# Cities, Skills, and Sectors in Developing Economies*

Donald R. Davis[†]
Columbia and NBER

Jonathan I. Dingel[‡]
Chicago Booth and NBER

Antonio Miscio[§]
BCG

December 31, 2016

PRELIMINARY AND INCOMPLETE

**Abstract**

In developed economies, larger cities are skill-abundant and specialize in skill-intensive activities. This paper characterizes the spatial distributions of skills and sectors in Brazil, China, and India. To facilitate comparisons with developed-economy findings, we construct metropolitan areas from finer geographic units for each economy. We then compare and contrast the spatial distributions of educational attainment, industrial employment, and occupational employment across Brazil, China, India, and the United States.

*We thank Luis Costa, Kevin Dano, Shirley Yarin, and Yue Yuan for excellent research assistance. Dingel thanks the Kathryn and Grant Swick Faculty Research Fund at the University of Chicago Booth School of Business for supporting this work.

[†]drdavis@columbia.edu
[‡]jdingel@chicagobooth.edu
[§]miscio.antonio@bcg.com

# 1 Introduction

This paper studies the distribution of skills across cities of different sizes in three large developing economies: Brazil, China, and India. These three countries jointly account for approximately 40 percent of world 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 developing economies' cities by 2050. While urbanization does not imply growth, the two are nonetheless strongly linked (Henderson, 2014).

In developed economies, larger cities are skill-abundant and specialize in skill-intensive activities. The positive relationship between the fraction of college graduates and metropolitan population has long been documented. In recent work, Davis and Dingel (2014) introduce a theoretical framework characterizing the distribution of heterogeneous skills and sectors across cities of different sizes for more than two skill groups. In their model, the comparative advantage of cities is jointly governed by the comparative advantage of individuals and their locational choices. The key testable predictions are that larger cities are skill-abundant and specialize in skill-intensive activities, which means that more skilled demographic groups and more skill-intensive sectors have higher population elasticities. These are clear features of the data for the United States in 2000.

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 of necessity 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 skill-abundant and specialize in skill-intensive activities. We begin to tackle this question by examining these three large developing economies.

Studying the distribution of skills and sectors across metropolitan areas in Brazil, China, and India necessitates constructing metropolitan areas, which are not readily available in these countries. Economic theory treats a city as a highly – if imperfectly – integrated labor market. For 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 often fragment economically integrated areas into distinct cities or circumscribe places, including rural areas, that are not integrated metropolises. To evaluate whether developing economies exhibit an urban hierarchy of skills and sectors similar to that of developed economies requires an appropriate geography defining cities' sizes and economic characteristics.

In this paper, we develop a methodology 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 a few steps. First, we show that, with appropriately selected light-intensity cutoffs, our nightlights-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 nightlights are available. Third, we show that the application of our nightlights-based approach to China and India eliminates anomalies in their city-size distribution. While spatial units defined by administrative boundaries in these countries seem to deviate from a power-law distribution (Chauvin et al., 2016), our nightlight-based definitions of cities accord much better with this empirical regularity, suggesting an advantage of the geographic units we define.

Using these definitions of metropolitan areas, we aggregate census data to characterize the spatial distributions of skills and sectors following the theoretical lens of Davis and Dingel (2014). That theory, in short, predicts that a linear regression of log skill-group population on log total metropolitan population will yield a larger slope coefficient for a more skilled group. Similarly, more skill-intensive sectors should have higher population elasticities. With further assumptions, the model also predicts that all these elasticities are positive, contrary to models of completely specialized cities such as Henderson (1974).

In all three developing economies, larger cities are skill-abundant. We use four educational categories for each economy and find that population elasticities are monotonically increasing in years of schooling. 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 estimating using the administrative definitions of spatial units that have been commonly used in previous research. In this preliminary and incomplete draft, we find that larger cities specialize in skill-intensive economic activity in Brazil.

