Journal of Urban Economics 125 (2021) 103174

Contents lists available at ScienceDirect

Journal of Urban Economics

journal homepage: www.elsevier.com/locate/jue

Cities, lights, and skills in developing economies[☆]

Jonathan I. Dingel[a,*], Antonio Miscio[b], Donald R. Davis[c]

[a] Booth School of Business, University of Chicago and NBER, United States
[b] The Boston Consulting Group, United States
[c] Columbia University and NBER, United States

ARTICLE INFO

JEL classification:
R1
O1
O18
C8

Keywords:
Cities
Metropolitan areas
Satellite images
Skill-biased agglomeration
Zipf's law

ABSTRACT

In developed economies, agglomeration is skill-biased: larger cities are skill-abundant and exhibit higher skilled wage premia. This paper characterizes the spatial distributions of skills in Brazil, China, and India. To facilitate comparisons with developed-economy findings, we construct metropolitan areas for each of these economies by aggregating finer geographic units on the basis of contiguous areas of light in nighttime satellite images. Our results validate this procedure. These lights-based metropolitan areas mirror commuting-based definitions in the United States and Brazil. In China and India, which lack commuting-based definitions, lights-based metropolitan populations follow a power law, while administrative units do not. Examining variation in relative quantities and prices of skill across these metropolitan areas, we conclude that agglomeration is also skill-biased in Brazil, China, and India.

1. Introduction

This paper studies the distribution of skills across and within cities in three large developing economies: Brazil, China, and India. These three countries jointly account for approximately 40% of the world's population and are diverse in their levels of income. The process of urbanization in developing economies is important due to both the number of people involved and the opportunity to shape outcomes. The World Bank projects that 2.7 billion additional people will live in cities in developing economies by 2050. While urbanization does not necessarily imply growth, the two are nonetheless strongly linked (Henderson, 2014).

In developed economies, agglomeration appears to be skill-biased. There is a well-known positive relationship between metropolitan population and the share of the population possessing a college degree (Costa and Kahn, 2000; Moretti, 2004; Bacolod et al., 2009). Davis and Dingel (2017) study more narrowly defined educational categories and document that larger cities are skill-abundant and specialize in skill-intensive activities. Despite this greater relative supply of skill in larger cities, the college wage premium is also higher in larger cities (Baum-Snow and Pavan, 2013; Davis and Dingel, 2019). The implied greater relative demand for skill in larger cities suggests that agglomeration complements skill in production (Giannone, 2018). Within US metropolitan areas, more skilled residents have increasingly moved to city centers in the last two decades (Couture and Handbury, 2017).

Do the urban systems of developing economies also exhibit these spatial patterns? This paper is a first step in characterizing the spatial distributions of skills and sectors in Brazil, China, and India. Cities in developing economies will not necessarily mirror corresponding characteristics of developed economies. The existence of cities still requires agglomeration and dispersion forces, but the technologies and conditions of production and consumption in cities can diverge sharply. It is an empirical question whether developing economies' larger cities are populated by more skilled residents who earn relatively higher wages and live near the city center. We begin to tackle this question by examining some of these patterns in three large developing economies.

Studying the distribution of skills across and within metropolitan areas in Brazil, China, and India necessitates constructing metropolitan areas consistent with our economic inquiry. Economic theory treats a city as a highly – if imperfectly – integrated labor market. For

[☆] We are grateful to Stuart Rosenthal and two anonymous referees for very valuable suggestions. We thank Luis Costa, Kevin Dano, Shirley Yarin, Yue Yuan, and especially Dylan Clarke for excellent research assistance. We thank Andrei Potlogea for a helpful discussion and Juan Pablo Chauvin for sharing suggestions and data. Dingel thanks the Kathryn and Grant Swick Faculty Research Fund and the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business for supporting this work.

[*] Corresponding author.
E-mail addresses: jdingel@chicagobooth.edu (J.I. Dingel), miscio.antonio@bcg.com (A. Miscio), drdavis@columbia.edu (D.R. Davis).
URL: http://faculty.chicagobooth.edu/jonathan.dingel/ (J.I. Dingel), http://www.columbia.edu/~drd28 (D.R. Davis)

https://doi.org/10.1016/j.jue.2019.05.005
Received 1 July 2018; Received in revised form 20 May 2019
Available online 23 May 2019
0094-1190/© 2019 Elsevier Inc. All rights reserved.

this and other reasons, statistical agencies in developed economies overwhelmingly define metropolitan areas on the basis of commuting flows (Duranton, 2015). Unfortunately, such commuting flow measures are not always available to define metropolitan areas in developing economies. This is the case in China and India. In practice, researchers studying cities in developing economies have employed a variety of measures of the relevant geographies, often using off-the-shelf administrative definitions of cities. These spatial units often do not correspond to the metropolitan areas employed in research describing cities in developed economies. Administrative or political boundaries can fragment economically integrated areas into distinct cities or circumscribe places, including rural areas, that are not integrated metropolises. Assessing whether developing economies exhibit spatial patterns of skills similar to those of developed economies requires an appropriate geography defining cities' sizes and economic characteristics.

In Section 2, we develop a method to define metropolitan areas in the absence of commuting data by using satellite images. Our approach aggregates spatial units into metropolitan areas on the basis of lights at night. When municipalities or towns are part of a sufficiently bright, contiguous area of light, they belong to that metropolitan area. We demonstrate the feasibility and value of such an approach in three steps. First, we show that, with appropriately selected light-intensity thresholds, our night-lights–based method produces metropolitan areas that match commuting-defined US metropolitan areas very well. Second, we show that this is also true in a developing-economy setting, Brazil, where data on both commuting flows and night lights are available. Third, the application of our night-lights–based approach to China and India eliminates anomalies in their city-size distributions. While spatial units defined by administrative boundaries in these countries seem to deviate from a power-law distribution (Chauvin et al., 2017), our night-lights–based definitions of cities accord much better with the empirical regularity exhibited in developed economies.

Using these definitions of metropolitan areas, we aggregate census data to characterize the spatial distributions of skills across metropolitan areas in Section 3. We characterize spatial variation in relative quantities of skill by employing a linear regression implied by the theory of Davis and Dingel (2017). In all three developing economies, larger cities are skill-abundant. This result is robust to our choice of the light-intensity threshold employed in our algorithm defining metropolitan areas. However, we obtain substantially different population elasticities in some cases when using the administrative definitions of spatial units that have been commonly used in previous research.

We also characterize within-metro variation in quantities of skill for Brazil and China (Section 4) and spatial patterns of wages for Brazil (Section 5). In both Brazil and China, more skilled residents tend to live closer to the center of metropolitan areas. In Brazil, college wage premia are higher in more populous cities, consistent with developed-economy patterns and the hypothesis that agglomeration increases productivity in a skill-biased manner. The limited scope of this part of our investigation is dictated by data availability. Studies that use satellite imagery to both define urban markets and measure outcomes, such as Baragwanath Vogel et al. (2018), do not face such limitations. But in the absence of satellite-based means of measuring skill-related outcomes, we must employ both satellite and administrative data to answer fundamental questions about the urban systems of developing economies.

Our paper belongs to a growing literature on urbanization in developing economies. Perhaps most closely related are Henderson (1991) and Chauvin et al. (2017), who also focus on urban development in Brazil, China, and India, and Hu et al. (2014), who study China. In particular, Chauvin et al. (2017) examine whether stylized facts about metropolitan areas in the United States also hold true in Brazil, China, and India using administrative spatial units commonly available in government data releases. Hu et al. (2014) examine the predictions of Davis and Dingel (2017) for China using administrative spatial units. Our investigation complements these studies by focusing on the spatial distribution of skills and developing definitions of metropolitan areas that are more comparable to the economically integrated entities studied in research on developed-economy cities.

Our night-lights–based approach to defining metropolitan areas is distinct from the administrative units defined by government statistical agencies, a commuting-based algorithm introduced by Duranton (2015), and a distance-based clustering algorithm introduced by Rozenfeld et al. (2011). The administrative units defined by government agencies often do not correspond to the integrated metropolitan areas of interest to economists. The commuting-based approach is ideal, but its application is constrained by the absence of economy-wide commuting data in many countries. The city-clustering algorithm of Rozenfeld et al. (2011) aggregates adjacent spatial units on the basis of proximity without exploiting information about the contiguity of economic activity. We use night lights, which are available at very fine spatial resolution, to inform the aggregation of spatial units for which socioeconomic data are available.

Our employment of satellite imagery to define metropolitan areas belongs to a rapidly expanding economics literature exploiting satellite data, recently surveyed by Donaldson and Storeygard (2016). The use of satellite imagery to infer urban extent dates at least to Welch (1980), who inferred Chinese cities' populations from their built-up areas in the absence of a population census. Much of the recent economics research, such as Bleakley and Lin (2012), Henderson et al. (2012) and Storeygard (2016), has utilized night lights as a proxy for local economic activity at a finer resolution than typically documented in administrative data. We use night lights as a basis for identifying contiguous areas of economic activity that define metropolitan areas and then characterize those metropolitan areas' socioeconomic characteristics by aggregating spatial units available in more traditional data sources. Our application to India is similar to Harari (2017), who defines Indian cities' spatial extent using night lights, aggregates population counts for these footprints, and relates cities' economic outcomes to their compactness. Relative to her work, we validate the lights-based approach by comparing it to commuting-based definitions, show that lights-based metropolitan areas differ substantially from the geographic units used in much prior research on Brazil, China, and India, and characterize the spatial distributions of skills in these three economies.

2. Defining metropolitan areas

In order to characterize the spatial distribution of skills and sectors, we construct metropolitan areas from finer geographic units for Brazil, China, and India. Research describing cities in the United States and other developed economies typically uses spatial units defined by economic integration rather than legal jurisdictions or administrative boundaries. Agglomeration forces, commuting flows, and other economic linkages do not stop at municipal, county, or state borders, so using these boundaries to define the unit of analysis would fragment economically integrated metropolitan areas.[1] In Brazil, China, and India, however, prior research describing urbanization has used spatial units defined by administrative boundaries due to the absence of spatial units analogous to US metropolitan statistical areas in these countries.

We propose a method for constructing metropolitan areas from smaller geographic units based on night lights. First, we validate our method by showing that applying it to the United States yields spatial units very similar to those defined by the government statistical agency based on commuting flows. Second, we apply both our night-lights–based method and a commuting-flow method to Brazil, for which both

[1] As described by Duranton (2015), most commonly used definitions of metropolitan areas emphasize commuting flows as the relevant economic linkage, treating metropolitan areas as integrated labor markets. Employing administrative definitions that fragment these entities can alter research conclusions. For example, when workplaces employ a mix of skills but there is residential sorting by skill, calling such units cities would overstate between-city skill differences and understate within-city skill sorting.

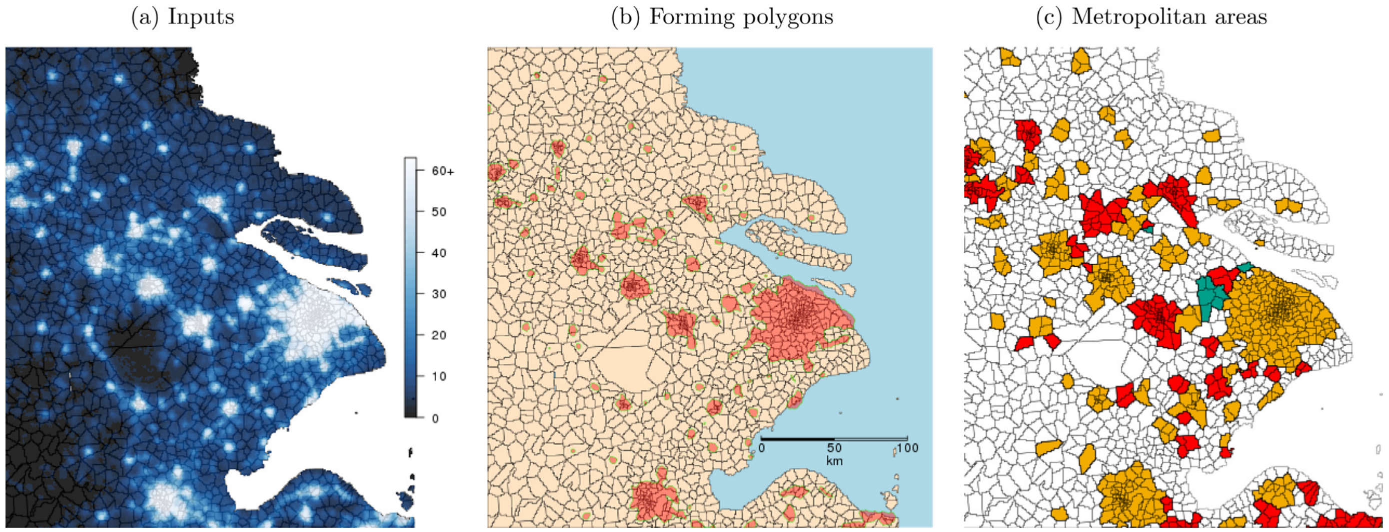


Fig. 1. Building metropolitan areas by aggregating smaller units based on lights at night.
Notes: This figure illustrates our procedure for combining satellite imagery of lights at night with administrative spatial units to build metropolitan areas. These panels depict a portion of the eastern coast of China in 2000. The administrative spatial units are townships. The polygons in the middle panel are areas of contiguous light brighter than 30. Aggregating the townships that intersect these polygons produces the metropolitan areas depicted in the right panel. Adjacent townships are often assigned to distinct metropolitan areas.

types of data are available, and find that they yield similar outcomes. Third, we construct metropolitan areas for China and India using satellite images of night lights, since commuting data are not available in these two countries.

In a number of cases, these metropolitan areas differ from urban units defined by political boundaries. These differences are sufficiently large that they affect conclusions about the distribution of population and economic activity across space. For example, we show that the city-size distribution in China conforms reasonably well to Zipf's law when we use night-lights–based metropolitan areas, while Chauvin et al. (2017) have shown substantial deviations from Zipf's law when using administrative units that incorporate substantial rural territories.[2]

2.1. Building metropolitan areas from satellite data

We propose a method for aggregating spatial units into a "metropolitan area" defined by a contiguous area of lights at night. Fig. 1 illustrates the procedure for a portion of the eastern coast of China along the East China Sea in 2000.

The two inputs to the algorithm are a satellite (raster) image of the country at night and a shapefile of the administrative units for which socioeconomic characteristics are reported. In the raster image, each pixel has a light intensity that is reported as an integer between 0 (no light) and 63 (top-coded value). The left panel of Fig. 1 depicts these values as a "heatmap" over the administrative boundaries of Chinese townships.

Upon selecting a light-intensity threshold, we identify contiguous areas of light brighter than the selected threshold. This yields polygons, as demonstrated in the middle panel of Fig. 1, which uses a light-intensity threshold of 30. Note that the polygons themselves are formed without reference to administrative boundaries. The largest polygon in that panel corresponds to the city of Shanghai. Our assumption is that contiguity of lights at night is informative about integration of economic activity.

