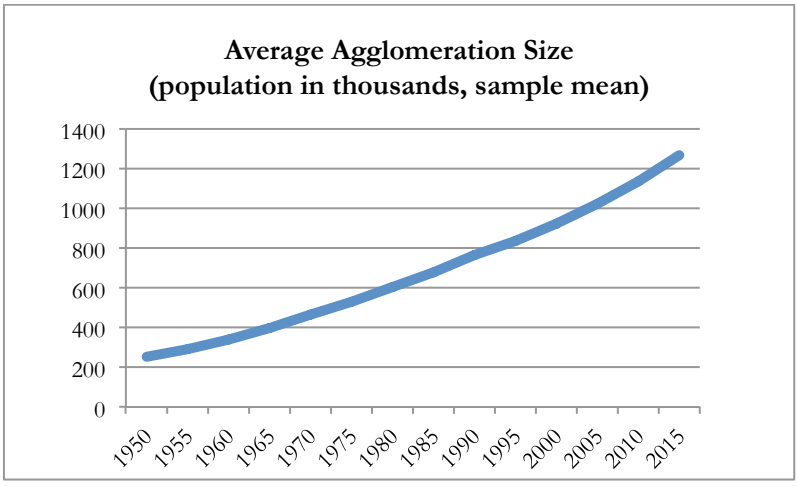
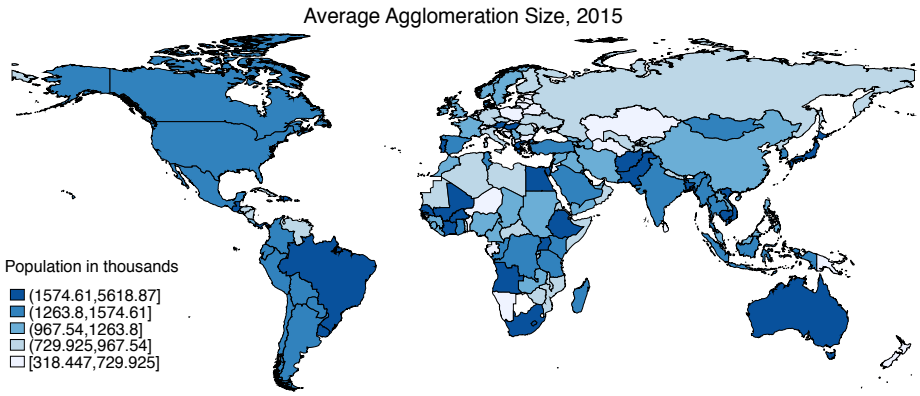



Figure 2: average agglomeration size around the world in 2015




A final stylised fact relates to the association between our key variables. Panel A in Figure 3 shows an inverted-U relationship between income and inequality levels, reflecting the Kuznets hypothesis. Panel B shows a different quadratic relationship beyond Kuznets', that between *Ave Agg Size* and *Inequality*. A U-shaped relationship emerges: inequality first declines and then increases with average agglomeration size.[6]

[6] Plots in Figure 3 consider all panel data. Annex D shows similar plots, displaying values for inequality levels in 2010 and average agglomeration size and income per capita in 1960.

Figure 3 Panel A: income per capita and inequality

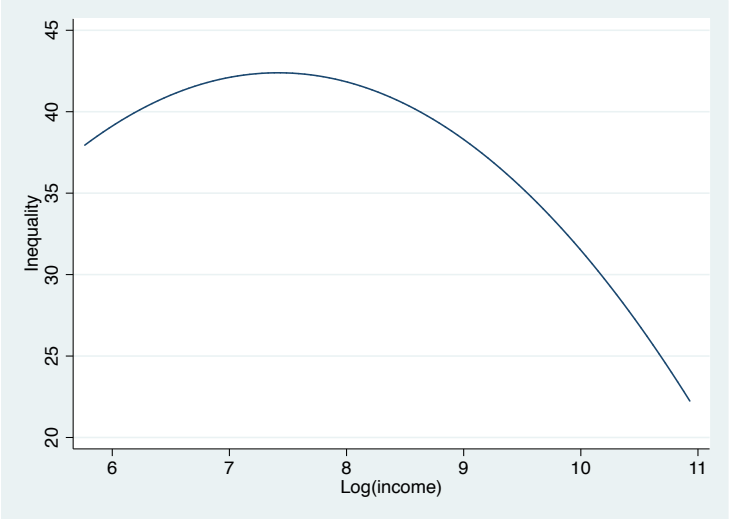
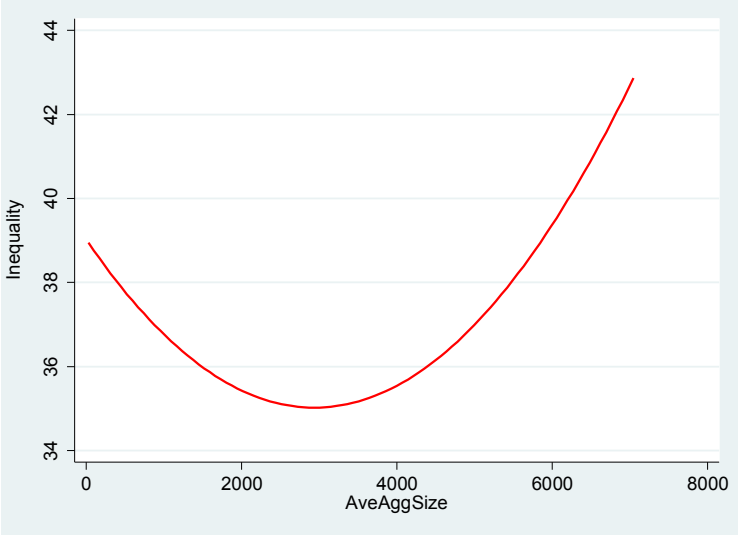