Our paper belongs to a growing literature on urbanization in developing economies. Perhaps most closely related are Henderson (1991) and Chauvin et al. (2016), who also focus on urban development in Brazil, China, and India, and Hu et al. (2014). In particular, Chauvin et al. (2016) 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 (2014) for China using administrative spatial units. Our investigation complements these studies by focusing on the spatial distribution of skills and sectors and developing definitions of metropolitan areas that are more comparable to the economically integrated entities studied in research on US cities.

Our nightlights-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 nightlights, 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 is one part of a rapidly expanding economics literature exploiting satellite data, recently surveyed by Donaldson and Storeygard (2016). Most prior research, such as Bleakley and Lin (2012) and Henderson et al. (2012), has utilized nightlights as a proxy for local economic activity at a finer resolution than typically documented in administrative data. Our innovation is to use nightlights as a basis for identifying contiguous areas of economic activity that define metropolitan areas and then characterizing those metropolitan areas' socioeconomic characteristics by aggregating spatial units available in more traditional data sources.

# 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. 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.

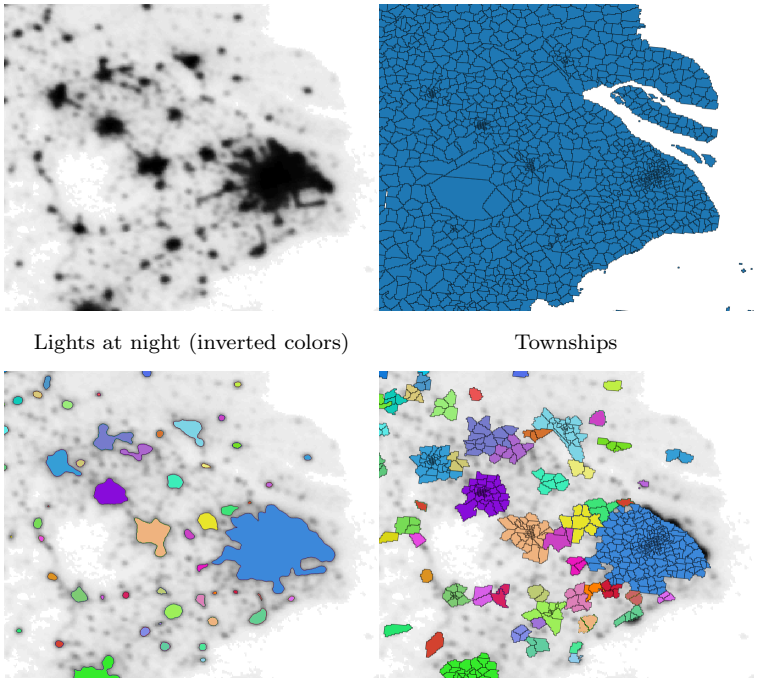
We propose a method for constructing metropolitan areas from smaller geographic units based on night lights data. First, we validate our method by showing that applying it to the United States yields spatial units very similar to those defined by the government statistical agency based on commuting flows. Second, we apply both our nightlights-based method and a commuting-flow method to Brazil, for which both types of data are available, and find that they yield similar outcomes. Third, we construct metropolitan areas for China and India using satellite night lights data, since commuting data are not available in these two countries.

In a number of cases, these metropolitan areas differ from agglomerations 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 nightlights-based metropolitan areas, while Chauvin et al. (2016) have shown substantial deviations from Zipf's law when using administrative units that incorporate substantial rural territories.

## 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. Figure 1 illustrates the procedure for a portion of the eastern coast of China along the East China Sea.