We use the intersection of the night-lights–based polygons and the spatial units to construct metropolitan areas. A township that intersects one light polygon is assigned to that polygon. In the case of multiple intersections, a township is assigned to the light polygon containing the greatest area of the township.[3] The union of the spatial units assigned to a light polygon constitutes a metropolitan area. The right panel of Fig. 1 depicts the metropolitan areas that result from applying our procedure to Chinese townships. Note that, unlike a distance-based clustering algorithm, our procedure often assigns adjacent townships to different metropolitan areas.

Finally, we impose a minimum population size to include a metropolitan area in our analysis of metropolitan economic outcomes. Following the literature (e.g., Chauvin et al. 2017), we focus on metropolitan areas with populations greater than 100,000. A metropolitan area's population is the sum of the constituent spatial units' populations.

The choice of the light-intensity threshold, which governs the definitions of the resulting metropolitan areas, is not pinned down by economic theory or prior empirical research. The relationship between the threshold and the number of metropolitan areas defined is ambiguous. On the one hand, a lower threshold aggregates more spatial units into a given polygon, potentially defining more metropolitan areas with populations greater than 100,000. On the other hand, a lower threshold may cause distinct areas of light to be combined into one polygon, reducing the number of metropolitan areas. Later, we show that the number of metropolitan areas is indeed non-monotone in the light-intensity threshold. Beyond this extensive margin, the choice of threshold affects the composition of these metropolitan areas' characteristics. To address this issue, we report results for a variety of light-intensity thresholds and examine whether they are sensitive to this choice. Our qualitative conclusions about the spatial distribution of skills do not depend upon the particular threshold used.

[2] Zipf's law for cities (the number of cities larger than L is proportionate to $1/L$) is an empirical regularity found to hold in many countries and time periods (Gabaix and Ioannides, 2004), although numerous deviations have also been documented (Ades and Glaeser, 1995; Soo, 2005; Findeisen and Südekum, 2008). Theoretical models attribute this power-law distribution to random growth (Gabaix, 1999), a process of urban industrial churn (Duranton, 2007), or the product of multiple random factors (Lee and Li, 2013).

[3] Since townships are the finest spatial unit available, land area is the only characteristic upon which assignment criteria can depend. Our assignment of the entire township to one polygon makes our procedure comparable to other algorithms doing likewise, such as Duranton (2015).

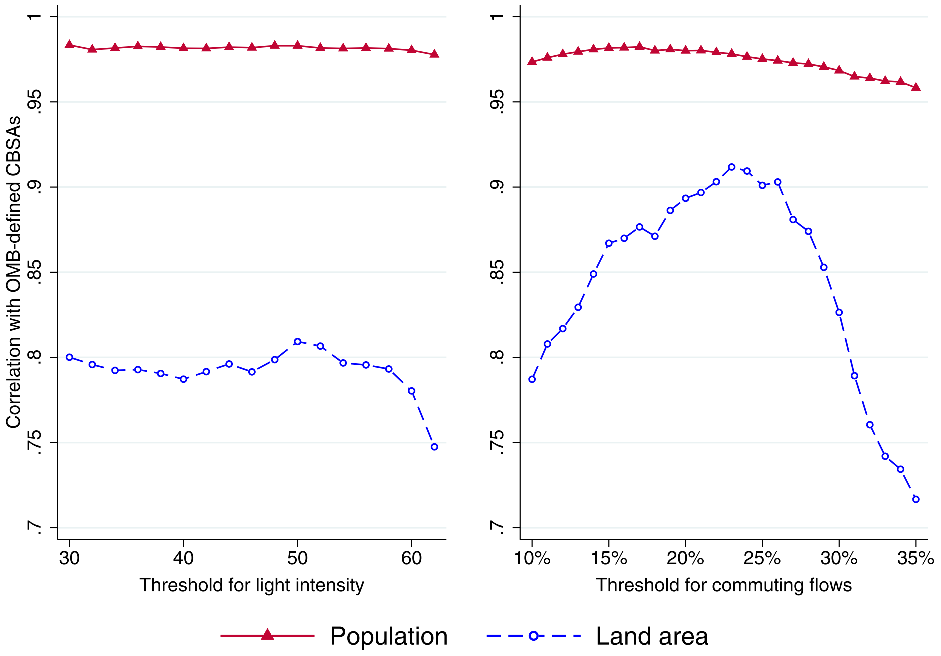


Fig. 2. Comparing population and land area across US metropolitan-area definitions, 2010.
Notes: The left panel depicts correlations of log population and log land area between metropolitan areas defined by contiguous areas of lights at night and 377 OMB-defined core-based statistical areas (CBSAs) with population above 100,000 in the 2010 US Census of Population. The right panel depicts correlations of log population and log land area between metropolitan areas defined by commuting flows and the OMB-defined CBSAs. The horizontal axes vary the thresholds for light intensity (left panel) and commuting flows (right panel) used to define metropolitan areas in our procedure and the Duranton (2015) procedure, respectively. Footnote 8 describes how we pair CBSAs with comparison counterparts.

As shown in the middle panel of Fig. 1, the night-lights–based polygons can intersect with sets of townships in a variety of ways. Given the spatial resolution of these polygons and administrative units, the edges of the resulting metropolitan areas may be defined with significant error.[4] We cannot really improve upon this, given the absence of data on within-township variation in economic characteristics. While we show that our empirical results are not sensitive to the choice of light-intensity threshold, research questions particularly focused on the “urban fringe” of metropolitan areas may be more sensitive to such choices.

2.2. US metropolitan areas

While this paper focuses on the spatial distribution of economic activity in developing economies, we use the United States as a testing ground to validate the night-lights–based method we develop to construct metropolitan areas in the absence of commuting data. In the United States, the Office of Management and Budget (OMB) aggregates US counties that meet certain requirements into a set of core-based statistical areas (CBSAs), which are designated metropolitan or micropolitan statistical areas depending on their size. The core is an urban population area of sufficiently large size. Outlying counties are adjoined to the central counties constituting this urban core on the basis of commuting ties.[5] Counties that do not meet these requirements are not included in any CBSA.

Recently, Duranton (2015) proposed an algorithm for defining metropolitan areas by the iterative aggregation of spatial units on the basis of commuting ties without requiring the initial designation of an urban core. Duranton applied this method to Colombia; here we apply it to US data to construct an alternative geography of US metropolitan areas. Our purpose is to establish that the Duranton (2015) method, which we will apply to Brazil, produces metropolitan areas similar to those defined by the OMB.[6]

We aggregate US counties into metropolitan areas on the basis of county-to-county commuting flows reported in the 2009–2013 American Community Survey.[7]

Our night-lights–based approach is a departure from these commuting-based methods. When we apply our night-lights–based method to the US, aggregating counties to build metropolitan areas, we obtain definitions of US metropolitan areas that are very similar to OMB-defined core-based statistical areas. To demonstrate this, we take the 377 OMB-defined CBSAs with a population above 100,000 as our baseline and match each one of them to the best corresponding metropolitan areas defined by the alternative methods based on commuting flows and night lights.[8] We then compare log population and log land area across agglomeration schemes, a comparison made by Rozenfeld et al. (2011) to validate their method.

The left panel of Fig. 2 shows that the correlation of log population between CBSAs and their night-lights–based counterparts is about 0.98 and relatively insensitive to the choice of light-intensity threshold. Similarly, the right panel shows that the correlations of log population between CBSAs and their commuting-flow-based counterparts

[4] This concern seems intuitively sensible, although we lack a precise notion of the urban boundary that we would define even if we possessed ideal data.

[5] An outlying county is aggregated into a CBSA if either of the following criteria is met: (i) at least 25% of the workers living in the outlying county work in the CBSA core; or (ii) at least 25% of the employment in the county is accounted for by workers who reside in the CBSA core. See Office of Management and Budget (2010) for a complete explanation. Prior to 2010, “local opinion” was an input into defining CBSAs.

[6] In cases of disagreement, it is not obvious which one should prefer. While the OMB definitions are widely used, Duranton (2015) provides reasons to prefer an algorithmic approach that does not require the designation of an initial core.

[7] The iterative algorithm requires the choices of a minimum commuting threshold to combine counties that are sufficiently connected by commuting ties. As discussed in Duranton (2015), the choice of a threshold depends on the size of the units to be aggregated as well as the level of economic development and quality of transportation systems. While a threshold of 10% was deemed appropriate for Colombian municipios (median land area 288 km^2), these criteria suggest that higher thresholds seem appropriate for the United States despite the much larger size of its counties (median land area 1,594 km^2). We report the results of constructing metropolitan areas using a range of commuting thresholds.

[8] This matching is not one-to-one in all cases. Counties from multiple CBSAs may be assigned to the same light-based metro, and different counties within one CBSA may be assigned to different light-based metros. We compare each of our light-based metropolitan areas to the CBSA with the largest population assigned to that metro, provided that CBSA does not have a greater population assigned to another light-based metro.

exceed 0.96 and vary little with the minimum commuting threshold used in the Duranton (2015) algorithm. Both the night-lights–based and commuting-based metropolitan areas exhibit larger discrepancies with OMB-defined CBSA in terms of land area, where the correlations average about 0.8. This is natural, given that definitions of these metropolitan areas are more likely to differ in their inclusion of boundary areas that have low population densities. Given our focus on the pattern of economic activity in terms of skills and sectoral employment, the alignment of population levels is more important for our purposes than the alignment of land area.

Our summary of these outcomes is that the US metropolitan-area population distribution can be well approximated by either of the alternative geographies and the quality of this approximation is not particularly sensitive to the threshold employed to define the agglomerations. This is our first finding validating our night-lights–based method, albeit in a developed-economy context. The fact that these methods are not particularly sensitive to their threshold parameters is encouraging for their application to settings where we cannot tune those parameters to replicate some (non-existent) official definition.

2.3. Brazilian metropolitan areas

Among the three developing economies that we study, only Brazil makes nationwide commuting data available, permitting us to implement more than one approach to defining metropolitan areas there. We will use this setting to compare our night-lights–based approach to the results of the approach based on commuting flows in a developing-economy context. Validating our night-lights–based approach in this setting is important because commuting-flow data is not available in the Chinese and Indian contexts.

Brazil is partitioned by a hierarchy of increasingly fine geographic units: states (26), mesoregions (137), microregions (558), and municipios (5565). The states and municipios are political entities. The mesoregions and microregions are areas defined by the Brazilian Institute of Geography and Statistics (IBGE) for statistical purposes and do not constitute autonomous political or administrative entities. The IBGE defines microregions according to shared forms of economic activity but not explicitly on the basis of commuting.[9] Our commuting-based and night-lights–based methods will be applied to municipios, the finest geographic unit available, in order to define metropolitan areas.

Prior research on local labor markets in Brazil has used four different geographic units. First, a number of papers have used microregions as the unit of analysis.[10] We will compare and contrast microregions with our commuting- and night-lights–based metropolitan areas below in Section 2.6. Second, a few researchers (e.g., Bustos et al. 2016; Cavalcanti et al. 2016) have used municipios as their spatial unit. This is appropriate for some research questions, but raises potential problems if the outcomes of interest depend on economic interactions at a supra-municipio level (e.g., local labor markets linked by commuting). Third, the IBGE recently defined *arranjos populacionais* by aggregating municipios on the basis of urban density and flows to work or school (IBGE, 2016). A few researchers have employed these units (Chauvin, 2017; Díaz-Lanchas et al., 2018; Scherer and Folch, 2017), and we will compare them to our metropolitan areas when describing the spatial distribution of skills in Sections 3 and 4.

Fourth, a less popular approach has employed definitions of metropolitan areas that the states themselves have developed.[11] These are known as *Regiões metropolitanas*. This has three problems. The first, again, is that agglomerations may cross state borders and the definitions of metropolitan areas do not include these cross-boundary areas. This problem was officially recognized by federal authorities in 1998 and solved with the introduction of a new type of metropolitan area that may cross state boundaries. The latter are called *Regiões integradas de desenvolvimento econômico* or *RIDE*. The second problem is that the criteria for inclusion are state-specific. As the following example illustrates, these legal definitions are subject to the vagaries of the legislative process, so they are not consistent across states nor time: the southern state of Santa Caterina suppressed five of its six *Regiões metropolitanas* in 2007, only to re-create all of them and a few more in 2010. The third problem is that by definition each *Região metropolitana* and *RIDE* must contain at least two municipios. This results in the exclusion of large agglomerations contained within one municipio. Finally, most states have used a high population cutoff for inclusion as a metropolitan area, with the consequence that many agglomerations, including some with populations of nearly half a million people, are excluded from these data.

Our first approach to building metropolitan areas in Brazil applies the Duranton (2015) method to 2010 Brazilian Census data on commuting flows between municipios.[12] We aggregate municipios into endogenously defined metropolitan areas using an iterative process that depends on our choice of a minimum commuting threshold. In our preferred specification, we use a threshold of 10% of the local working population.[13] We work with metropolitan areas with a minimum population of 100,000.[14]

Our second approach to building metropolitan areas in Brazil is based on satellite data characterizing lights at night, as described in Section 2.1. We construct encompassing polygons that depend on the choice of a light-intensity threshold. We then assign municipios to these polygons in order to define metropolitan areas. If a municipio intersects with a single polygon, it is assigned to the corresponding metropolitan area. If a municipio intersects multiple polygons, it is assigned to the polygon with which it has the largest overlap.

Our commuting-based and night-lights–based methods produce quite similar metropolitan areas. Taking the 10% commuting threshold as our preferred specification, we compare metropolitan areas defined by night lights and alternative commuting thresholds in terms of the correlation of log population and log land area. As Fig. 3 shows, the correlations for population are very high, exceeding 97%, across all the reported thresholds. That is, in terms of population, the commuting-based and night-lights–based metropolitan areas with populations above 100,000 are quite robust to the choice of agglomeration-method parameters. As in the US case, the correlations for land area are weaker but still quite informative, exceeding 80% for all light-intensity thresholds and 90% for all commuting thresholds. This is quite sensible because the municipios included or excluded are those at the boundary of the metropolitan areas, which typically have lower population densities and larger physical areas. The correlation is greater than in the US case because Brazilian municipios are typically smaller geographic areas than US counties.

The key result of our comparison of Brazilian metropolitan areas constructed on the basis of commuting and satellite data is their similarity.

[9] See the criteria employed at http://www.ngb.ibge.gov.br/Default.aspx?pagina=divisao.

[10] See for instance Kovak (2013); Dix-Carneiro and Kovak (2015); Costa et al. (2016); Chauvin et al. (2017).

[11] See, for instance, Hoffmann (2003). More generally, any study that relies on data from the Brazilian statistical agency (IBGE) aggregated by metropolitan area has indirectly used this definition, including commonly used data such as the National Sample Survey of Households (PNAD) and the Urban Labor Force Survey (PME).

[12] These commuting-flow data are not available for earlier years.