Figure 3 Panel B: average agglomeration size and inequality




## 4. Estimation and results

Equation 1 is estimated considering as many countries as possible (up to 131 in main estimations) and the longest time span depending on data availability (usually considering data from 1960 to 2010 and splitting the data on five-year periods). All right-hand-side variables are included one period before to reduce problems of reverse causality. As data to measure income inequality comes from Solt (2014), all estimations are done using multiple-imputation estimates (100 imputations) and small-sample adjustment. Time effects are included to control for global shocks. Several panel data techniques are implemented, including Ordinary Least Squares (pooled-OLS) and country-Fixed Effects (FE), in order to control for country-specific characteristics. All estimations are done with robust standard errors.

Table 1 presents main results. Column 1 only considers *Ave Agg Size* and presents pooled-OLS estimates. Results yield a negative and significant effect, indicating that the higher the average agglomeration size of a country the lower its level of income inequality. Column 2 considers *Ave Agg Size* and its square to control for non-linearities. Results yield a negative effect for the linear term and a positive for the quadratic, being both highly significant, and suggesting that inequality first decreases and then increases with average agglomeration size. Column 3 introduces income per capita (in logs) and its square to capture Kuznets' inverted-U. All coefficients are highly significant and have the expected signs, reflecting an inverted-U relationship between income and inequality (Kuznets), but also a U-shaped relationship between average agglomeration size and inequality (our hypothesis). Column 4 introduces country fixed effects. Results hold for *Ave Agg Size* and its square, but the coefficients for income are no longer significant. Finally, columns 5 and 6 introduce further controls (column 5 presents pooled-OLS estimates while column 6 introduces fixed effects). Controls have the expected sign (although coefficients are not always significant). *Ave Agg Size* and its square still display significant coefficients, negative the first and positive the second.[7]

Table 1: Main results

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dependent variable: *Inequality* (Gini Coefficient) | | | | | | |
| *Ave Agg Size* | -0.0012** | -0.0045** | -0.0045** | -0.0039* | -0.0028** | -0.0052** |
| | (0.0005) | (0.0010) | (0.0010) | (0.0023) | (0.0011) | (0.0026) |
| *Ave Agg Size²* | | 0.0001*** | 0.0001*** | 0.0001*** | 0.0001*** | 0.0001*** |
| | | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| Log*(income)* | | | 31.6839*** | 11.5567 | 32.5722*** | 10.799 |
| | | | (3.4654) | (8.3669) | (3.5497) | (8.8662) |
| Log*(income)²* | | | -2.0863*** | -0.5605 | -2.1052*** | -0.4389 |
| | | | (0.2108) | (0.4457) | (0.2140) | (0.4728) |
| *Eco growth* | | | | | -0.5060*** | 0.0789 |
| | | | | | (0.1302) | (0.0940) |
| Investment (*ki*) | | | | | -0.0191 | -0.0407 |
| | | | | | (0.0401) | (0.0526) |
| Gov spend (*kg*) | | | | | -0.0622 | -0.1569 |
| | | | | | (0.0793) | (0.1739) |
| Education (*schooling*) | | | | | -1.0209*** | -1.2582* |
| | | | | | (0.2850) | (0.6623) |
| Year FE | YES | YES | YES | YES | YES | YES |
| Country FE | NO | NO | NO | YES | NO | YES |
| Controls | NO | NO | NO | NO | YES | YES |
| Observations | 828 | 752 | 752 | 752 | 690 | 690 |
| No. of countries | 131 | 131 | 131 | 131 | 111 | 111 |

Note: All right-hand-side variables are lagged one period. *Econ growth*, *ki*, *kg* and *schooling* are calculated as averages over 5 years. All remaining variables are measured at the beginning of the period. The time span goes from 1970 to 2010. All estimations are done with multiple-estimation regressions (100 imputations) and small-sample correction. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

[7] Economic growth, as one of the controls, deserves special attention. Regressing economic growth on average agglomeration size and its square yields significant coefficients: economic growth increases and then declines with average agglomeration size (results available upon request). This result is expected according to the urban economics literature, given agglomeration benefits and congestion costs that come with city size, and are in line with Frick and Rodriguez-Pose (2016).