Figure 1: Building metropolitan areas by aggregating smaller units based on lights at night



Lights at night (inverted colors) Townships

Polygons of contiguous light above 30 Township-based metropolitan areas

The starting point is a satellite image of the country at night. Each pixel has a brightness that is reported as an integer between 0 (no light) and 63 (top-coded value). In the upper left panel, we invert the image colors, so that darker pixels correspond to brighter lights at night. Upon selecting a brightness threshold, we identify contiguous areas of light brighter than the selected threshold. This yields polygons, as in the lower left of Figure 1, which uses a brightness threshold of 30. Note that the polygons themselves are formed without reference to administrative boundaries. The largest polygon, corresponds to the city of Shanghai. Our assumption is that contiguity of lights at night is informative about contiguity of economic activity.

The second crucial data source is a shapefile describing the spatial units for which socioeconomic data are available. For China, these are the townships depicted in the upper right panel of Figure 1.

We use the intersection of the nightlights-based polygons and the spatial units to construct metropolitan areas. A township that intersects one nightlight polygon is assigned to that polygon. In the case of multiple intersections, a township is assigned to the nightlight polygon containing the greatest area of the township. The union of the spatial units assigned to a nightlight polygon constitutes a metropolitan area. The lower right of Figure 1 depicts the metropolitan areas that result from applying our procedure to Chinese townships.

Finally, we impose a minimum population size to include a metropolitan area in our economic analysis. Following the literature (e.g. Chauvin et al. 2016), 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.

## 2.2 US metropolitan areas

While this paper is about the spatial distribution of economic activity in developing economies, we use the United States as a testing ground to validate the nightlights-based methodology we develop to construct metropolitan areas in the absence of commuting data. In the United States, metropolitan statistical areas (MSAs) are defined by the Office of Management and Budget (OMB). In its most recent (2010) standard the OMB aggregates US counties that meet certain requirements into a set of core-based statistical areas (CBSAs), which are designed metropolitan and 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. Counties that don't meet these requirements are not included in any CBSA.[1]

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 methodology 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) methodology, which we will apply to Brazil, produces metropolitan areas similar to those defined by the OMB. We aggregate US counties into metropolitan areas on the basis of county-to-county commuting flows reported in the 2009-2013 American Community Survey.[2]

Our nightlights-based methodology is a departure from these commuting-based methods. When we apply our nightlights-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 metropolitan statistical areas. 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 urban agglomerations defined by the alternative methods based on commuting flows and night lights.[3] We then compare log population and log land area across agglomeration schemes, a comparison made by Rozenfeld et al. (2011) to validate their methodology.

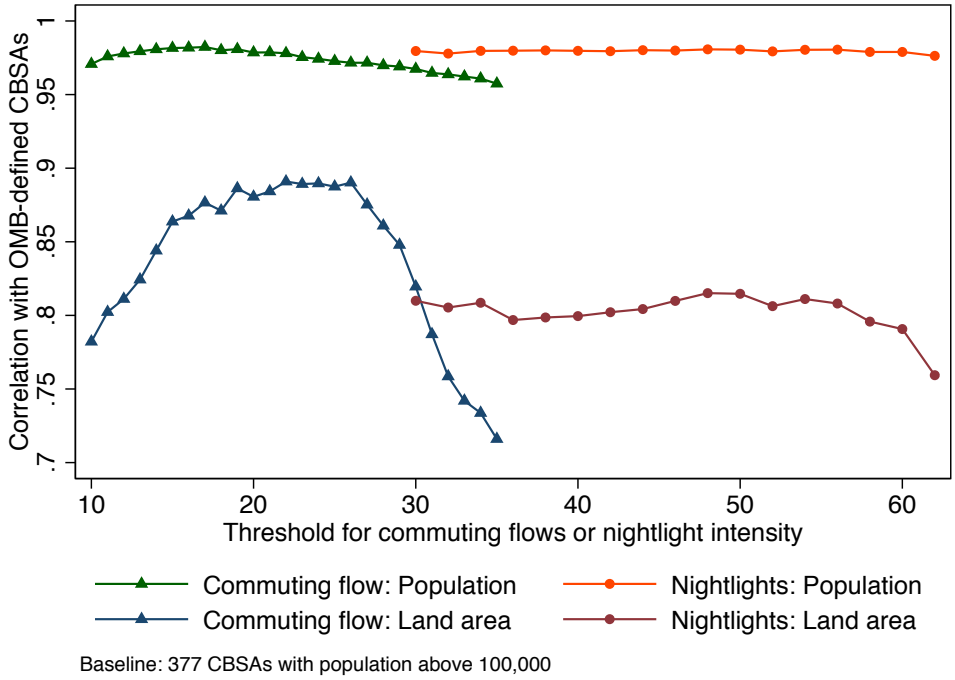
Figure 2 shows that the correlation of log population between CBSAs and their nightlights-based counterparts is about 0.98 and relatively insensitive to the choice of nightlight intensity threshold. Similarly, the correlation of log population between CBSAs and their commuting-flow-based counterparts exceeds 0.96 and varies little with the minimum commuting threshold used in the Duranton (2015) algorithm. There are larger discrepancies in terms of land