[13] As mentioned in footnote 7, the threshold choice depends on economic development, transport infrastructure, and geographic size. Compared to Colombia, Brazil's larger municipios (median area 416 km^2 vs 288 km^2) suggest a lower threshold, while its higher GDP per capita ($8,600 vs $6,000 in 2015) pushes in the opposite direction. Hence, we choose to apply to Brazil the same 10% threshold applied to Colombia by Duranton (2015).

[14] With the 10% threshold, we obtain 4807 metropolitan areas with populations ranging from 805 residents to 19 million. 192 of these metropolitan areas have populations greater than 100,000, and they contain 60% of Brazil's total population and 68% of its urban population.

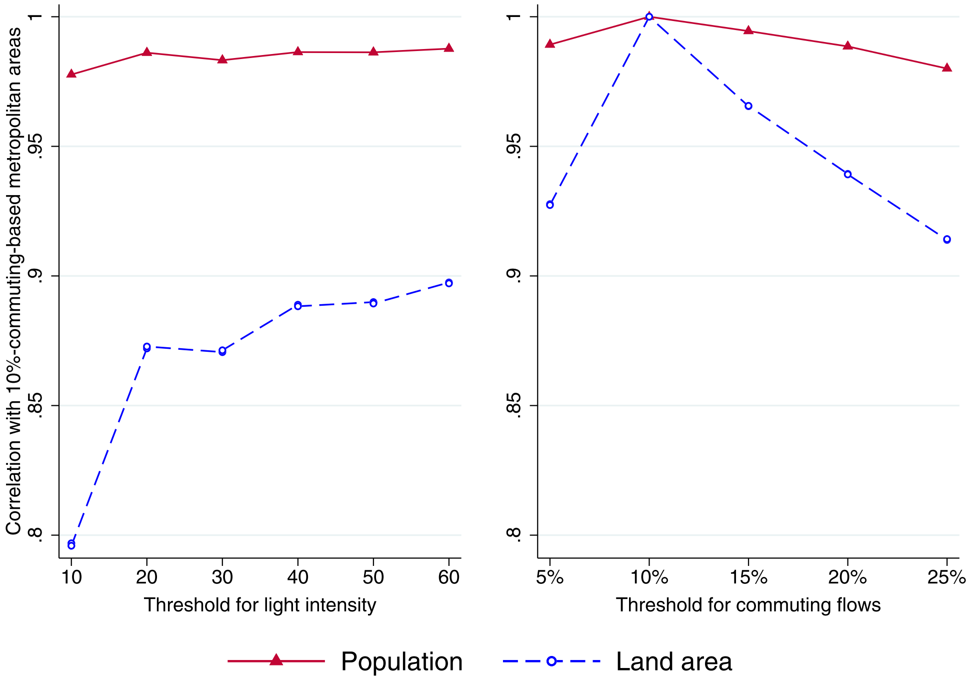


Fig. 3. Comparing population and land area across Brazilian metropolitan-area definitions, 2010.

Notes: The left panel depicts correlations of population and land area between metropolitan areas defined by contiguous areas of lights at night using different thresholds. The right panel depicts the same for metropolitan areas defined by commuting flows using different thresholds. The baseline for comparison in both panels is metropolitan areas defined by commuting flows in 2010 with a 10% threshold, per Duranton (2015). Thus, the perfect correlation at the 10% threshold in the right panel is tautological. Footnote 8 describes how we pair metropolitan areas with their baseline counterparts. The sample is restricted to metropolitan areas with population above 100,000.

There is a close correspondence between the preferred approach based on commuting data, which we will use as our baseline definition in our work on Brazil, and the night-lights–based approach that can be applied to all countries. This correspondence is relatively insensitive to the light-intensity threshold used. This should give us confidence that when we use satellite data in China and India, where we do not have commuting data, we will obtain sensible definitions of metropolitan areas. The weaker relation with physical area is of little consequence for the research questions we address here, as they do not depend on densities in an important way.

2.4. Chinese metropolitan areas

The basic geographical units in mainland China are provinces (31), prefectures (333, as of 2013), counties (2853), and townships (40,497). The first three are geographic partitions of the country. Townships, roughly speaking, partition the populated geography of the country, since the only areas excluded from townships have very small populations. Urban administrative units mirror this hierarchical governance structure: there are provincial-level cities, deputy-provincial cities, provincial capitals, prefecture-level cities, and county-level cities (Chan, 2010). These designations are time-varying and endogenous, as local jurisdictions can be promoted to higher administrative ranks as they grow. The four provincial-level cities, Beijing, Chongqing, Shanghai, and Tianjin, have the same administrative status as provinces and span many thousands of square kilometers in area. Prefecture-level cities are equivalent in administrative status to prefectures, span similarly large areas, and over time have come to dominate this administrative level as prefectures have been converted to prefecture-level cities. Today, nearly 300 of China's prefecture-equivalent administrative units are prefecture-level cities; less than a dozen are designated as prefectures.[15]

The spatial units most commonly used in prior research on Chinese urbanization have been the urban administrative units ranked as prefecture-level cities or higher. These offer one huge advantage: administrative cities are often the most conveniently available data (and in early periods may be the only form available). Yet there are large downsides. First, because prefectures differ dramatically in population, the set of prefecture-level cities includes some very small prefecture-level cities and does not recognize some very large cities that lack the prefecture-level designation. Second, provincial-level and prefecture-level cities incorporate both substantial rural areas and distinct urban areas not necessarily economically integrated with the prefecture-level city's urban core.[16] Third, the prefecture-level cities are necessarily bounded by the prefecture, whereas economically integrated metropolitan areas need not be. A particularly problematic example is the pair of prefecture-level cities of Guangzhou and Foshan. While administratively separate, they are geographically proximate; the distance from downtown Guangzhou to downtown Foshan is only about 18 miles. The two cities share connected subway lines, and it is not uncommon for people to live in Foshan and work in Guangzhou.

We use night lights to build Chinese metropolitan areas. While the preferred approach to defining an economically integrated labor market in economies such as the United States relies on commuting data, this method cannot be applied to China due to Chinese commuting data only being available for a quite limited set of areas. Based on our finding that commuting-based and night-lights–based methods delivered similar results when applied to Brazil, we apply the night-lights–based approach to China. We build metropolitan areas by aggregating counties or townships.[17] The latter is preferable, because it addresses the problems of erroneously including economically disconnected areas and rural areas in the defined metropolitan areas. Unfortunately, township-level data for 2010 are not yet publicly available for many socioeconomic characteristics of interest.

[15] In addition to prefectures and prefecture-level cities, this administrative level includes "leagues" of Inner Mongolia and "autonomous prefectures." Chan (2007) warns that the "system of urban definitions used in Mainland China appears to be the world's most complicated and confusing."

[16] Chan (2007) reports that "the current administrative boundaries of a great majority of large Chinese cities extend far beyond the familiar 'metropolitan area' or 'city proper' patterns by including rural counties, some with dense farming populations."

[17] In year 2000 definitions, the median township had a land area of 72 km^2. In year 2000 definitions, the median county had a land area of 1582 km^2.

Table 1
Comparing Chinese township- and county-based metropolitan areas, 2000.

		Correlation with township-based					
		Intensity: 10		Intensity: 30		Intensity: 50	
Metropolitan scheme	N	Pop'n	Land	Pop'n	Land	Pop'n	Land
County-based, intensity 10	1705	0.76	0.41	0.78	0.42	0.81	0.38
County-based, intensity 20	1464	0.71	0.28	0.73	0.31	0.82	0.35
County-based, intensity 30	1167	0.70	0.20	0.70	0.27	0.78	0.28
County-based, intensity 40	811	0.74	0.08	0.72	0.22	0.76	0.24
County-based, intensity 50	501	0.77	0.05	0.75	0.15	0.75	0.26
County-based, intensity 60	185	0.84	0.09	0.85	0.22	0.86	0.26
Township-based, intensity 10	1139			0.93	0.76	0.91	0.66
Township-based, intensity 20	960	0.92	0.81	0.98	0.87	0.95	0.75
Township-based, intensity 30	805	0.90	0.75			0.96	0.80
Township-based, intensity 40	599	0.89	0.66	0.95	0.82	0.98	0.88
Township-based, intensity 50	405	0.86	0.60	0.90	0.74		
Township-based, intensity 60	151	0.84	0.50	0.88	0.62	0.91	0.65

Notes: The first column reports the number of metropolitan areas with population exceeding 100,000 that are defined by that row's metropolitan scheme. Each cell in the following six columns reports the correlation coefficient for log population or log land area between the metropolitan scheme identified in the row and the metropolitan scheme identified in the column pairs for China in 2000. We pair metropolitan areas for comparison as described in footnote 8.

There are substantial differences between metropolitan areas obtained by aggregating townships and those obtained by aggregating counties. Table 1 illustrates these differences in two dimensions, reporting the correlations of log population and log land areas across comparable locations under different metropolitan-area definitions. The metropolitan areas obtained by aggregating townships are relatively consistent across different choices of the light-intensity threshold. The level of correlation typically exceeds 0.8 for population and 0.6 for land area. In contrast, the correlation between county-based and township-based metropolitan areas are typically below 0.8 for population and 0.4 for land area. This is unsurprising, as there are an order of magnitude more townships than counties in China, and townships cover only populated areas while counties partition the entire landmass. This strongly favors using township- over county-based metropolitan areas when possible. For total population and land area, Table 1 demonstrates the possibility that metropolitan characteristics may not be sensitive to the choice of light-intensity threshold.

2.5. *Indian metropolitan areas*

India is partitioned by a hierarchy of increasingly fine geographic units: states (35), districts (640), and sub-districts (5564).[18] Distinctly, the Census of India divides the country into urban and rural areas, with urban areas being comprised of two types of towns, "statutory towns" defined by their political character and places that are sufficiently populous, non-agricultural, and dense to be declared "census towns."[19] The Census furthermore defines an "urban agglomeration" (UA) as one or more physically contiguous towns with at least 20,000 residents. There were 384 UAs in 2001 and 475 UAs in 2011. Towns and urban agglomerations can span subdistrict and district borders, but by definition they do not cross state borders. This results in major metropolitan areas composed of multiple urban agglomerations. For example, Chandigarh is a city and union territory that is the capital of the states of Haryana and Punjab that is part of the "tricity" Chandigarh Capital Region, which has a regional planning board to coordinate an economically integrated area that spans three states.

Most prior research on urbanization in India has used (the urban population of) districts as the geographic units of interest. This has two immediate shortcomings. The first is that the towns within a district need not themselves be contiguous or have strong economic connections. This is non-trivial since an Indian district is roughly twice the size of a US county. The second is that there may be strong connections between contiguous urban areas in different districts that are ignored in this approach. Each of these problems finds a partial solution in the Indian statistical agencies' definition of "urban agglomerations."

We consider two different methods for defining Indian metropolitan areas, each imperfect in some respects. The first is to apply our night-lights–based approach to the urban populations of subdistricts, the finest spatial unit for which both a geographic shapefile and socioeconomic characteristics are publicly available.[20] However, only a limited set of socioeconomic characteristics are reported for subdistricts.[21] The second is to use administratively defined urban agglomerations and cities, agglomerated across state borders on the basis of night lights.[22] Socioeconomic characteristics are available for urban agglomerations' component census towns of population greater than 100,000. Unfortunately, we are not aware of publicly available shapefiles for towns and villages, which we would need to apply our night-lights–based approach to geographically finer administrative units.

[18] Here we use "states" to refer to "states and union territories." There were 35 states prior to 2014, when a new state, Telangana, was created, constituted by ten districts formerly in northwestern Andhra Pradesh. Sub-districts are known by names that vary across states, including mandal, tahsil, taluk, and block. See "Statement showing the Nomenclature and Number of Sub-Districts in States/UTs ".

[19] Statutory towns are administrative units defined to be urban, such as municipal corporations, municipalities, and so forth. In 2011, a "census town" was a place with population greater than 5000 persons, at least 75% of male laborers working outside agriculture, and population density greater than 400 persons per square kilometer. See Census of India 2011, Provisional Population Totals, Urban Agglomerations and Cities.

[20] In year 2001 definitions, the median sub-district had a land area of 374 km^2.

[21] Chauvin (2017), Harari (2017), and other researchers have addressed this shortcoming by using district-level averages as proxies for city-level averages when studying economic outcomes.

[22] We aggregate urban agglomerations and towns across state borders using contiguous areas with light intensity exceeding 20 defined by collections of subdistricts. When such a polygon crosses state borders, we aggregate the urban agglomerations and sufficiently large towns belonging to that polygon into a single metropolitan area. This produces two metropolitan areas that span three states: Greater Delhi and Chandigarh Tricity.

2.6. Comparison with administrative units

Prior work on urbanization in Brazil, China, and India has typically relied upon administrative units, such as microregions in Brazil and prefecture-level cities in China, that do not necessarily coincide with economically integrated metropolitan areas. In this section, we compare our definitions of metropolitan areas to the geographic units employed in previous research.

For Brazil, comparing our commuting-based metropolitan areas to the microregions used in prior research reveals substantial discrepancies. Microregions may be defined too narrowly or too broadly for such purposes. The former occurs frequently when agglomerations cross state boundaries, since microregions are defined to be strict subsets within a single state. The latter occurs when there are multiple small agglomerations of similar economic activity grouped into a single microregion even though these components are not significantly integrated by commuting. For example, Fig. 4 shows all the commuting-based metropolitan areas (color-coded) with a population above 100,000 in northeastern Brazil and microregion boundaries (dashed). We can spot several metropolitan areas that cross microregion boundaries, as well as one microregion that contains two distinct metropolitan areas. Moreover, we can see that most microregions containing a metropolitan area also encompass large areas that are not integrated to the metropolitan area by commuting ties. This mismatch between microregion boundaries and commuting-based metropolitan areas occurs in other areas of Brazil as well. 44 of the 192 metropolitan areas with population greater than 100,000, containing 59% of the population of such locations, span multiple microregions. 34 of the 208 microregions containing municipios that are part of a metropolitan area with population greater than 100,000 contain municipios assigned to more than one metropolitan area. Insofar as we think the economic integration implied by commuting should inform definitions of metropolitan areas, this casts doubt on interpreting microregions as metropolitan areas or local labor markets.

Given the close correspondence between our commuting-based and night-lights–based metropolitan areas for Brazil, the contrasts between our night-lights–based metropolitan areas and microregions are similar.

Prior work on China has used prefecture-level cities, the administrative capitals described in Section 2.4. Notably, Chauvin et al. (2017) find that the Chinese city-size distribution is poorly described by Zipf's law when using prefecture-level cities. The left panel of Fig. 5, taken from their work, shows a rank-size relationship that is more log-quadratic than log-linear. They describe this result as finding that "China has fewer ultra-large cities than the US city size distribution would predict" and suggest a number of possible explanations. These include that China's city-size distribution may be far from steady state, may be significantly distorted by urban planning, may be shaped by disamenities unique to extreme population sizes over 20 million, or that "China and India may be better seen as continents rather than standard countries." Another potential explanation is that the finding is simply a statistical artifact of the geographic units used to characterize the Chinese city-size distribution.