Estimates imply a U-shaped relationship between average agglomeration size and inequality. This relationship between the two variables suggests an optimal level of average agglomeration size. This level changes depending on the estimation, falling between 2 and 3 million inhabitants. In other words, everything else equal, an average agglomeration size between 2 and 3 million inhabitants minimizes the overall level of national inequality. An average agglomeration size of 3 million inhabitants turns to be a relatively high value. Most countries in our sample have levels of average agglomeration size below 3 and even 2 millions. In any case, countries differ greatly in what refers to the functional characteristics of their urban agglomerations (see for instance Castells-Quintana 2016), which is likely to influence the relationship between average agglomeration size and inequality. Consequently, we can expect each country to have its optimal level of average agglomeration size (something that arises as interesting for further research).

## 5. Further robustness checks

### *Confounding factors and additional controls*

As further robustness checks we can consider potential "confounding controls"; variables potentially correlated with average agglomeration size that may influence income inequality also in a non-linear way. In column 1 of Table 2 we introduce *poplargest*, the population of the largest city of the country, and its square. In column 2 we introduce *urb1m*, the percentage of total population living in cities of more than 1 million inhabitants, and its square. In column 3 we introduce *primacy*, the percentage of urban population living in the largest city. Primacy captures how concentrated is urban population in a country, which may be interesting to control for, to disentangle the effect of average agglomeration size from that of the urban structure of the country.[8] In all three cases the coefficients for *Ave Agg Size* and its square remain significant, negative the first and positive the second.

We can also consider many additional variables that may be relevant to explain income inequality (besides the already included controls) but at the expense of losing observations. Column 4 of Table 2 introduces urban rates, fertility rates, coal rents, exports, and the size of the agricultural sector (these last three variables as percentage of GDP). *Ave Agg Size* and its square remain with the correct sign, negative the first and positive the second, although now only the quadratic effect remains statistically significant.[9]

### *Dynamic specification*

Income inequality at country level has been shown to be very persistent over time. We can even expect the evolution of inequality to depend on previous levels. Taking this into account, one may want to consider a dynamic model, in which inequality in time $t$ depends on inequality in $t-1$. Column 5 of Table 2 does this by introducing lagged values of inequality as an additional control. The coefficient for the lagged dependent variable is positive and highly significant, confirming the persistence of inequality. Nevertheless, the coefficients for our key variables, *Ave Agg Size* and its square, remain significant and their values hardly change.

[8] Primacy has also been shown to be relevant for economic growth (i.e., Henderson 2003; Castells-Quintana 2016). It can be interesting to also examine its role in income inequality (something not done before). Results suggest that if one controls for average agglomeration size primacy plays no significant role in income inequality.

[9] I also checked robustness of the results to excluding potential outliers. Results hold.

Table 2: Robustness checks

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Dependent variable: *Inequality* (Gini Coefficient) | | | | | |
| *Ave Agg Size* | -0.0063** | -0.0079** | -0.0059** | -0.0052 | -0.0033* |
| | (0.0029) | (0.0035) | (0.0025) | (0.0046) | (0.0020) |
| *Ave Agg Size²* | 0.0001*** | 0.0001** | 0.0001*** | 0.0001** | 0.0001** |
| | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| *Log(income)* | 9.6732 | 10.1103 | 12.1773 | 32.1168*** | 8.2685 |
| | (8.8599) | (8.5044) | (8.9994) | (11.2565) | (8.1415) |
| *Log(income)²* | -0.3787 | -0.4181 | -0.5361 | -1.8180*** | -0.3535 |
| | (0.4802) | (0.4574) | (0.4832) | (0.6286) | (0.4432) |
| *Pop largest city* | 0.0001 | | | | |
| | (0.0004) | | | | |
| *Pop largest city²* | 0.0001 | | | | |
| | (0.0001) | | | | |
| *Urb 1m* | | 0.3598 | | | |
| | | (0.3562) | | | |
| *Urb 1m²* | | -0.0028 | | | |
| | | (0.0042) | | | |
| *Primacy* | | | 0.2400 | | |
| | | | (0.2299) | | |
| *Primacy²* | | | -0.0019 | | |
| | | | (0.0029) | | |
| $Inequality_{t-1}$ | | | | | 0.3845*** |
| | | | | | (0.0735) |
| Year FE | YES | YES | YES | YES | YES |
| Country FE | YES | YES | YES | YES | YES |
| Controls | YES | YES | YES | YES | YES |
| Additional controls | NO | NO | NO | YES | NO |
| Observations | 690 | 690 | 688 | 524 | 588 |
| No. of countries | 111 | 111 | 110 | 107 | 107 |
| Average RVI | 21.527 | 22.994 | 33.762 | 4.382 | 4.265 |
| Largest FMI | 0.283 | 0.116 | 0.262 | 0.148 | 0.155 |