[1] An outlying county is aggregated into a CBSA if either of the following criteria is met: (i) at least 25 percent of the workers living in the outlying county work in the CBSA core; (ii) at least 25 percent 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.

[2] 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 percent was deemed appropriate for Colombian municipios (median land area 288 square km), 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 square km). We report the results of constructing metropolitan areas using a range of commuting thresholds.

[3] This matching is not one-to-one in all cases. For example, two CBSAs may be merged into a single polygon based on nightlights. In these cases, we follow Rozenfeld et al. (2011) and select the corresponding CBSA with the largest population.

Figure 2: Comparing population and land area across US metropolitan area definitions




area, where the correlations average about 0.8. 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 nightlights-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 both commuting data and nightlight data available, such that we can implement more than one approach to defining metropolitan areas. We will use this setting to compare our nightlights-based approach to the results of the approach based on commuting flows in a developing-economy context. Validating our nightlights-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 for statistical purposes and do not constitute au-

tonomous political or administrative entities. The IBGE defines microregions according to shared forms of economic activity but not explicitly on the basis of commuting.[4] Our commuting-based and nightlights-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 three different geographic units. First, a number of papers have used microregions as the unit of analysis.[5] We will compare and contrast microregions with our commuting- and nightlights-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, a less frequent approach has employed definitions of metropolitan areas that the states themselves have developed.[6] These are known as *Regiões metropolitanas*. This has three problems. The first, again, is that agglomerations may cross states border 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 those 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) methodology to 2010 Brazilian Census data on commuting flows between municipios.[7] 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 percent of the local working population.[8] We work with

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

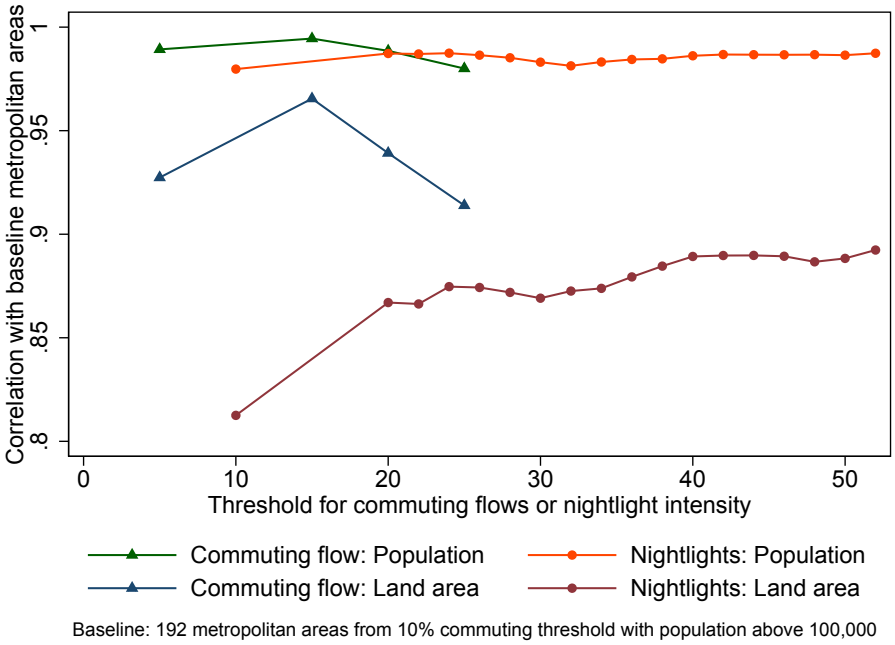
[5] See for instance Kovak (2013); Dix-Carneiro and Kovak (2015); Costa et al. (2016); Chauvin et al. (2016).