There are considerable differences between Chinese administrative cities and the metropolitan areas we define based on lights at night. While there are a few hundred administrative cities, our aggregations of townships yield twice as many or more metropolitan areas with population greater than 100,000. In addition, the metropolitan areas that correspond to locations for which prefecture-level cities are defined differ meaningfully in terms of their populations and land coverage. Fig. 6 reports the correlation of log population and log land area between metropolitan areas defined at various light-intensity thresholds and their prefecture-level-city counterparts. The correlation for log population never exceeds 0.8, and the correlation for log land area is always below 0.4. Given these contrasts, using different geographic units may yield very different conclusions about the spatial distribution of economic activity in China.

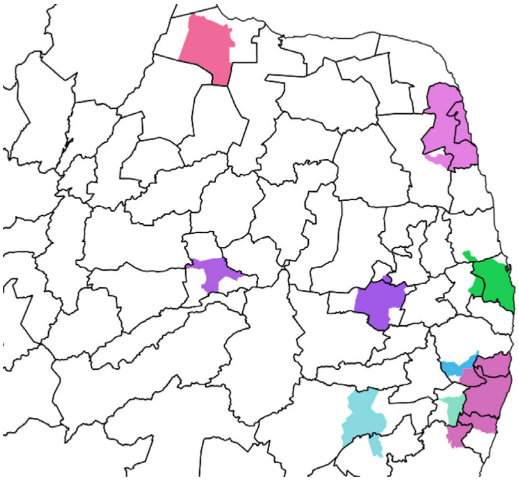


Fig. 4. Brazilian microregions and commuting-based metropolitan areas, 2010. *Notes*: This figure depicts northeastern Brazil, including the states of Rio Grande do Norte and Pernambuco. Microregion boundaries are depicted by black lines. Commuting-based metropolitan areas (population greater than 100,000) are depicted by colored polygons. These metropolitan areas are defined by commuting ties between municipios in 2010, using the Duranton (2015) algorithm with a 10% threshold.

When measured using night-lights–based metropolitan areas, China's city-size distribution is well described by a power law, and this fit is not very sensitive to the light-intensity threshold used to construct the metropolitan areas. Fig. 7 depicts China's city-size distribution for a light-intensity threshold of 30. While the slope coefficient is statistically distinct from the value of −1.0 that defines Zipf's law, the rank-size relationship fits a log-linear power-law specification quite well, with an R^2 of more than 99%.[23] Table 2 shows that this result is relatively invariant to the choice of light-intensity threshold. For threshold values from 10 to 50, the log-linear specification yields an R^2 of 98% of higher. The log-quadratic shape found by Chauvin et al. (2017) seems primarily due to their choice of geographic unit.[24]

Chauvin et al. (2017) suggest that China has a shortage of "ultra-large cities" relative to a power-law distribution, but their use of administrative units plays an important role in this result. The largest metropolitan area produced by our night-lights–based procedure corresponds to the Pearl River Delta, the largest urban area in the world (World Bank Group, 2015, 21). The Pearl River Delta is an administratively fragmented urban area spanning Dongguan, Foshan, Guangzhou, and Shenzhen that has no dominant central city but rather "several original centers that over time merge across boundaries" (World Bank Group, 2015, 36). This multi-jurisdictional urban area, which by its nature does not appear in prefecture-level city data, had about 42 million residents in 2010, and "is a unique kind of settlement in its immense scale as well as its form" (World Bank Group, 2015, 75).

As in China, the Indian city-size distribution looks different when we use metropolitan areas rather than administrative units. The distribution in Fig. 5 depicting the urban populations of Indian districts exhibits curvature, suggesting a log-quadratic rather than log-linear relationship between population size and population rank. Fig. 8 depicts this relationship using urban agglomerations as the geographic units. This distribution is much closer to the expected power-law relationship,

[23] Gabaix and Ioannides (2004) warn against excessive focus on statistically rejecting the null hypothesis of −1.0 and suggest focusing on fit.

[24] To be clear, Zipf's law is an empirical regularity observed in many countries, not a "law" that should serve as the sole criterion for defining metropolitan areas in the Chinese context. The contrast in results shows that how one delineates metropolitan areas is important, but viewed in isolation it would not be a strong reason to prefer our approach.

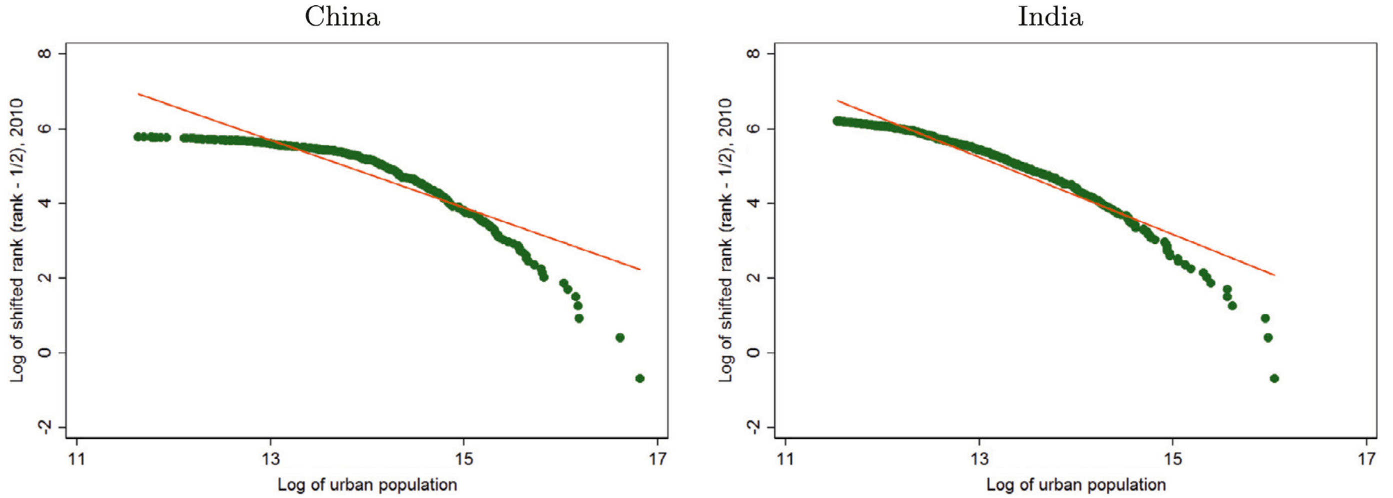


Fig. 5. City-size distributions with administrative units, 2010 (Chauvin et al., 2017).
Notes: These two panels are taken from Fig. 2 in Chauvin et al. (2017). The left panel depicts 326 prefecture-level cities in China (slope −0.91, $R^2 = 0.79$). The right panel depicts 495 districts in India (slope −1.03, $R^2 = 0.92$).

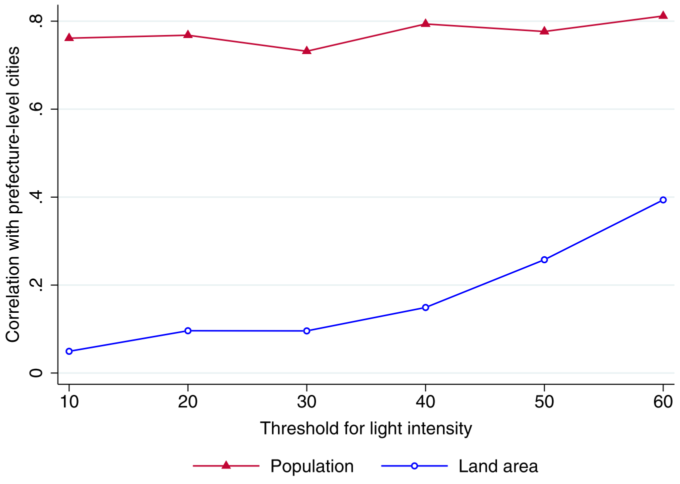
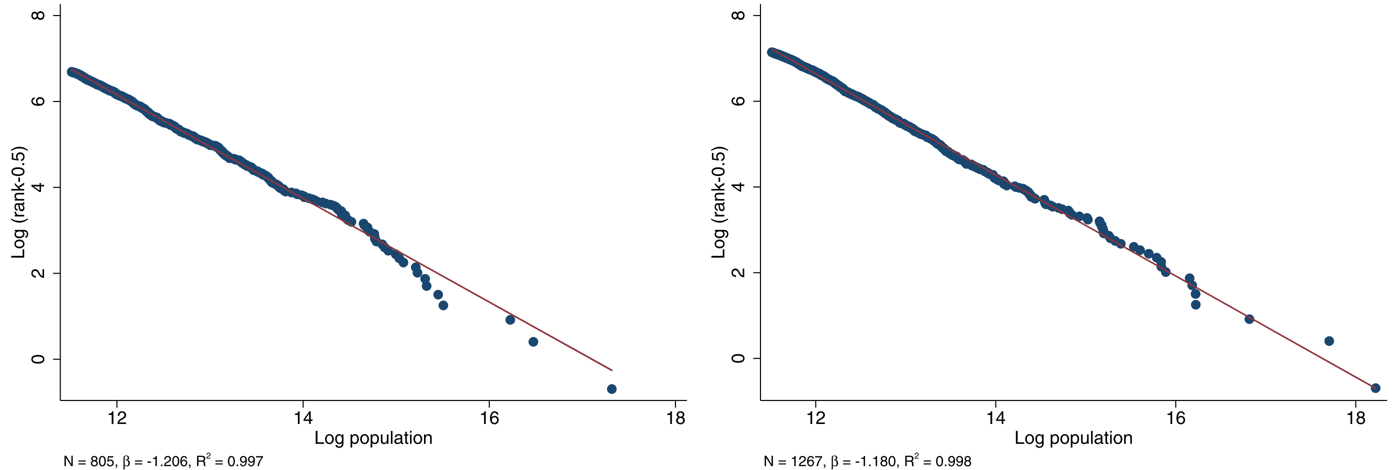


Fig. 6. Comparing Chinese night-lights–based metropolitan areas to prefecture-level cities, 2000.

with the log-linear specification yielding an R^2 greater than 99% in both 2001 and 2011. Aggregating subdistricts' urban populations to define metropolitan areas based on night lights yields similar results, in the sense that the city-size distribution is well characterized by a power-law relationship with a very high R^2. As shown in Table 3, this result is quite stable across a broad range of light-intensity thresholds used to define the metropolitan areas.

In all three developing economies we examine, there are substantial differences between the administrative units typically employed in prior research and the metropolitan areas that we construct on the basis of contiguous lights at night. In Brazil, we find that our night-lights–based method produces metropolitan areas quite similar to those produced by a commuting-flow-based algorithm. For both China and India, the power-law relationship that characterizes developed economies' city-size distribution fits considerably better when we use our night-lights–based approach to build metropolitan areas. Having built these metropolitan areas, we now turn to examining the distribution of skills across and within these metropolitan areas.



N = 805, β = -1.206, R^2 = 0.997
Census 2000. Metropolitan areas defined by aggregating townships based on lights at night.

N = 1267, β = -1.180, R^2 = 0.998
Census 2010. Metropolitan areas defined by aggregating townships based on lights at night.

Fig. 7. China's city-size distribution with night-lights–based units, 2000 and 2010.
Notes: The sample is Chinese metropolitan areas with population greater than 100,000. Metropolitan areas are defined by aggregating townships in areas of contiguous night lights with intensity greater than 30. Left panel depicts 2000; right panel 2010.

Table 2
China's city-size distribution with night-lights–based units, 2000 and 2010.

	2000				2010			
Metropolitan scheme	β	s.e.	R^2	N	β	s.e.	R^2	N
Light intensity 10	−1.175	(0.049)	0.998	1139	−1.033	(0.044)	0.995	1117
Light intensity 20	−1.211	(0.055)	0.996	960	−1.150	(0.045)	0.998	1313
Light intensity 30	−1.206	(0.060)	0.997	805	−1.180	(0.047)	0.998	1267
Light intensity 40	−1.157	(0.067)	0.994	599	−1.163	(0.049)	0.997	1140
Light intensity 50	−1.091	(0.077)	0.989	405	−1.091	(0.052)	0.995	876
Light intensity 60	−0.859	(0.099)	0.945	151	−0.987	(0.066)	0.988	454

Notes: This table reports the coefficient β, standard error, and R^2 from a linear regression of the form
$\ln(\text{rank}_i - 0.5) = \alpha + \beta \ln \text{population}_i + \epsilon_i$
where rank_i is the population rank of metropolitan area i and the standard error is $\sqrt{2/N}|\hat{\beta}|$ (Gabaix and Ibragimov, 2011). The sample for each regression is a set of Chinese metropolitan areas in 2000 or 2010 with population greater than 100,000. Night-lights–based metropolitan areas are defined by aggregating townships in contiguous areas with light intensity exceeding the listed threshold.

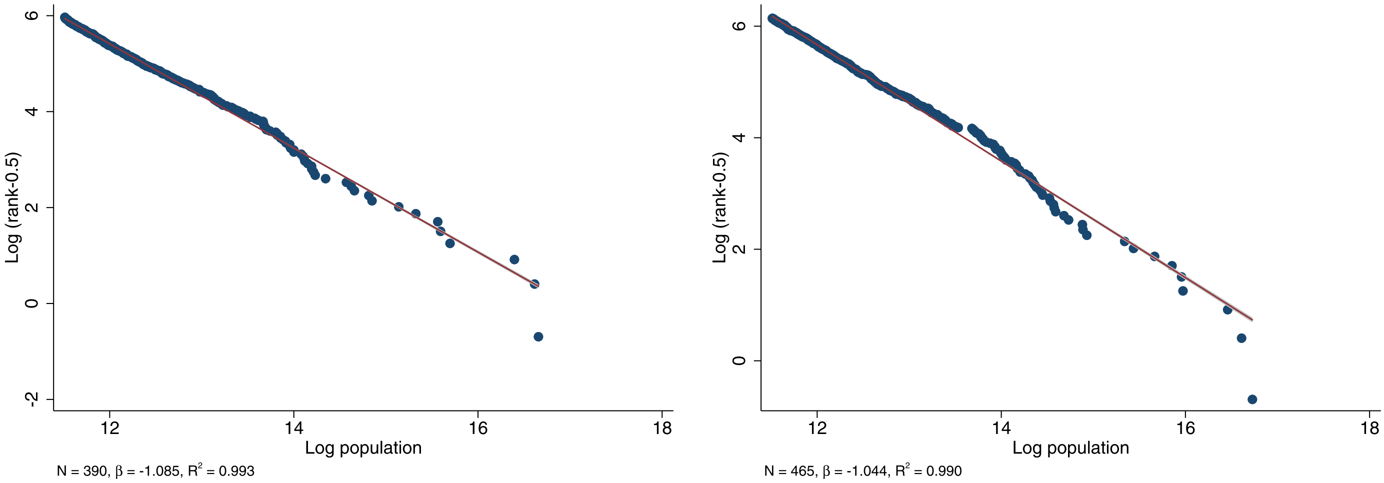


Fig. 8. India's city-size distribution, urban agglomerations, 2001 and 2011.