Note: All right-hand-side variables are lagged one period. Controls include: *econ growth*, *ki*, *kg* and *schooling*. Additional controls include: *urbrate*, *fertility*, *coal*, *export*s, and *agriculture*. The time span goes from 1970 to 2010. All estimations are done with multiple-estimation regressions (100 imputations) and small-sample correction. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

## *Sorting and Endogeneity*

So far results point towards a U-shaped relationship between average agglomeration size and income inequality at country level, robust to a long list of controls. A relationship that is interesting in itself, and so far overlooked in the literature. Does this relationship imply a causal effect from average agglomeration size to income inequality? Papers working with income (or income inequality) at city level face a problem of sorting across cities: these papers need to disentangle the *true* effect of city size on income (or income inequality) from the one produced by the fact that larger cities attract people with different abilities and skills. With much less mobility across countries (and most probably not driven by cross-country differences in average city size), this problem is much lower when we work with income inequality at country level. But we can still face endogeneity concerns. First, due to reverse causality: it could be that higher inequality at country level leads to higher average agglomeration size, for instance if more unequal places grow at a faster rate - higher inequality has usually been associated with higher fertility rates (i.e., Barro 2000). Second, we may suffer from endogeneity due to relevant omitted variables. These concerns have already been taken into account. Estimations in Tables 1 and 2 introduced *Ave Agg Size*, and its square, lagged 5 years with respect to income inequality, in order to reduce reverse causality. Estimations in Table 2 also considered several additional controls potentially correlated with both average agglomeration size and income inequality. However, to further check for endogeneity we can perform alternative estimation techniques. Furthermore, a potential dynamic structure of the

data, suggested by results in column 1 of Table 3, implies that our FE results could be inconsistent - calling for a different estimation strategy if we want to get closer to a causal relationship. In this line two things are done. One is to first difference equation (1), in order to remove unobserved time-invariant country-specific characteristics that may be correlated with both average agglomeration size and income inequality. Column 1 of Table 3 shows first-differences (FD) estimates. Results are very similar to those in column 6 of Table 1.[10] A first-differences specification then allows us to use lags of *Ave Agg Size*, and its square, to predict first-differences and perform Instrumental Variables (FD-IV) estimations.[11] Consistency of IV estimates depends on the validity of the instruments. For lags of *Ave Agg Size* to be valid instruments they should not only be relevant (that is, explain first-differences in *Ave Agg Size*) but also exogenous and affect inequality only through first-differences in *Ave Agg Size* (the exclusion restriction). First-stage results in Appendix E show second and third lagged levels of *Ave Agg Size*, and its square, displaying significant power to predict first-differences. To test for the exclusion restriction we can estimate residuals from the first and second stage and then run residuals of the second stage on those from the first stage. Results are not significant, indicating that the two residuals are not correlated, and providing evidence to support the exclusion restriction. Table 3 reports additional tests that support the validity of the instruments. Column 2 uses second and third lagged levels of *Ave Agg Size*, and its square, as instruments. Column 3 uses third and fourth lagged levels. In both cases, FD-IV estimates yield significant coefficients for *Ave Agg Size*, and its square. These estimates point towards a causal effect of average agglomeration size on income inequality at country level, although this result should be taken with caution and could invite further research.

Table 3: First Differences and Instrumental Variables estimations

| | (1) FD | (2) FD-IV | (3) FD-IV |
|---|---|---|---|
| Dependent variable: Δ*Inequality* (Gini Coefficient) | | | |
| Δ*Ave Agg Size* | -0.0049* | -0.0096*** | -0.0110** |
| | (0.0025) | (0.0036) | (0.0044) |
| Δ*Ave Agg Size²* | 0.0001** | 0.0001*** | 0.0001** |
| | (0.0000) | (0.0000) | (0.0000) |
| Δ*Log(income)* | -1.4778 | -1.4644 | -1.3262 |
| | (2.9273) | (2.8637) | (2.8666) |
| Δ*Log(income)²* | 8.3129 | 7.7509 | 6.8004 |
| | (8.0442) | (7.7727) | (7.8528) |
| Year FE | YES | YES | YES |
| Controls | YES | YES | YES |
| Observations | 503 | 503 | 493 |
| No. of countries | 111 | 111 | 111 |
| AP first-stage F-stats p-value | | 0.000; 0.000 | 0.000; 0.012 |
| Kleibergen-Paap F-stat | | 38.09 | 7.171 |
| Kleibergen-Paap LM-stat | | 34.70** | 24.51*** |
| Hansen J stat p-value | | 0.699 | 0.531 |

Note: Controls include: Δ*econ growth*, Δ*ki*, Δ*kg* and Δ*schooling*. Instruments in column 2 are second and third lags of *Ave Agg Size*, and its square. Instruments in column 3 are third and forth lags of *Ave Agg Size*, and its square. Angrist-Pischke (AP) F tests the significance of excluded instruments. Kleibergen-Paap F-stat tests for weak instruments. Kleibergen-Paap LM-stat tests the null hypothesis that the equation is underidentified. Hansen J tests that the excluded instruments are uncorrelated with the error term. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

[10] In static models first differencing is almost equivalent to introducing fixed effect (see Wooldridge 2010).
[11] Gonzalez-Navarro and Turner (2016) also work with panel data on city-level population across the world, and use a similar identification strategy building on Olley and Pakes (1991) and Arellano and Bond (1991).