[6] 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).

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

[8] As argued in Duranton (2015), the choice of the minimum commuting threshold depends on the size of the underlying units to be aggregated as well as the level of development and the quality of transportation systems. In Colombia, where the median municipio is 288 square km and nominal GDP per capita was $6,000 in 2015, Duranton's preferred threshold to aggregate municipios is 10 percent. In Brazil, the larger size of the median municipio (422 square km) should lead us to choose a lower threshold. However, Brazil's higher GDP per capita ($8,600 in 2015) is a factor pushing us in the opposite direction. Hence, we chose to maintain Duranton's preferred threshold of 10 percent for Brazil.

Figure 3: Comparing population and land area across Brazilian metropolitan area definitions




metropolitan areas with a minimum population of 100,000.[9]

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 nightlights-based methods produce quite similar metropolitan areas. Taking the 10% commuting threshold as our preferred specification, we compare metropolitan areas defined by nightlights and alternative commuting thresholds in terms of the correlation of log population and log land area. As Figure 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 nightlights-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, since 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 key result of our comparison of Brazilian metropolitan areas constructed on the basis of commuting and satellite data is their similarity. There is a close correspondence

[9]With the 10-percent threshold, we obtain 4,807 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.

between the preferred approach based on commuting data, which we will use as our baseline definition in our work on Brazil, and the nightlights-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 represent 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. In addition to these geographic units, there are administrative definitions of "cities" that are used in some distributions of data and thus have been relied upon by researchers. First is a set of four very large "provincial cities" – Beijing, Shanghai, Chongqing, and Tianjin – that may spread across multiple counties. Second is a set of "prefecture-level cities," which are the administrative capital of the prefecture, typically the largest city there, and which may include one or more counties. Sometimes these prefecture-level cities will be listed as a single county in a list of counties; other times, the constituent counties will be disaggregated. Unfortunately one doesn't always know which is at work. Additionally, a distinction is sometimes drawn between urban and rural counties, where the former are labeled "districts" and only the rural are termed "counties."

The spatial unit corresponding to "cities" most commonly used in prior research on Chinese urbanization has been the collection of provincial-level cities and prefecture-level cities. These offer one huge advantage – namely, administrative cities are often the most conveniently available data (and in early periods may be the only form available). Yet there are large downsides. The first is that, because prefectures differ dramatically in population, the set of provincial- and prefecture-level cities will include some very small prefecture-level cities and exclude some very large cities that lack the prefecture-level designation. Second, since counties are a partition of the country, many prefecture-level cities will have substantial rural areas included, as well as distinct urban areas not necessarily economically integrated with the prefecture-level city. 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; 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 nightlights 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 Brazilian results that showed similar results when applying commuting-based and nightlights-based methods to a developing economy, we apply the nightlights-based approach to China. We build