Table 3
India's city-size distribution, subdistrict-night-lights–based metropolitan areas.

	2001				2011			
Metropolitan scheme	β	s.e.	R^2	N	β	s.e.	R^2	N
Light intensity 10	−1.021	(0.076)	0.992	358	−0.967	(0.072)	0.995	359
Light intensity 20	−1.155	(0.078)	0.994	438	−1.096	(0.070)	0.992	494
Light intensity 30	−1.157	(0.080)	0.994	422	−1.125	(0.071)	0.992	503
Light intensity 40	−1.133	(0.083)	0.993	374	−1.122	(0.072)	0.991	481
Light intensity 50	−1.084	(0.087)	0.984	309	−1.104	(0.075)	0.988	436
Light intensity 60	−1.006	(0.117)	0.943	148	−1.035	(0.089)	0.968	272

Notes: This table reports the coefficient and R^2 from a log-linear rank-size regression, as described in the notes of Table 2. The sample for each regression is a set of Indian metropolitan areas in 2001 or 2011 with population greater than 100,000. Night-lights–based metropolitan areas are defined by aggregating subdistricts in contiguous areas with light intensity exceeding the listed threshold.

3. Skill distributions across metropolitan areas

For each country, we characterize the distribution of skill across metropolitan areas using four categories of educational attainment. Following Davis and Dingel (2017), we regress each skill's log population in a city on that city's log total population to estimate skill-specific "population elasticities." These regressions are of the form

$$\ln L(\nu, c) = \alpha_\nu + \beta_\nu \ln L(c) + \epsilon_{\nu,c}, \tag{1}$$

where $L(\nu, c)$ denotes the number of individuals in city c of skill ν, $L(c)$ is that city's total population, α_ν are fixed effects, and β_ν is skill ν's population elasticity. Where possible, we report population elasticities for a variety of metropolitan-area definitions to assess the sensitivity of cross-metropolitan skill patterns to how metropolitan areas are constructed.

The theoretical model in Davis and Dingel (2017) implies that more skilled groups have higher population elasticities, and our empirical estimates of β_ν are indeed monotonically increasing in skill.[25] In Appendix C.1, we also implement a non-parametric approach proposed in Davis and Dingel (2017) to characterize the distribution of

[25] In the frictionless model of Davis and Dingel (2017), all spatial difference in human capital are due to individuals' locational choices. When migration is costly, spatial variation in the production of human capital may also contribute to these differences.

Table 4
Brazil: Population shares for educational categories, 2010.

Brazil	All	Metro
No schooling	.49	.40
Elementary Graduate	.15	.16
High School Graduate	.25	.29
College Graduate	.11	.15

skill across metropolitan areas. Those results match the conclusion of the population-elasticities approach: larger metropolitan areas are skill-abundant in all three countries.

Our focus on metropolitan-level variation in this section follows a large literature treating cities as the relevant spatial unit for human capital externalities (Moretti, 2004). While the spatial scale of the skill-biased agglomeration economies may be much finer (c.f. Arzaghi and Henderson 2008, Rosenthal and Strange 2008, Ahlfeldt et al. 2015, and Kerr and Kominers 2015), the large volume of research on cross-city variation makes this a natural starting point. We will examine within-metropolitan variation in Section 4.

3.1. Brazil

For Brazil, we construct metropolitan population counts for educational categories by aggregating (with appropriate sampling weights) individual-level observations from the 2010 Census. The four educational categories are "no schooling," "elementary school graduate," "high school graduate," and "college graduate." These four categories are unavoidably unequal in size due to the very large fraction of the population that has no schooling. As reported in Table 4, about half of Brazil's population has no schooling. This number falls to 40% when we restrict attention to metropolitan areas with at least 100,000 residents.

The contrast between the two columns in Table 4 already suggests that metropolitan areas are more skilled, as the difference between the two columns is increasing in educational attainment. Our population elasticity regressions will, effectively, examine variation in population shares within the latter column across metropolitan areas of different sizes.

Table 5 reports population elasticities for these four skill groups for eight different definitions of metropolitan areas. The first three columns use commuting-based metropolitan areas, and the next three use night-lights–based metropolitan areas. The final two columns use the *arranjos populacionais* and microregions defined by the IBGE, the latter being the geographic unit most commonly employed in prior studies of Brazil. A few patterns are immediately evident. Within any column, the order of the population elasticities conforms to the prediction of the model in Davis and Dingel (2017): more skilled groups exhibit higher population elasticities. Comparing across the commuting-based and night-lights–based columns, the estimated elasticities are quite stable. As suggested by the comparisons in Fig. 3, the patterns of economic activity are not sensitive to the threshold employed in defining the metropolitan areas and the two different methods yield metropolitan areas that exhibit similar patterns. The results for *arranjos populacionais* are quite similar. There is a notable contrast between the results for microregions and the first seven columns. These population elasticities are also increasing in skill level, but that variation is considerably larger in magnitude. These values suggest considerably larger difference in skill composition across microregions of different population sizes than across economically integrated metropolitan areas of different sizes. Thus, conclusions about the spatial distribution of skills are sensitive to whether and how we aggregate spatial units.

Fig. 9 relaxes the linear specification employed in Table 5 by plotting a local mean smoother. The population level for each skill group is demeaned, so as to facilitate comparisons across metropolitan areas of different sizes. Plotting each series for commuting-based metropolitan areas with a 10% commuting threshold amounts to a non-parametric version of the log-linear regression slope coefficients reported in the second column of Table 5. The slope at each point of the series is the "local population elasticity." For almost all of the variation, the log-linear approximation fits the data very well. Only at the extreme of the city-size distribution, where there are only two metropolitan areas with population greater than 5 million and thus the local smoother amounts to little more than a data point, does the local smoother deviate considerably from the log-linear approximation. Thus, the first-order approximation appears to be an apt summary of the relationship between metropolitan population size and skill composition in Brazil, as Davis and Dingel (2017) found for US metropolitan areas.[26]

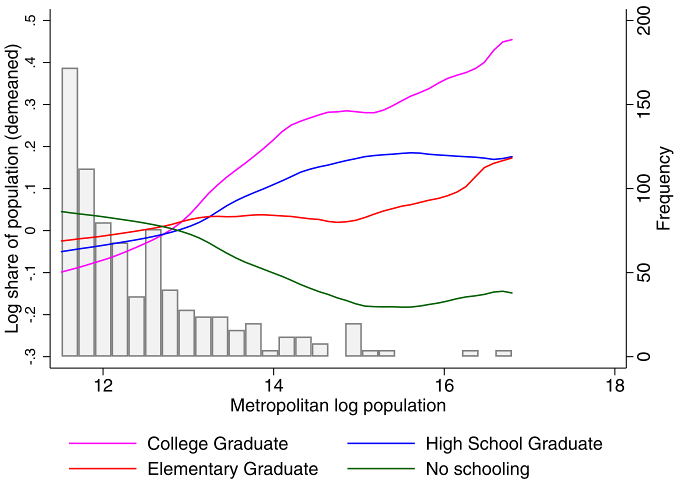


Fig. 9. Brazil: Non-parametric population elasticities for educational categories, 2010.
Notes: Each series plots a local mean smoother using an Epanechnikov kernel. Metropolitan areas are defined by commuting ties between municipios, using Duranton (2015) algorithm with 10% threshold. The histogram bars depict the number of metropolitan areas (on the right vertical axis).

In sum, larger cities are skill-abundant in Brazil, and these differences are exaggerated if one defines cities to be microregions.

3.2. China

For China, we construct metropolitan population counts for educational categories by aggregating township-level tabulations from the 2000 Census. While year-2010 township-level population counts are available, year-2010 data describing educational attainment is currently only available at the county level. As we show below, characterizations of the spatial distribution of skills are sensitive to whether we use metropolitan areas based on aggregating townships or counties.

For China, the four educational categories are "primary school or less," "middle school," "high school," and "college or university." These four categories are unavoidably unequal in size due to the fact that, at the most granular level reported, the primary-school and middle-school categories have the two largest population shares and jointly account for about two-thirds of the metropolitan population, as shown in Table 6. More detail is available for the "college or university" educational levels, but this skill group represents only 4% of China's total population.[27]

Table 7 reports population elasticities for these four skill groups for four different definitions of metropolitan areas. The first three columns

[26] These results are not driven by the spatial distribution of age cohorts. Population elasticities monotonically increase with skill when estimated separately for ages 25–34, 35–44, and 45–54.

[27] While average educational attainment is increasingly rapidly in China, the largest shift from 2000 to 2010 is from primary school to middle school, so the population shares are also quite unequal in size in the 2010 data.

Table 5
Brazil: Population elasticities for educational categories, 2010.

	Commuting			Nightlights			Arranjos	Microregions
	5	15	25	10	30	50	NA	NA
No schooling (β_1)	0.912	0.921	0.914	0.931	0.915	0.912	0.917	0.858
	(0.009)	(0.010)	(0.010)	(0.011)	(0.011)	(0.011)	(0.010)	(0.009)
Elementary graduate (β_2)	1.041	1.033	1.027	1.044	1.034	1.028	1.028	1.115
	(0.010)	(0.011)	(0.011)	(0.012)	(0.011)	(0.011)	(0.011)	(0.013)
High school graduate (β_3)	1.102	1.087	1.086	1.092	1.095	1.096	1.090	1.217
	(0.012)	(0.012)	(0.013)	(0.014)	(0.013)	(0.013)	(0.012)	(0.016)
College graduate (β_4)	1.178	1.168	1.179	1.163	1.173	1.181	1.155	1.302
	(0.024)	(0.025)	(0.026)	(0.027)	(0.026)	(0.026)	(0.023)	(0.027)
Observations	816	768	880	616	676	728	740	1672
Number of metropolitan areas	204	192	220	154	169	182	185	418

Notes: Each column reports OLS estimates of β_v from a regression defined by equation (1). Skill fixed effects α_v are not reported. Standard errors are clustered by geographic unit. Each sample contains geographic units with population greater than 100,000.

Table 6
China: Population shares for educational categories, 2000.

China	All	Metro
Primary school or less	.48	.30
Middle school	.37	.38
High school	.12	.22
College or university	.04	.10

use metropolitan areas obtained by aggregating townships on the basis of night lights, while the fourth column aggregates counties. Using the township-based metropolitan areas, we find that more skilled groups exhibit higher population elasticities. The estimated elasticities are not particularly sensitive to the light-intensity threshold employed to define the metropolitan areas. The differences in population elasticities across skill groups are comparable to those found for Brazil, though this comparison should be tempered by the fact that the educational categories defining the four skill groups are not necessarily comparable across countries.

The population elasticities estimated when employing county-based metropolitan areas differ considerably. First, the elasticities vary much less, as the elasticities for the least- and most-skilled groups are both closer to one. Second, the population elasticities are no longer monotonically increasing in educational attainment: the junior middle school and senior middle school are not statistically distinguishable (and the point estimates are in the "wrong" order). By grouping together both urban and rural areas and possibly grouping together distinct metropolitan areas of different sizes, the county-based metropolitan areas would lead us to substantially understate spatial variation in skill distributions. Since at the moment educational attainment data for 2010 is only available at the county level, we cannot yet reliably characterize spatial variation in skill distributions using the 2010 Census data.

In sum, we find that larger cities are skill-abundant in China when measuring skills using four educational categories. These results are sensitive to the precision of the spatial units used to define metropolitan areas and their characteristics. We find larger differences in population elasticities when building metropolitan areas from more precise geographic units.

3.3. India

For India, we construct metropolitan population counts for educational categories by aggregating town-level tabulations of main workers by educational level from the 2001 Census for the metropolitan areas described in Section 2.5. The four educational categories are "illiterate," "primary," "secondary," and "college graduate." These four populations are of roughly equal size, at least when we restrict attention to urban agglomerations and towns with more than 100,000 residents, as shown in Table 8.

We face some data limitations imposed by the Census of India data describing educational attainment. Educational attainment data are not available at the sub-district level, so we cannot use the metropolitan areas that were produced by aggregating sub-districts on the basis of night lights. We therefore use the definition of metropolitan areas that is the union of urban agglomerations (aggregated across state borders on the basis of night lights) and census towns of sufficient population size. The data source that we employ describes educational attainment

Table 7
China: Population elasticities for educational categories, 2000.

	Township-based			County-based		
Light intensity threshold:	10	30	50	10	30	50
Primary school or less (β_1)	0.914	0.897	0.906	0.938	0.970	0.962
	(0.008)	(0.009)	(0.010)	(0.007)	(0.012)	(0.018)
Middle school (β_2)	1.003	0.985	0.970	1.066	1.017	1.021
	(0.005)	(0.006)	(0.009)	(0.007)	(0.007)	(0.009)
High school (β_3)	1.124	1.096	1.073	1.046	0.992	0.994
	(0.012)	(0.012)	(0.013)	(0.014)	(0.021)	(0.026)
College or university (β_4)	1.344	1.327	1.314	1.140	1.090	1.105
	(0.024)	(0.027)	(0.031)	(0.026)	(0.039)	(0.050)
Observations	4556	3220	1620	6820	4668	2004
Number of metropolitan areas	1139	805	405	1705	1167	501

Notes: Each column reports OLS estimates of β_v from a regression defined by equation (1). Skill fixed effects α_v are not reported. Standard errors are clustered by geographic unit. Each sample contains geographic units with population greater than 100,000.

Table 8
India: Population shares for educational categories, 2001.

India	All	Metro
No education	.43	.22
Primary	.26	.22
Secondary	.24	.36
College graduate	.08	.21

Table 9
India: Population elasticities for educational categories, 2001.

Inclusion threshold:	None	0.8	0.95
No education (β_1)	0.960	0.980	0.970
	(0.027)	(0.018)	(0.027)
Primary (β_2)	0.972	0.987	1.009
	(0.022)	(0.020)	(0.029)
Secondary (β_3)	1.014	1.035	1.049
	(0.018)	(0.015)	(0.022)
College graduate (β_4)	1.027	1.063	1.061
	(0.022)	(0.018)	(0.028)
Observations	1320	1152	808
Number of metropolitan areas	330	288	202

Notes: Each column reports OLS estimates of β_ν from a regression defined by equation (1). Skill fixed effects α_ν are not reported. Standard errors are clustered by geographic unit. Each sample contains the union of urban agglomerations and census towns with population greater than 100,000. Across columns, there is variation in the inclusion threshold, which is the fraction of the urban agglomerations' population for which educational attainment data on constituent components is available.

for constituent components of these metropolitan areas when they are of sufficient population size. We therefore report results for samples that differ in the degree to which we require that the constituent components account for the total population of the metropolitan area.

Table 9 reports population elasticities for the four skill groups for three different definitions of metropolitan areas. The first column includes all metropolitan areas regardless of the fraction of their population covered in the educational-attainment data, while the second and third columns impose minima of 80% and 95%, respectively. In all three columns, skill groups' population elasticities increase with the level of educational attainment. Thus, India's metropolitan areas that are more populous are more skill-abundant, and this finding is robust across various samples that we consider in order to address limitations of the underlying data sources. The range of variation between the least- and most-skilled groups' population elasticities is greater when we restrict the sample to observations with better coverage.