### *Cross-section specification*

Finally, there are questions as to whether panel methods are the most appropriate when working with variables that are fairly stable over time, as is the case with inequality (see for instance Easterly 2007). An alternative approach is to estimate equation (1) using a simple 'deep' cross-section, regressing inequality measured in 2010 on right-hand-side variables measured in 1960. This is another strategy to further reduce problems of reverse causality and consider a long-run association (50 years) between average agglomeration size and income inequality.[12] Columns 1 to 3 in Table 4 show estimates by OLS. Column 1 controls for the Kuznets' hypothesis, column 2 includes further controls, and column 3 also includes dummies for Latin America and the Caribbean, and Sub-Saharan African countries, which tend to display higher levels of inequality. Lastly, column 4 performs a simple IV, using as regressor levels of average agglomeration size in 2010 instrumented with levels in 1960. In all four columns the coefficients for *Ave Agg Size* and its square remain significant and in line with our panel results.[13]

Table 4: Cross-section results

| | (1) OLS | (2) OLS | (3) OLS | (4) IV |
|---|---|---|---|---|
| Dependent variable: *Inequality* (Gini Coefficient in 2010) | | | | |
| *Ave Agg Size* | -0.0118** | -0.0142*** | -0.0091** | -0.0068* |
| | (0.0053) | (0.0049) | (0.0044) | (0.0036) |
| *Ave Agg Size²* | 0.0001* | 0.0001** | 0.0001* | 0.0001* |
| | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| Log*(income)*$_{1960}$ | 37.1296*** | 45.6298*** | 35.7961*** | 45.2419*** |
| | (11.0606) | (11.2276) | (13.2371) | (12.5504) |
| Log*(income)*$^2_{1960}$ | -2.5164*** | -3.0940*** | -2.4826*** | -3.1567*** |
| | (0.6878) | (0.7050) | (0.8176) | (0.8089) |
| LAC dummy | | | 5.9140*** | 6.5128*** |
| | | | (1.8142) | (2.0386) |
| SSA dummy | | | 7.4859* | 8.4023*** |
| | | | (4.1869) | (2.3865) |
| Controls | NO | YES | YES | YES |
| Observations | 70 | 66 | 66 | 66 |
| Average RVI | 0.5708 | 0.3099 | 0.3383 | 0.322 |
| Largest FMI | 0.1691 | 0.1876 | 0.185 | 0.187 |
| AP first-stage F-stats p-value | | | | 0.006; 0.001 |
| Kleibergen-Paap F-stat | | | | 13.81 |
| Kleibergen-Paap LM-stat | | | | 15.50*** |

Note: All right-hand-side variables are measured in 1960. Controls include e*con growth*, *ki*, *kg* and *schooling*. In columns 1 to 3 *Ave Agg Size* and its square are measured in 1960. In column 4 *Ave Agg Size* and its square are measured in 2010 and instrumented with 1960 values. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

[12] Panel FE, or panel first-differences, estimates consider variation within countries over time, so results relate to the association between *changes* in average agglomeration size and *changes* in income inequality. A cross section setting considers variation between countries, so results relate to the association between *levels* in average agglomeration size, in this case in the past (1960), and *levels* in income inequality, in 2010.

[13] These cross-section regressions can also be estimated using Gini coefficients from the World Bank, rather than using Solt (2014) data, which depends on multiple imputations. Results are very similar, which reassures us about the robustness of the results to using alternative data for inequality. Results are available upon demand.

## 6. Discussion and policy implications

This paper has studied a neglected relationship in the development economics literature; that between average agglomeration size and income inequality. While the literature has emphasized the relationship between economic development (and urbanization) and income inequality, is has not paid much attention to the potential role of differences across economies and over time in urban agglomeration sizes. In order to address this issue, this paper has combined the literature on the determinants of income inequality at country level with the literature focusing on the relationship between city size and city-level inequality.