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

| | Correlation with township-nightlight-based | | | | | |
|---|---|---|---|---|---|---|
| | Intensity: 10 | | Intensity: 30 | | Intensity: 50 | |
| MSA scheme | Pop'n | Land | Pop'n | Land | Pop'n | Land |
| County-based nightlights intensity 10 | 0.74 | 0.39 | 0.62 | 0.30 | 0.60 | 0.22 |
| County-based nightlights intensity 20 | 0.62 | 0.22 | 0.62 | 0.25 | 0.68 | 0.29 |
| County-based nightlights intensity 30 | 0.58 | 0.17 | 0.65 | 0.21 | 0.74 | 0.26 |
| County-based nightlights intensity 40 | 0.60 | 0.06 | 0.66 | 0.16 | 0.74 | 0.24 |
| County-based nightlights intensity 50 | 0.64 | 0.07 | 0.74 | 0.11 | 0.72 | 0.21 |
| County-based nightlights intensity 60 | 0.79 | 0.12 | 0.85 | 0.27 | 0.78 | 0.22 |
| Township-based nightlights intensity 10 | | | 0.80 | 0.59 | 0.72 | 0.50 |
| Township-based nightlights intensity 20 | 0.87 | 0.75 | 0.91 | 0.81 | 0.82 | 0.67 |
| Township-based nightlights intensity 30 | 0.80 | 0.63 | | | 0.92 | 0.79 |
| Township-based nightlights intensity 40 | 0.78 | 0.51 | 0.95 | 0.82 | 0.97 | 0.86 |
| Township-based nightlights intensity 50 | 0.76 | 0.50 | 0.93 | 0.77 | | |
| Township-based nightlights intensity 60 | 0.86 | 0.66 | 0.96 | 0.63 | 0.96 | 0.68 |

NOTES: Each cell reports the correlation coefficient for log population or log land area between the MSA scheme identified in the row and the MSA scheme identified in the column pairs for China in 2000.

metropolitan areas by aggregating counties or townships. 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 available for many socioeconomic characteristics of interest.

There are substantial difference between metropolitan areas obtained by aggregating townships and those obtained by aggregating counties. Table 1 illustrates these difference in one dimension, 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. The results in Table 1 suggest that metropolitan characteristics should not be strongly sensitive to the choice of nightlight 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).[10] The Census of India designates 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".[11] An "urban agglomeration" (UA) is one or more physically contiguous towns with at least 20,000 residents. There were 384 UAs in 2001 and 475 UAs in 2011. Urban agglomerations are contiguous areas that may span 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".

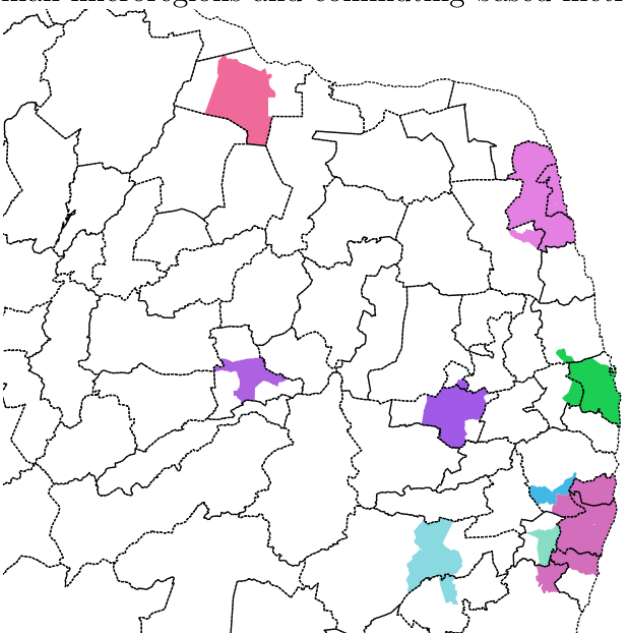
In this preliminary draft, we consider two different methodologies for defining Indian metropolitan areas, each imperfect in some respects. The first is to apply our nightlights-based methodology to the urban populations of subdistricts, the finest spatial unit for which both a geographic shapefile and socioeconomic characteristics are publicly available. Unfortunately, only a limited set of socioeconomic characteristics are reported for subdistricts. The second is to use administratively defined urban agglomerations and cities, agglomerated across state borders on the basis of nightlights.[12] Socioeconomic characteristics are available for urban agglomerations' component census towns of population greater than 100,000. Going forward, we hope to obtain a shapefile for all census towns and villages, which would allow us to apply our nightlights-based approach to a geographically fine administrative unit for which population counts are available.