In sum, when we examine whether the population distribution is log-supermodular in skill and metropolitan population, we find that larger cities are indeed skill-abundant in Brazil, China, and India. The quantitative magnitudes of these findings are, in some cases, sensitive to using metropolitan areas defined by contiguity of night lights rather than administrative or political boundaries. Relative to prior work characterizing the spatial distribution of human capital in terms of two skill groups, we show that larger cities are skill-abundant in a high-dimensional sense.

4. Within-metropolitan variation in skills

In this section, we examine how the skill composition within metropolitan areas varies with distance to the city center in Brazil and China.[28] Even in the simplest monocentric city model, this skill gradient is theoretically ambiguous because it depends on the relative income elasticities of housing demand and commuting costs (LeRoy and Sonstelie, 1983). Prior empirical work on patterns of residential sorting within cities has overwhelmingly focused on the United States and European countries (Duranton and Puga, 2015). In the United States, residents near the metropolitan center are typically poorer than suburban residents (Rosenthal and Ross, 2015), a pattern dating to at least 1930 (Lee and Lin, 2018). One potential explanation for this pattern is that poor households locate in central cities to access public transport (Glaeser et al., 2008).

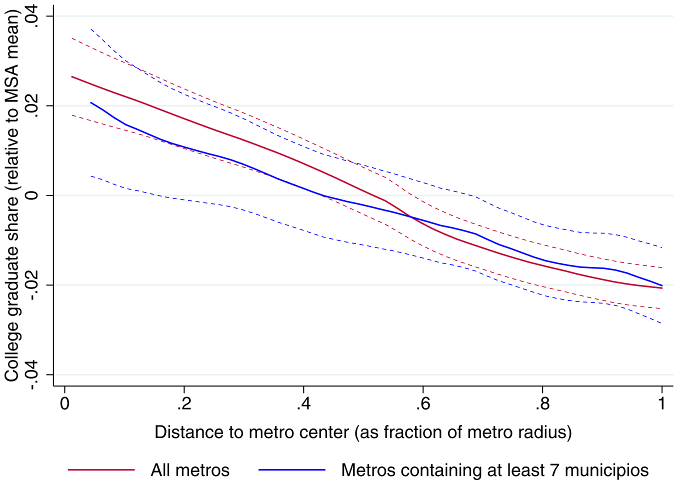


Fig. 10. Skill gradient in Brazilian metropolitan areas, 2010.
Notes: The two series are local mean smoothers of the share of residents who are college graduates in a municipio as a function of distance to metro center using an Epanechnikov kernel. Dashed lines depict 95% confidence intervals. Metropolitan areas are defined by aggregating municipios using a light-intensity threshold of 30.

In both Brazil and China, we find that residents living closer to the center of metropolitan areas are more skilled. For each metropolitan area, we define the city center as the population-weighted average of the latitude-longitude coordinates of its constituent components.[29] We then plot or regress the fraction of residents who are college graduates in each constituent component on its distance from the city center. To make these gradients comparable across metropolitan areas of different size and average skill, we measure the college share relative to the metropolitan mean and distance relative to the most distant constituent component in the metropolitan area.

For Brazil, Table 10 shows that the skill gradient is negative, precisely estimated, and not sensitive to the light-intensity threshold employed to define metropolitan areas. The non-parametric plot of this relationship in Fig. 10 shows that the linear relationship imposed in the regression specification fits the data well. The skill gradient estimated when using microregions is also negative but of substantially smaller magnitude.[30]

The skill gradient is also negative in China, as Table 11 shows. While the skill gradients observed in Chinese metropolitan areas are qualitatively similar whether we use township- or county-level observations of educational attainment, they differ quantitatively. For intermediate

[28] Unfortunately, Indian population counts by educational category are not reported for sufficiently fine geographic units to study this outcome for Indian urban agglomerations.

[29] Alternatively, we could define the city center as the spatial unit within the metropolitan area with the highest population density. Doing so yields qualitatively similar results, but the coefficients are about 20%–30% smaller in absolute magnitude than those reported in Tables 10 and 11.

[30] This is not merely due to the set of microregions containing many more places. Restricting attention to microregions in which the largest municipio is also the largest municipio under the metropolitan areas defined by using a light-intensity threshold of 30 yields an estimated coefficient of −0.050, which is still meaningfully less than −0.0715.

Table 10
Skill gradient in Brazilian metropolitan areas, 2010.

Light threshold	10	20	30	40	50
Panel A: Lights-based metros					
Distance to metro center	−0.0470	−0.0747	−0.0763	−0.0794	−0.0762
	(0.00834)	(0.0121)	(0.0111)	(0.0107)	(0.0129)
Number of municipios	958	519	435	385	330
Number of metropolitan areas	96	86	77	78	71
Microregion *p*-value	0.024	0.000	0.000	0.000	0.000
Microregion clusters	426	425	423	429	430
Arranjo *p*-value	0.000	0.120	0.120	0.180	0.163
Arranjo clusters	142	126	118	119	116
Panel B: Other metro definitions					
		Commuting		Arranjos	Microregions
Threshold	5	15	25	NA	NA
Distance to metro center	−0.0750	−0.100	−0.114	−0.0921	−0.0288
	(0.00348)	(0.00604)	(0.00789)	(0.00506)	(0.00178)
Number of municipios	1074	418	223	557	4750
Number of metropolitan areas	157	84	53	102	417

Notes: The dependent variable is the share of residents who are college graduates in a municipio. Distance to metro center is measured from the municipio centroid to the population-weighted average of constituent-municipio centroids as a share of the greatest distance. The sample is restricted to metropolitan areas containing at least two municipios. Standard errors, clustered by metropolitan area, are in parentheses. The reported *p*-values test the null hypotheses that the coefficients estimated when using night-lights–based metropolitan areas are equal to the coefficients estimated when using arranjos or microregions. The table also reports the number of clusters used when computing those test statistics; see Appendix C.2 for details.

Table 11
Skill gradient in Chinese metropolitan areas, 2000.

Light threshold:	10	20	30	40	50	60
Panel A: Township-based metropolitan areas						
Distance to metro center	−0.110	−0.133	−0.140	−0.144	−0.135	−0.122
	(0.00934)	(0.00712)	(0.00733)	(0.00823)	(0.00944)	(0.0117)
Number of townships	13,900	9672	7761	6224	4775	2584
Number of metropolitan areas	1116	901	720	501	332	135
Panel B: County-based metropolitan areas						
Distance to metro center	−0.0841	−0.0916	−0.0988	−0.0945	−0.0970	−0.107
	(0.0128)	(0.0127)	(0.0132)	(0.0137)	(0.0147)	(0.0171)
Number of counties	992	818	719	663	556	335
Number of metropolitan areas	210	205	186	173	143	79
p-value for difference	0.0386	0.0001	0.0001	0.0000	0.0013	0.0794
Clusters	1128	950	780	578	395	173

Notes: The dependent variable is the share of residents who are college graduates in a constituent component, which is a township in the upper panel and a county in the lower panel. Distance to metro center is measured from the component centroid to the population-weighted average of constituent-component centroids as a share of the greatest such distance in the metropolitan area. The sample is restricted to metropolitan areas containing at least two constituent spatial units. Standard errors, clustered by metropolitan area, in parentheses. The reported *p*-values test the null hypothesis that the coefficients in the two panels within a column are equal. The last line reports the number of clusters used in computing that test statistic. See Appendix C.2 for details.

light-intensity thresholds, the skill gradients obtained from township-based metropolitan areas in the upper panel of Table 11 are about 50% steeper than those obtained from county-based metropolitan areas in the lower panel.

5. Spatial variation in nominal wages across Brazil

In this section, we examine how spatial variation in nominal wages across Brazilian metropolitan areas relates to skills.[31] A very robust finding of the empirical literature is that nominal wages paid to observationally similar workers are higher in more populous and more skilled cities (Moretti, 2004; Combes and Gobillon, 2015). We confirm this finding using our definitions of Brazilian metropolitan areas. A recent finding in developed economies is that the relative price of skill is higher in more populous cities (Baum-Snow and Pavan, 2013; Davis and Dingel, 2019). We also find this pattern in Brazil.

Table 12 shows that nominal wages are higher in Brazilian metropolitan areas with larger populations and a higher share of residents with a college degree. These differences in nominal wages across metropolitan areas control for individual demographics (gender, age, race, and educational attainment). Such findings are usually interpreted as suggesting agglomeration economies and human capital externalities that increase productivity.[32] The estimated coefficients are not sensitive to the thresholds used to define night-lights– and commuting-based metropolitan areas. While the coefficient on log population estimated

[31] Unfortunately, data on wages by educational category are not available for sufficiently fine geographic units to study this outcome for Chinese and Indian metropolitan areas.

[32] Glaeser and Mare (2001) and Rauch (1993) are early influential studies showing that metropolitan population and educational attainment, respectively, are positively correlated with nominal wages after controlling for individual characteristics.

Table 12
Average nominal wages across Brazilian metropolitan areas, 2010.

Light threshold	10	20	30	40	50
Panel A: Lights-based metros					
College graduate share	0.0259	0.0230	0.0222	0.0209	0.0213
	(0.00270)	(0.00283)	(0.00281)	(0.00281)	(0.00276)
Log population	0.0707	0.0710	0.0723	0.0702	0.0688
	(0.00622)	(0.00655)	(0.00674)	(0.00709)	(0.00711)
Observations	139,653	146,460	151,771	160,913	161,936
Full Sample	3,600,519	3,170,748	3,106,347	3,056,954	2,997,105
Number of metropolitan areas	154	162	169	180	182
Panel B: Other metro definitions					
		Commuting		Arranjos	Microregions
Threshold	5	15	25	NA	NA
College graduate share	0.0246	0.0236	0.0164	0.0218	0.0354
	(0.00264)	(0.00273)	(0.00192)	(0.00300)	(0.00173)
Log population	0.0715	0.0694	0.0781	0.0732	0.0769
	(0.00677)	(0.00732)	(0.00809)	(0.00718)	(0.00815)
Observations	183,619	171,273	195,235	166,919	372,908
Number of metropolitan areas	204	192	220	185	418

Notes: The dependent variable is the average nominal hourly wage in a metropolitan-area × gender × age × race × education cell. The college graduate share takes values between 0 and 100. Unreported controls are fixed effects for gender, age, race, and educational attainment. Standard errors, clustered by metropolitan area, in parentheses.

Table 13
Skill premia in Brazilian metropolitan areas, 2010.

Light threshold	10	20	30	40	50
Panel A: Lights-based metros					
Metro log population	0.0399	0.0427	0.0456	0.0501	0.0497
	(0.00760)	(0.00809)	(0.00801)	(0.00827)	(0.00820)
Number of metropolitan areas	154	162	169	180	182
Panel B: Other metro definitions					
		Commuting		Arranjos	Microregions
Threshold	5	15	25	NA	NA
Metro log population	0.0469	0.0446	0.0489	0.0425	0.0517
	(0.00727)	(0.00761)	(0.00782)	(0.00743)	(0.00591)
Number of metropolitan areas	204	192	220	185	418

Notes: The dependent variable is a metropolitan area's difference in average log hourly wages between college graduates and high school graduates.

when using microregions is similar to the estimates obtained when using commuting- and night-lights–based metropolitan definitions, the coefficient on the college graduate share is about 50% greater when estimated using microregions.[33]

Are college wage premia higher in more populous metropolitan areas in Brazil? In the United States, larger cities exhibit both higher relative quantities and higher relative prices of skill, as measured by the share of college graduates and the college wage premium (Davis and Dingel, 2019). The implied greater relative demand for college graduates in larger cities suggested that the productivity benefits of agglomeration are skill-biased.

Table 13 shows that college wage premia are higher in bigger cities in Brazil. We define the college wage premium as the difference in average log hourly wages between college graduates and high school graduates and estimate that its population elasticity is between 4% and 5%. Thus, the productivity benefits of agglomeration economies in Brazil appear to be skill-biased. The population elasticity of the college wage premium in Brazil is larger than the 3% elasticity estimated for US metropolitan areas in 2000 (Davis and Dingel, 2019).

Our estimates in Tables 12 and 13 are quantitatively similar across night-lights–based metropolitan areas, commuting-based metropolitan areas, arranjos populacionais, and microregions. By contrast, our estimates describing spatial variation in the quantities of skill in Tables 5 and 10 are substantially different when we employ microregions rather than our preferred definitions of metropolitan areas. This contrast between spatial variation in quantities and prices of skill in their sensitivity to geographic definitions is interesting, but we do not have reason to believe that this pattern will necessarily generalize to other countries.

6. Sectoral distributions across Brazilian cities

In the theoretical model of Davis and Dingel (2017), larger cities are relatively more skilled, cities' equilibrium productivity differences are Hicks-neutral, and sectors can be ordered by their skill intensity, so larger cities employ relatively more labor in skill-intensive sectors. The results of Section 3 show that larger cities are relatively more skilled in Brazil, China, and India. We now examine whether larger cities are relatively specialized in skill-intensive sectors, using employment levels in both occupations and industries. Due to data limitations, we restrict attention to Brazil.

To characterize the spatial distribution of occupational and industrial employment across Brazilian metropolitan areas, we plot each sector's estimated population elasticity against its skill intensity, measured as the average years of schooling of individuals employed in that sector. Each sector's bubble size is proportionate to its employment share.

Fig. 11 depicts the results of using 10 occupational categories to define sectors. In the left panel, the very low population elasticity of agri-

[33] The coefficient on the college graduate share estimated using microregions is comparable to that estimated by Chauvin et al. (2017, Table 10), who restrict their sample to prime-age males.