Using cross-country panel data for as many countries and for the longest time span as possible, results support an inverted-U relationship between development and inequality; inequality first increases and then declines with development (the Kuznets' hypothesis). But, beyond the Kuznets' inverted-U, results also suggest a U-shaped relationship between average agglomeration size and inequality; inequality first declines and then increases with average agglomeration size. This U-shaped relationship is found to be robust to several estimation techniques and a long list of controls, and is in line with insights from the urban economics literature and recent papers analyzing the association between different types of urbanization and inequality. These papers suggest that while urbanization in small and medium-sized cities is associated with decreasing inequality, urbanization in large cities is expected to increase inequality (Behrens and Robert-Nicoud 2014 and Castells-Quintana and Royuela 2015)

The policy implications from the results are straightforward. Larger average agglomeration size may be desirable when cities are small. In this case larger size is likely to lead to better economic performance, as cities benefit from agglomeration economies. Also income inequality is expected to fall. However, a very high average agglomeration size is undesirable. On the one hand, continuous growth of very large cities has been argued to reduce overall economic performance, mostly due to increasing congestion costs. On the other hand, as results in this paper show, excessive average agglomeration size is associated with increases in inequality. High inequality has been found to be detrimental for long-run economic growth, but also to hinder the benefits from agglomeration (Castells-Quintana and Royuela 2014). Consequently, results reinforce the idea that medium-sized cities may be more desirable for economic development: they may be associated with stronger long-run economic performance and to more cohesive societies. Nevertheless, as the urban economics literature has emphasized, to properly study the desirability of larger or smaller cities, it is important to consider further characteristics of cities beyond size. In this line, further research is needed to better understand the relationship between city size (and what happens in cities) and the overall level of inequality.

## References


Arellano, M. and Bond, S. 1991. Some test of specification for panel data: Monte Carlo evidence and an application to employment, *Review of Economic Studies* 58(2): 277-297.

Barro, R. J. 2000. Inequality and growth in a panel of countries, *Journal of Economic Growth,* 5: 5-32.

Baum-Snow, N., and Pavan, R. 2013. Inequality and city size, *The Review of Economics and Statistics* 95(5): 1535-1548.

Behrens, K., and Robert-Nicoud, F. 2014. Survival of the fittest in cities: Urbanisation and inequality, *The Economic Journal* 124(581): 1371-1400.

Cairo-Cespedes, G. and Castells-Quintana, D. 2016. Dimensions of the current systemic crisis, *Progress in Development Studies* 16(1): 1-23.

Castells-Quintana, D. 2016. Malthus living in a slum: urban concentration, infrastructure and economic growth, *Journal of Urban Economics*. DOI:10.1016/j.jue.2016.02.003

Castells-Quintana, D. and Larrú, J.M. 2015. Does aid reduce inequality? Evidence for Latin America, *European Journal of Development Research* 27: 826-849.

Castells-Quintana , D., Ramos, R. and Royuela, V. 2015. Inequality in European Regions: recent trends and determinants, *Review of Regional Research* 35: 123-146.