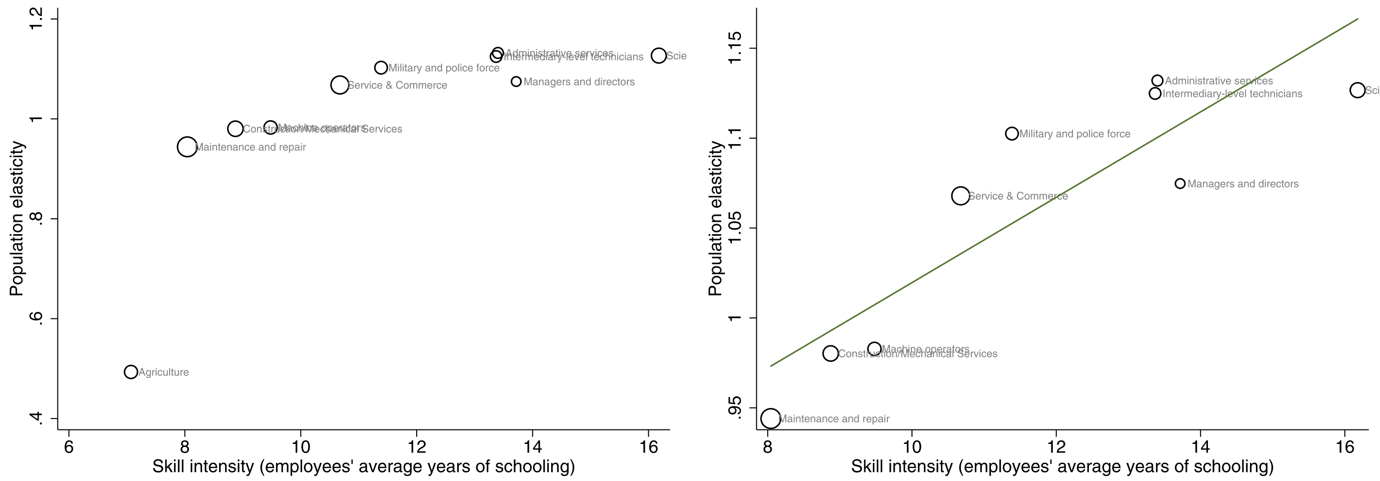


Fig. 11. Brazil: Occupational employment population elasticities, 2010.
Notes: Each observation is an occupational category. The population elasticity of employment is estimated by linear regression. Skill intensity is the average years of schooling of persons employed in that occupational category. Bubble sizes are proportionate to the occupational category's share of employment. Metropolitan areas are defined by commuting flows between municipios, using the Duranton (2015) algorithm with a 10% threshold. Left panel includes all occupations; right panel omits agriculture.

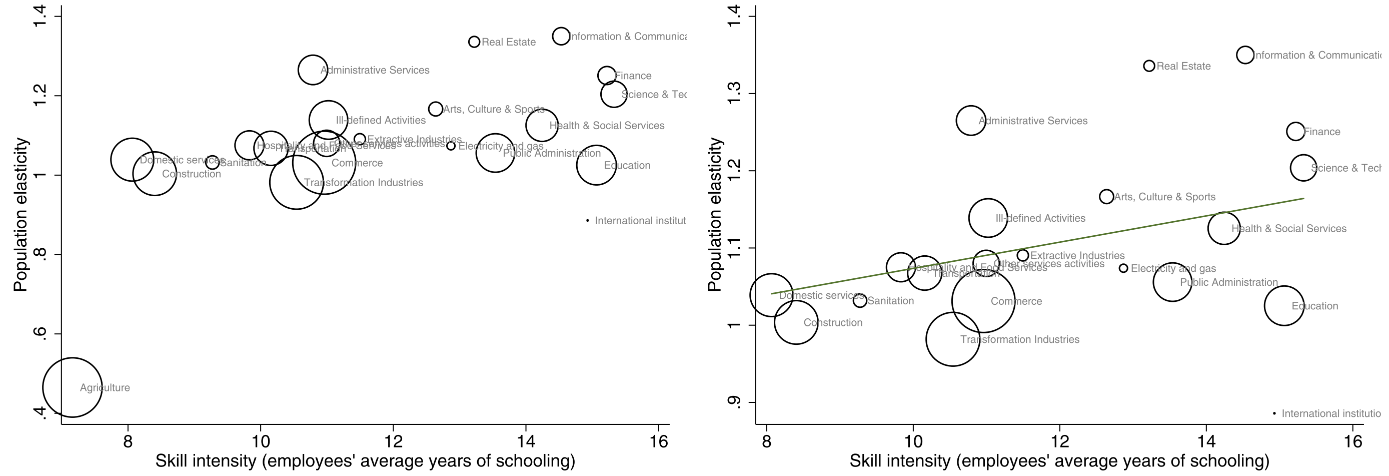


Fig. 12. Brazil: Industrial employment population elasticities, 2010.
Notes: Each observation is an industrial category. The population elasticity of employment is estimated by linear regression. Skill intensity is the average years of schooling of persons employed in that industrial category. Bubble sizes are proportionate to the industrial category's share of employment. Metropolitan areas defined by commuting ties between municipios, using Duranton (2015) algorithm with 10% threshold.

cultural employment masks the rest of the variation depicted, so the right panel omits agriculture and depicts the line of best fit. The model of Davis and Dingel (2017) predicts that the population elasticity of occupational employment should rise with skill intensity and indeed we see a clear positive relationship in Fig. 11.

Fig. 12 depicts the results of using 22 industrial categories to define sectors. Again, we omit agriculture from the right panel in order to better depict the remaining variation across industries. Industrial population elasticities generally increase with skill intensity. The most notable outliers from the central tendency of the data are education (high skilled, low elasticity) and administrative services (low skilled, high elasticity). The fact that the population elasticity of education is quite close to one despite its employment of highly educated individuals may reflect the fact that educational services are typically non-traded. The low skill intensity associated with administrative services as an industry contrasts with the higher average years of schooling associated with administrative services as an occupation.

For both occupations and industries, the estimated population elasticities reveal a broad tendency for more populous metropolitan areas to employ relatively more individuals in skill-intensive sectors. Using alternative commuting thresholds or any of the night-lights–based metropolitan areas delivers very similar elasticity estimates. The population elasticities estimated when using microregions are similarly ordered, but they exhibit substantially greater variation in magnitude, similar to the pattern in Table 5.

7. Conclusion

We study spatial variation in skills in Brazil, China, and India. Our goal is to characterize whether urban systems of developing economies exhibit spatial patterns similar to those found in developed economies, where agglomeration appears to be skill-biased. In order to do so, we aggregate smaller administrative spatial units, such as municipios and townships, belonging to contiguously lit areas in satellite imagery to define metropolitan areas. We intend these metropolitan areas to be comparable to those employed in research studying developed economies. We find that larger cities are more skill-abundant, more skilled residents live closer to the city center, and larger cities exhibit higher skilled wage premia. In short, the productivity benefits of agglomeration appear to be skill-biased in developing economies.

Unlike research designs that employ satellite imagery to both define metropolitan areas and measure economic outcomes, our strategy relies on both satellite imagery and conventional administrative data. Studying the relationship between skills and agglomeration necessarily requires observing individuals' educational attainment (or some other proxy for skills) and cannot be done using satellite imagery alone. Thus, our inquiry is constrained by the availability of data on socioeconomic outcomes for fine geographic units, and this limitation is often binding in India.

Our use of satellite imagery to construct metropolitan areas is preferable to using administrative units that do not correspond to integrated economic entities and alleviates the need for comprehensive commuting data, which are not available in many developing economies. For Brazil and the United States, we find that our night-lights–based metropolitan areas are similar to those defined based on commuting flows. For China and India, where commuting data are not available, our night-lights–based approach eliminates substantial deviations from a power-law distribution for city sizes. Since satellite images cover the entire globe and are becoming available in finer resolutions, our method for defining metropolitan areas should facilitate studies of urbanization and local labor markets in many different contexts.

The spatial patterns that we observe suggest that agglomeration is skill-biased. In all three developing economies, larger cities are skill-abundant. Across four educational categories, the estimated population elasticity monotonically increases with skill. This finding is robust to varying the light-intensity threshold used to define metropolitan areas, but the estimated elasticities can substantially differ from those obtained when using administrative definitions of cities. Where data permit we also study within-metropolitan variation in quantities and across-metropolitan variation in wages and sectoral employment. In Brazil, college wage premia and employment in skill-intensive sectors are relatively greater in more populous cities. These patterns echo recent evidence that agglomeration is skill-biased in the United States and other developed economies.

Appendix A. Defining metropolitan areas

A1. Building metropolitan areas from satellite data

This section provides a few more details of the procedure described in Section 2.1 of the main text.

We extract contour lines for selected light-intensity values from the night lights raster layer and convert those contour lines into polygons. To extract contour lines we project the raster into an azimuthal equidistant projection defined by a latitude and longitude of origin. This projection preserves distance and direction relative to the latitude and longitude of the projection center that we specify for each country. Occasionally these contour lines produce polygons-within-polygons, which can lead to erroneous assignments in subsequent steps. This happens, for instance, when the night lights reveal a sufficiently large (dark) park or lake entirely surrounded by a metropolitan area. We obtain contiguous areas by dissolving these smaller polygons into the larger ones that entirely contain them.

To obtain the intersection of these contiguous area with spatial units for which socioeconomic data is available, we perform a spatial join. Before joining them, we project both the night-light polygons and the administrative spatial units into an Albers equal-area conic projection centered on the same latitude and longitude of origin, specifying standard parallels for each country to minimize distortions. When an administrative spatial unit intersects multiple night-light polygons, we assign the spatial unit to the polygon with which it has the largest area of overlap. Thus, each administrative spatial unit is assigned to one metropolitan area, if any.

A2. Building metropolitan areas from commuting data

This section briefly describes the iterative algorithm introduced by Duranton (2015) to define metropolitan areas on the basis of commuting flows between smaller geographic units, call them "microunits," which are in the US case and municipios in the Brazilian case. Using the algorithm requires the choice of a minimum commuting threshold. We initialize the algorithm by aggregating the two microunits with the largest commuting tie. At each successive iteration of the algorithm, we recompute the commuting flow between any microunit that is not already assigned to a metropolitan area and each metropolitan area. We recursively aggregate microunits to the metropolitan area with which they share the strongest commuting tie that exceeds the minimum commuting threshold. The algorithm stops when there are no more microunits to be aggregated.

Appendix B. Data description

B1. Satellite image data

Night lights raster data is available from NOAA's Earth Observation Group. We use observations from the Version 4 DMSP-OLS Nighttime Lights Time Series for the years 2000, 2001, 2010, and 2011 from the "average visible, stable lights" series.[34] These data have a resolution of 30 arc-seconds, which is roughly one square kilometer.

B2. Brazil

B2.1. Geography

We build metropolitan areas by aggregating municipios. In year 2010 definitions, the 5th, 50th, and 95th percentiles of municipio land area were 83, 416, and 5344 km^2, respectively.

To construct metropolitan areas based on commuting flows, we use anonymized individual-level microdata from the 2010 Census to construct a commuting flows matrix between origin municipio and destination municipio. We then select a commuting share threshold and implement the Duranton (2015) algorithm to construct metropolitan areas.

To construct metropolitan areas based on night lights, we use the night lights raster data described above and shapefiles for Brazilian municipios. We use a spatial coordinate system recommended by the Instituto Brasileiro de Geografia e Estatística (IBGE). Their website recommends a latitude and longitude of origin $(-12°, -54°)$ and standard parallels of −2° and −22°. Our code pulls these parameters directly from https://spatialreference.org/ref/sr-org/7823/. The azimuthal equidistant projection depends only on the latitude and longitude of origin. The Albers conic projection employs this origin and also requires the standard parallels.

B2.2. Skills and sectors

Anonymized individual-level microdata from the 2010 Census is available from the Instituto Brasileiro de Geografia e Estatística (IBGE) website. We aggregate these observations, using the individual sampling weights, to produce municipio-level counts of the population older than 25 by educational attainment, industry, and occupation. We use these same observations to compute average years of schooling by industry and occupation. We compute average hourly wages and skill premia using income and hours at the main job for individuals between 25 and 65 years old with an identified educational attainment and race with wages between the 1st and 99th percentiles.

[34] These are filenames of the form `F1?YYYY_v4?_stable_lights.avg_vis.tif` for F152000, F152001, F182010, and F182011.

Table C.1
Pairwise comparisons for educational categories.

Brazil (2010)				China (2000)				India (2001)			
Bins	Pairings	Success rates		Bins	Pairings	Success rates		Bins	Pairings	Success rates	
2	6	1.00	1.00	2	6	1.00	1.00	2	6	1.00	1.00
8	168	0.91	0.98	5	60	0.88	0.91	5	60	0.78	0.83
16	720	0.82	0.95	10	270	0.86	0.85	11	330	0.65	0.75
64	12,096	0.72	0.88	50	7350	0.77	0.80	33	3168	0.58	0.64
96	27,360	0.68	0.83	150	67,050	0.70	0.73	110	35,970	0.53	0.56
192	110,016	0.63	0.77	805	1,941,660	0.62	0.61	330	323,736	0.52	0.54
Weighted			√				√				√

Notes: The samples contain geographic units with population greater than 100,000: 192 Brazilian metropolitan areas defined by commuting with 10% threshold, 800 Chinese metropolitan areas defined by night lights with 30 intensity threshold, and 330 Indian urban agglomerations and census towns for which educational attainment data are available. Weighted success rates are comparison outcomes weighted by the product of the difference in log population sizes and product of educational category's population shares.

B3. China

B3.1. Geography

We build both county-based and township-based metropolitan areas for the years 2000 and 2010. County- and township-level shapefiles are available via the China Data Center at the University of Michigan. In year 2000 definitions, the 5th, 50th, and 95th percentiles of township land area were 4, 72, and 435 km^2, respectively.

In year 2000 definitions, the 5th, 50th, and 95th percentiles of county land area were 42, 1582, and 11,053 km^2, respectively.

To implement our night-lights–based approach, we use the night lights raster data for 2000 and 2010 described above and apply light-intensity thresholds ranging from 10 to 60 in increments of 10. We use an azimuthal equidistant projection centered on (35°, 105°) and an Albers equal-area conic projection centered on (35°, 105°) with standard parallels 27° and 45°, which is available at https://spatialreference.org/ref/sr-org/china-albers-equal-area-conic/proj4/.

B3.2. Skills

Data on township-level and county-level employment by educational level come from the 2000 and 2010 Population Census. Townships are considerably smaller than counties and therefore preferable where available. Population counts for both counties and townships are available for both 2000 and 2010. However, for the 2010 Census data, many socioeconomic characteristics, such as educational attainment, are thus far only available at the level of counties. The Chinese population census enumerates the *de facto* population of these geographic units, not the *de jure* population of households given by the *hukou* (household registration) system (Chan, 2007, p.392).

B4. India

B4.1. Geography

We define India metropolitan areas using two imperfect methods, due to the absence of a shapefile for India's towns and villages, which we have yet to acquire. First, we use subdistricts, for which a shapefile is available, and aggregate these subdistricts using our night-lights–based approach. In year 2001 definitions, the 5th, 50th, and 95th percentiles of sub-district land area were 92, 374, and 1512 km^2, respectively.

Second, we use urban agglomerations defined by the Census of India. The assignments of census towns to urban agglomerations is available from the Census of India's website. We use an azimuthal equidistant projection centered on (20°, 78°) and an Albers equal-area conic projection centered on (20°, 78°) with standard parallels 28° and 12°, which is available at https://spatialreference.org/ref/sr-org/albers-india/proj4/.

B4.2. Skills

Tables from the 2001 Census are available via the Government of India's website. Town-level employment by educational level is reported in Table B-9, "Main Workers by Educational Level, Age and Sex."

B5. United States

We define United States metropolitan areas by aggregating counties on the basis of lights at night in 2010. We use an equidistant conic projection centered on −96° with standard parallels 29.5° and 45.5° and an Albers equal-area conic projection centered on $(37.5°, -96°)$ with standard parallels 29.5° and 45.5° for the contiguous United States, which is available at https://spatialreference.org/ref/esri/usa-contiguous-albers-equal-area-conic/. We use distinct azimuthal equidistant and Albers equal-area conic projections for Alaska and Hawaii.