Castells-Quintana, D. and Royuela, V. 2015. Are increasing urbanization and inequalities symptoms of growth?, *Applied Spatial Analysis and Policy* 8(3): 291-308.
Castells-Quintana, D. and Royuela, V. 2014. Agglomeration, inequality and economic growth, *Annals of Regional Science* 52(2): 343-366.
Castells-Quintana, D. and Royuela, V. 2014. Tracking positive and negative effects of inequality on long-run growth. IREA-AQR Working Paper 2014/1.
Duncan, O., and Reiss, A. 1956. *Social Characteristics of Urban and Rural Communities, 1950*. John Wiley and Sons, New York.
Easterly, W. 2007. Inequality does cause underdevelopment: Insights from a new instrument, *Journal of Development Economics* 84: 755-776.
Fields, G.S. 1979. A Welfare Economic Approach to Growth and Distribution in the Dual Economy, *Quarterly Journal of Economics*, 93: 325-353.
Frick, S. and Rodriguez-Pose, A. 2016. Average city size and economic growth, *Cambridge Journal of Regions, Economy and Society*. DOI: 10.1093/cjres/rsw013
Glaeser, E., Resseger, M., and Tobio, K. 2015. Inequality in cities, *Journal of Regional Science* 49(4): 617-646.
Gonzalez-Navarro, M. and Turner, M. 2016. *Subways and Urban Growth: Evidence from Earth*. Unpublished manuscript.
Gustafsson, B. and Johansson, M. 1999. In search of smoking guns: What makes income inequality vary over time in different countries?, *American Sociological Review*, 64(4): 585-605.
Henderson, J.V. 2003. The urbanization process and economic growth: The so-what question, *Journal of Economic Growth* 8: 47-71.
Heston, A., Summers, R., Aten, B. 2012. Penn World Table Version 7.1, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania.
Kuznets, S. 1955. Economic Growth and Income Inequality, *American Economic Review* 45: 1–28.
Lewis, A. 1954. Economic development with unlimited supplies of labor, *Manchester School of Economics and Social Studies* 22: 139-191.
Li, H., Squire, L. and Zou, H. 1998. Explaining international and intertemporal variations in income inequality, *The Economic Journal* 108(446): 26-43.
Long et al. 1977. Income inequality and city size, *The Review of Economics and Statistics* 59(2): 244-246.
Ma, L., and Tang, Y. 2016. A tale of two tails: wage inequality and city size.
Marrero, G. and Rodríguez, J.G. (2014) "Inequality of opportunity and growth", *Journal of Development Economics* 104: 107-122.
Milanovic, B. 1994. Determinant of cross-country income inequality, *World Bank Policy Research Working Paper* 1246. World Bank.
Milanovic, B. 2011. Global Income Inequality by the Numbers: in History and Now - An Overview, *World Bank Policy Research Working Paper* No. 6259.
Nord, S. 1980. An Empirical Analysis of Income inequality and city size, *Southern Economic Journal* 46(3): 863-872.
Olley, G. and Pakes, A. 1991. The dynamics of productivity in the telecommunications equipment industry, *Econometrica* 64(6): 1263-1297.
Odedokun, M.O. and Round, J. 2004. Determinants of income inequality and its effects on economic growth: Evidence from African countries, *African Development Review*, 16(2): 287-327.
Perugini, C. and Martino, G. 2008. Income inequality within European regions: Determinants and effects on growth, *Review of Income and Wealth* 54(3): 373-406.
Richardson, H. 1973. *The Economics of Urban Size*. Saxon House: Westmead.
Rodriguez-Pose, A., Tselios, V. 2009. Education and income inequality in the regions of the European Union, *Journal of Regional Science*, 49: 411-437.
Roine, J., Vlachos, J. and Waldenstrom, D. 2009. The long-run determinants of inequality: what can we learn from top income data?, *Journal of Public Economics*, 93: 974-988.
Solt, F. 2009. Standardizing the World Income Inequality Database, *Social Science Quarterly* 90(2): 231-242.
Solt, F. 2014. The Standardized World Income Inequality Database. Working Paper. SWIID Version 5.0, October 2014.
Todaro, M. 1969. A model of labor migration and urban unemployment in less developed countries, *American Economic Review*, 59: 138-148.
Todaro, M. 1976. Urban job creation, induced migration and rising unemployment: A formula and simplified empirical test for LDCs, *Journal of Development Economics*, 3: 211-226.
Tselios, V. 2008. Income and educational inequalities in the regions of the European Union: geographical spillovers under welfare state restrictions" *Papers in Regional Science*, 87: 403-430.
Tselios, V. 2014. The Granger-causality between income and educational inequality: a spatial cross-regressive VAR framework, *Annals of Regional Science*, 53: 221-243.
United Nations, Department of Economic and Social Affairs, Population Division, 2015. World Urbanisation Prospects: The 2014 Revision. UN DESA Press.

Vanhoudt, P. 2000. An assessment of the macroeconomic determinants of inequality, *Applied Economics*, 32(7): 877-883.

Wooldridge, J. 2010. *Econometric Analysis of Cross Section and Panel Data*, second ed. MIT Press. Cambridge. MA.

World Bank. 2006. World Development Report 2006: Equity and development. The World Bank, Washington.

**Annex A: Variables' names, definitions and source**

| Main variables: | Description | Source |
|---|---|---|
| inequality | Income inequality measured by the Gini coefficient (Estimate in equivalised household market income) | SWIID v5.0 (Solt 2014) |
| Ave Agg Size | Average agglomeration size, in terms of population (thousand inhabitants) | Constructed with data from World Urbanization Prospects 2014. |
| income | Per capita GDP (in logs) | Constructed with data from PWT 7.1 (Heston et al. 2012), using real GDP chain data (rgdpch) |
| growth | Cumulative annual average per capita GDP growth rate | Constructed with data from PWT 7.1 (Heston et al. 2012), using real GDP chain data (rgdpch) |
| ki | Investment share (% of GDP) | PWT 7.1. (Heston et al. 2012) |
| kg | Government consumption (% of GDP) | PWT 7.1. (Heston et al. 2012) |
| schooling | Average years of secondary and tertiary schooling of adult population | Barro and Lee dataset |
| poplargest | Total population living in the largest city | World Urbanization Prospects 2014 |
| urb1m | Total population living in cities of more than 1 million inhabitants | World Bank - World Development Indicators |
| primacy | Population living in the largest city, as percentage of total urban population | World Bank - World Development Indicators |
| Further controls: | Description | Source |
| urbrate | Population living in urban areas, as percentage of total population. | World Urbanization Prospects 2014 |
| fertility | Fertility rates | World Bank - World Development Indicators |
| coal | Coal rents, as percentage of GDP | World Bank - World Development Indicators |
| exports | Total export, as percentage of GDP | World Bank - World Development Indicators |
| agriculture | Value added in agriculture, as percentage of GDP | World Bank - World Development Indicators |