Appendix C. Additional empirical results

C1. Pairwise comparisons of skills across metropolitan areas

Davis and Dingel (2017) introduce a theoretical model that predicts that larger cities are skill-abundant and delivers two methods of examining these patterns for an arbitrary number of skill groups. In Section 3, we report population elasticities estimated by linear regression. In this appendix, we report non-parametric pairwise comparisons of relative population levels of any two skills and any two cities, which test the theory's prediction that the more skilled group should be relatively larger in the more populous city.

The pairwise-comparison prediction in Davis and Dingel (2017) says that, if cities are divided into bins ordered by population sizes, then in any pairwise comparison of two bins and two skills, the bin containing more populous cities will have relatively more of the more skilled type.[35] Our pairwise-comparison test therefore reports, among all possible pairs of bins and pairs of skills, the fraction of comparisons in which the population of the more skilled group is relatively larger in the more populous cities. We compare this observed success rate to the null hypothesis that skills are uniformly distributed across cities. When doing

[35] Formally, if $L(\nu, c)$ is log-supermodular, C and C' are distinct sets, C is greater than C' ($\inf_{c\in C} L(c) > \sup_{c'\in C'} L(c')$), and n_C ($n_{C'}$) is the number of elements in C (C'), then

$$\frac{1}{n_C}\sum_{c\in C}\ln L(\nu,c)+\frac{1}{n_{C'}}\sum_{c'\in C'}\ln L(\nu',c')\geq\frac{1}{n_C}\sum_{c\in C}\ln L(\nu',c)+\frac{1}{n_{C'}}\sum_{c'\in C'}\ln L(\nu,c')\ \forall\nu>\nu'.$$

so, we can also weight the comparison outcome by the product of the difference in log population sizes and the product of the educational categories' population shares. Davis and Dingel (2017) show that, in the presence of additive random errors to $\ln L(\nu, c)$, the likelihood of a successful pairwise comparison increases with the difference in population size and the number of cities assigned to each bin. In the interest of brevity, we only report results for one definition of metropolitan areas for each economy.

Table C.1 reports the success rates for these comparisons. In all three economies, the success rate is higher when we use a smaller number of bins or weight the comparisons by population differences. Thus, the central tendency of the data are consistent with the patterns exhibited by the estimated population elasticities. More populous metropolitan areas are more skill-abundant, as captured by four educational-attainment categories.

If we are willing to assume that the comparison of any two educational categories is equally informative across the three economies, we can also compare the success rates across countries to gauge the degree to which larger cities are more skill-abundant. These comparisons are complicated by substantial cross-country differences in the size of geographic units that we aggregate and the number of metropolitan areas with population greater than 100,000. With these caveats in mind, the general pattern is that Brazil's pairwise-comparison success rates are higher than China's, which are higher than India's. Thus, broadly speaking, the distribution of skills that most closely matches the theoretical predictions and US empirical patterns in Davis and Dingel (2017) is that of Brazil, followed by China, and then India. This is similar to the finding of Chauvin et al. (2017), who conclude that, in terms of a variety of spatial patterns, Brazil is more like the US than China, which is more like the US than India. In terms of the spatial distribution of skills, however, we find that all three economies' populations are well described by the stylized fact that larger cities are skill-abundant.

C2. *Comparing skill gradients of different metropolitan definitions*

The p-values reported in Table 11 test the null hypothesis that the skill gradient coefficients estimated using township- and county-based metropolitan areas are equal. We implement this test as a difference-in-differences regression. Let Y_{icg} denote the college share of residents in administrative unit i (either a town or a county) within metropolitan area c in group g. A group g contains a county-based metropolitan area and the township-based metropolitan areas with which it spatially overlaps. We regress the college share Y_{icg} on a metropolitan area fixed effect α_c, the normalized distance to the metropolitan centroid X_{icg}, and the interaction of X_{icg} with a dummy indicating that i is a township.

$$Y_{icg} = \alpha_c + \beta X_{icg} + \gamma X_{icg} \times \text{township}_i + \epsilon_{icg}$$

We allow ϵ_{icg} to be correlated within group g when computing the standard error for γ. The reported p-values test the hypothesis that $\gamma = 0$.

The analogous procedure is employed to produce the p-values reported in Table 10, which test the null hypothesis that the skill gradient coefficients estimated using night-lights–based metropolitan areas are equal to those estimated using arranjos or microregions.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jue.2019.05.005

References

Ades, A.F., Glaeser, E.L., 1995. Trade and circuses: explaining urban giants. Quart. J. Econ. 110 (1), 195–227.

Ahlfeldt, G.M., Redding, S.J., Sturm, D.M., Wolf, N., 2015. The economics of density: evidence from the Berlin wall. Econometrica 83 (6), 2127–2189. doi:10.3982/ECTA10876.

Arzaghi, M., Henderson, J.V., 2008. Networking off madison avenue. Rev. Econ. Stud. 75 (4), 1011–1038. doi:10.1111/j.1467-937X.2008.00499.x.

Bacolod, M., Blum, B.S., Strange, W.C., 2009. Skills in the city. J. Urban Econ. 65 (2), 136–153. doi:10.1016/j.jue.2008.09.003.

Baragwanath Vogel, K., Goldblatt, R., Hanson, G.H., Khandelwal, A.K., 2018. Detecting Urban Markets with Satellite Imagery: An Application to India. Working Paper 24796. National Bureau of Economic Research doi:10.3386/w24796.

Baum-Snow, N., Pavan, R., 2013. Inequality and city size. Rev. Econ. Stat. 95 (5), 1535–1548.

Bleakley, H., Lin, J., 2012. Portage and path dependence. Quart. J. Econ. 127 (2), 587–644.

Bustos, P., Caprettini, B., Ponticelli, J., 2016. Agricultural productivity and structural transformation: evidence from brazil. Am. Econ. Rev. 106 (6), 1320–1365. doi:10.1257/aer.20131061.

Cavalcanti, T., Mata, D.D., Toscani, F.G., 2016. Winning the Oil Lottery; The Impact of Natural Resource Extraction on Growth. IMF Working Papers 16/61. International Monetary Fund.

Chan, K.W., 2007. Misconceptions and complexities in the study of China's cities: definitions, statistics, and implications. Eurasian Geogr. Econ. 48 (4), 383–412. doi:10.2747/1538-7216.48.4.383.

Chan, K.W., 2010. Fundamentals of China's urbanization and policy. China Rev. 10 (1), 63–93.

Chauvin, J.P., 2017. Gender-segmented labor markets and the effects of local demand shocks. Harvard University working paper (available at https://scholar.harvard.edu/files/chauvin/files/jpchauvin_jmp.pdf).

Chauvin, J.P., Glaeser, E., Ma, Y., Tobio, K., 2017. What is different about urbanization in rich and poor countries? cities in Brazil, China, India and the United States. J. Urban Econ. 98, 17–49. doi:10.1016/j.jue.2016.05.003.

Combes, P.-P., Gobillon, L., 2015. The Empirics of agglomeration economies. In: Duranton, G., Henderson, J.V., Strange, W.C. (Eds.), Handbook of Regional and Urban Economics, 5. Elsevier, pp. 247–348. doi:10.1016/B978-0-444-59517-1.00005-2.

Costa, D.L., Kahn, M.E., 2000. Power couples: changes in the locational choice of the college educated. Quart. J. Econ. 115 (4), 1287–1315. doi:10.1162/003355300555079.

Costa, F., Garred, J., Pessoa, J.P., 2016. Winners and losers from a commodities-for-manufactures trade boom. J. Int. Econ. 102, 50–69. doi:10.1016/j.jinteco.2016.04.005.

Couture, V., Handbury, J., 2017. Urban Revival in America, 2000–2010. Working Paper 24084. National Bureau of Economic Research doi:10.3386/w24084.

Davis, D.R., Dingel, J.I., 2017. The Comparative Advantage of Cities. Technical Report.

Davis, D.R., Dingel, J.I., 2019. A spatial knowledge economy. Am. Econ. Rev. 109 (1), 153–170.

Díaz-Lanchas, J., Llano, C., Minondo, A., Requena, F., 2018. Cities export specialization. Appl. Econ. Lett. 25 (1), 38–42. doi:10.1080/13504851.2017.1290784.

Dix-Carneiro, R., Kovak, B.K., 2015. Trade Reform and Regional Dynamics: Evidence From 25 Years of Brazilian Matched Employer-Employee Data. Working Paper 20908. National Bureau of Economic Research doi:10.3386/w20908.

Donaldson, D., Storeygard, A., 2016. The view from above: applications of satellite data in economics. J. Econ. Perspect. 30 (4), 171–198. doi:10.1257/jep.30.4.171.

Duranton, G., 2007. Urban evolutions: the fast, the slow, and the still. Am. Econ. Rev. 97 (1), 197–221. doi:10.1257/aer.97.1.197.

Duranton, G., 2015. Delineating metropolitan areas: Measuring spatial labour market networks through commuting patterns. In: Watanabe, T., Uesugi, I., Ono, A. (Eds.), The Economics of Interfirm Networks. Springer Japan, Tokyo, pp. 107–133. doi:10.1007/978-4-431-55390-8_6.

Duranton, G., Puga, D., 2015. Urban land use. In: Duranton, G., Henderson, J.V., Strange, W.C. (Eds.), Handbook of Regional and Urban Economics, 5. Elsevier, pp. 467–560. doi:10.1016/B978-0-444-59517-1.00008-8.

Findeisen, S., Südekum, J., 2008. Industry churning and the evolution of cities: evidence for germany. J. Urban Econ. 64 (2), 326–339.

Gabaix, X., 1999. Zipf'S law for cities: an explanation. Quart. J. Econ. 114 (3), 739–767.

Gabaix, X., Ibragimov, R., 2011. Rank - 1/2: A Simple way to improve the OLS estimation of tail exponents. J. Bus. Econ. Stat. 29 (1), 24–39. doi:10.1198/jbes.2009.06157.

Gabaix, X., Ioannides, Y., 2004. The evolution of city size distributions. In: Henderson, J.V., Thisse, J.F. (Eds.), Handbook of Regional and Urban Economics, Chapter 53, 4. Elsevier, pp. 2341–2378.

Giannone, E., 2018. Skill-biased technical change and regional convergence. Pennsylvania State University working paper (available at https://sites.google.com/view/elisagiannone/research/working-papers).

Glaeser, E.L., Kahn, M.E., Rappaport, J., 2008. Why do the poor live in cities? the role of public transportation. J. Urban Econ. 63 (1), 1–24. doi:10.1016/j.jue.2006.12.004.

Glaeser, E.L., Mare, D.C., 2001. Cities and skills. J. Labor Econ. 19 (2), 316–342.

Harari, M., 2017. Cities in bad shape: Urban geometry in India. The Wharton School, University of Pennsylvania working paper (available at http://real.wharton.upenn.edu/∼harari/Harari_Papers/CityShapeOct2017.pdf).

Henderson, J.V., 1991. Urban Development: Theory, Fact, and Illusion. Oxford University Press.

Henderson, J.V., 2014. Urbanization and the Geography of Development. Technical Report 6877. World Bank Policy Research Working Paper doi:10.1596/1813-9450-6877.

Henderson, J.V., Storeygard, A., Weil, D.N., 2012. Measuring economic growth from outer space. Am. Econ. Rev. 102 (2), 994–1028. doi:10.1257/aer.102.2.994.

Hoffmann, R., 2003. Inequality in Brazil: the contribution of pensions. Revista Brasileira de Economia 57, 755–773.

Hu, S., Brakman, S., van Marrewijk, C., 2014. Smart Cities are Big Cities - Comparative Advantage in Chinese Cities. CESifo Working Paper Series 5028. CESifo Group Munich.

IBGE, 2016. Arranjos Populacionais e Concentrações Urbanas do Brasil, 2nd ed. IBGE, Rio de Janeiro.

Kerr, W.R., Kominers, S.D., 2015. Agglomerative forces and cluster shapes. Rev. Econ. Stat. 97 (4), 877–899.

Kovak, B.K., 2013. Regional effects of trade reform: what is the correct measure of liberalization? Am. Econ. Rev. 103 (5), 1960–1976.

Lee, S., Li, Q., 2013. Uneven landscapes and city size distributions. J. Urban Econ. 78, 19–29. doi:10.1016/j.jue.2013.05.001.

Lee, S., Lin, J., 2018. Natural amenities, neighbourhood dynamics, and persistence in the spatial distribution of income. Rev. Econ. Stud. 85 (1), 663–694. doi:10.1093/restud/rdx018.

LeRoy, S.F., Sonstelie, J., 1983. Paradise lost and regained: transportation innovation, income, and residential location. J. Urban Econ. 13, 67–89.

Moretti, E., 2004. Human capital externalities in cities. In: Henderson, J.V., Thisse, J.-F. (Eds.), Cities and Geography, 4. Elsevier, pp. 2243–2291. doi:10.1016/S1574-0080(04)80008-7.

Office of Management and Budget, 2010. 2010 standards for delineating metropolitan and micropolitan statistical areas. In: Federal Register, 75, pp. 37246–37252.

Rauch, J.E., 1993. Productivity gains from geographic concentration of human capital: evidence from the cities. J. Urban Econ. 34 (3), 380–400.

Rosenthal, S.S., Ross, S.L., 2015. Change and persistence in the economic status of neighborhoods and cities. In: Duranton, G., Henderson, J.V., Strange, W.C. (Eds.), Handbook of Regional and Urban Economics, 5. Elsevier, pp. 1047–1120. doi:10.1016/B978-0-444-59531-7.00016-8.

Rosenthal, S.S., Strange, W.C., 2008. The attenuation of human capital Spillovers. J. Urban Econ. 64 (2), 373–389. doi:10.1016/j.jue.2008.02.006.

Rozenfeld, H.D., Rybski, D., Gabaix, X., Makse, H.A., 2011. The area and population of cities: new insights from a different perspective on cities. Am. Econ. Rev. 101 (5), 2205–2225. doi:10.1257/aer.101.5.2205.

Scherer, C. E. M., Folch, D. C., 2017. A comparative study of urban occupational structure: Brazil and united states. http://sisconev.com.br/Uploads/ENABER17/Trab01570036772017009_000000.pdf.

Soo, K.T., 2005. Zipf's law for cities: a cross-country investigation. Reg. Sci. Urban Econ. 35 (3), 239–263. doi:10.1016/j.regsciurbeco.2004.04.004.

Storeygard, A., 2016. Farther on down the road: transport costs, trade and urban growth in sub-saharan africa. Rev. Econ. Stud. 83 (3), 1263–1295. doi:10.1093/restud/rdw020.

Welch, R., 1980. Monitoring urban population and energy utilization patterns from satellite data. Remote Sens. Environ. 9 (1), 1–9. doi:10.1016/0034-4257(80)90043-7.

World Bank Group, 2015. East Asia's Changing Urban Landscape : Measuring a Decade of Spatial Growth. World Bank.