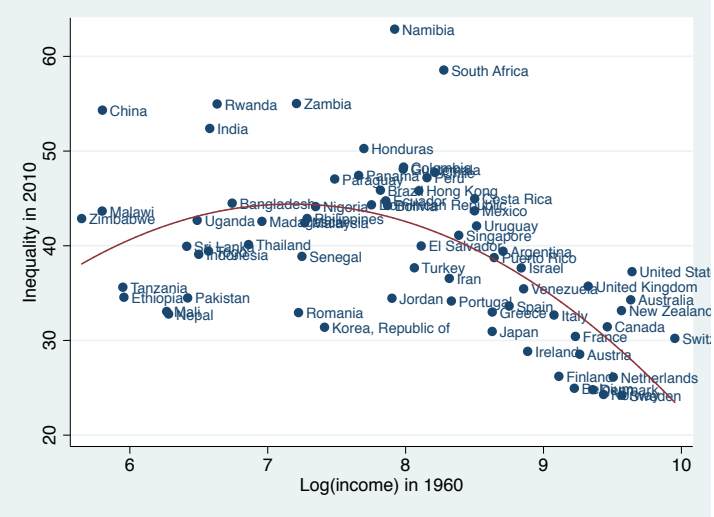
**Annex B: Descriptive statistics, main variables**

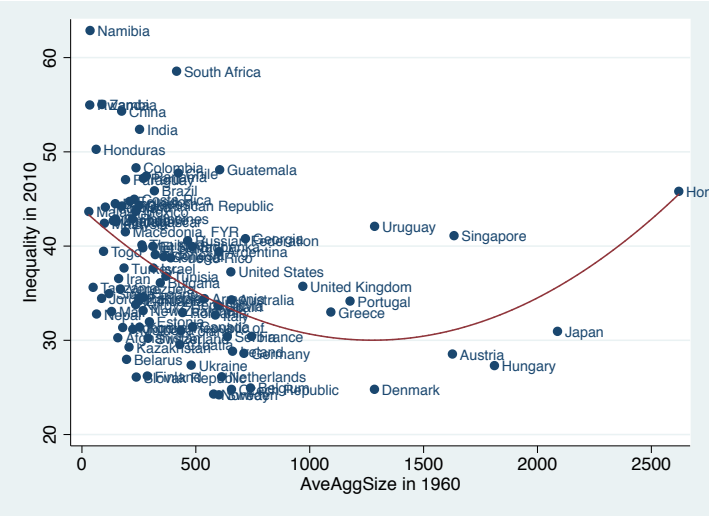
| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| *Ave Agg Size* | 2114 | 679.2489 | 678.1072 | 2.695 | 7313.557 |
| *ecogrowth* | 1223 | 1.974556 | 3.622458 | -20.65954 | 27.30487 |
| *income* | 1367 | 8.225289 | 1.312824 | 5.080978 | 11.82223 |
| *ki* | 1389 | 20.19026 | 10.30843 | 0.6262409 | 63.01598 |
| *kg* | 1533 | 9.575957 | 7.148799 | 0.0615228 | 56.41956 |
| *schooling* | 1360 | 2.022507 | 1.616491 | 0.02 | 8.06 |

**Annex C: Correlation matrix, main variables**

| | *Ave Agg Size* | *ecogrowth* | *income* | *ki* | *kg* | *schooling* |
|---|---|---|---|---|---|---|
| *Ave Agg Size* | 1 | | | | | |
| *ecogrowth* | 0.1077 | 1 | | | | |
| *income* | 0.3771 | 0.0489 | 1 | | | |
| *ki* | 0.1663 | 0.1844 | 0.3325 | 1 | | |
| *kg* | -0.2129 | -0.098 | -0.3159 | -0.187 | 1 | |
| *schooling* | 0.3754 | 0.0839 | 0.6916 | 0.1655 | -0.1956 | 1 |

**Annex D: Income per capita and inequality, and average agglomeration size and inequality**




**Annex E: First stage of IV regressions**

| | (1) | (2) |
|---|---|---|
| Dependent variable: | Δ*Ave Agg Size* | Δ*Ave Agg Size*$^2$ |
| *Ave Agg Size* $_{t-2}$ | 0.9695*** | 71.9257 |
| | (0.0885) | (56.6022) |
| *Ave Agg Size* $_{t-3}$ | -0.9798*** | -48.7362 |
| | (0.0942) | (57.3112) |
| *Ave Agg Size*$^2$ $_{t-2}$ | 0.0005 | 7.7114*** |
| | (0.0026) | (2.0159) |
| *Ave Agg Size*$^2$ $_{t-3}$ | -0.0006 | -8.8758*** |
| | (0.0034) | (2.4637) |
| Year FE | YES | YES |
| Controls | YES | YES |
| adj R square | 0.820 | 0.681 |
| Observations | 503 | 503 |
| No. of countries | 111 | 111 |
| AP first-stage F-stats p-value | 0.000 | 0.000 |

Note: Controls include: Δ*econ growth*, Δ*ki*, Δ*kg* and Δ*schooling*. Angrist-Pischke (AP) F tests the significance of excluded instruments. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1