[10] 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".

[11] In 2011, a "census town" was a place with population greater than 5,000 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.

[12] To aggregate UAs and towns across state borders, we visually inspect the polygons of each to detect whether it belongs to contiguous urban areas spanning state boundaries. This procedure led us to merge the following UA and towns: the National Capital Territory of Delhi, seven UAs and towns in Haryana, and 8 towns in Uttar Pradesh were merged to form Greater Delhi; the towns of Chandigarh (Chandigarh), Mohali (Punjab), Panchkula (Haryana) were merged to form Chandigarh Tricity. Our approach to merge the constituent UA and Towns of Greater Delhi into a single spatial unit is consistent with the method in http://www.prb.org/Publications/Articles/2007/delhi.aspx. As mentioned in the text, Chandigarh Tricity is a single urban area that spans three states.

Figure 4: Brazilian microregions and commuting-based metropolitan areas



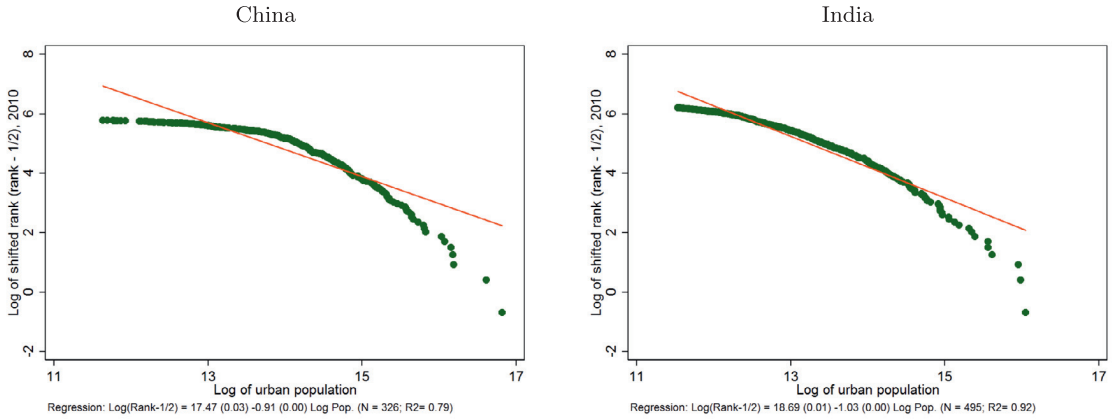
NOTES: This figure depicts northeastern Brazil, including the states of Rio Grande de Notre and Pernambuco. Commuting-based metropolitan areas (population > 100,000) are color-coded. Microregion boundaries are represented by dashed lines. Metropolitan areas defined by commuting ties between municipios, using Duranton (2015) algorithm with 10% threshold.

## 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. As it turns out, 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, Figure 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 an 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 popula-

Figure 5: City-size distributions with administrative units, 2010 (Chauvin *et al* 2016)




NOTES: These two panels are the lower half of Figure 2 in Chauvin et al. (2016).

tion 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 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 nightlights-based metropolitan areas for Brazil, the contrasts between our nightlights-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. (2016) find that the Chinese city-size distribution is poorly described by Zipf's law when using prefecture-level cities. The left panel of Figure 5, taken from their work, shows a rank-size relationship that is more log-quadratic than log-linear. They describe this results as finding that "China has fewer ultra-large cities than the U.S. 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 nights at light. 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. Figure 6 reports the correlation of log population and log land area between metropolitan areas defined at various nightlight-intensity threshold and their prefecture-level-city counterparts. The correlation for log population never exceeds

Figure 6: Comparing nightlights-based metropolitan areas to prefecture-level cities, 2000

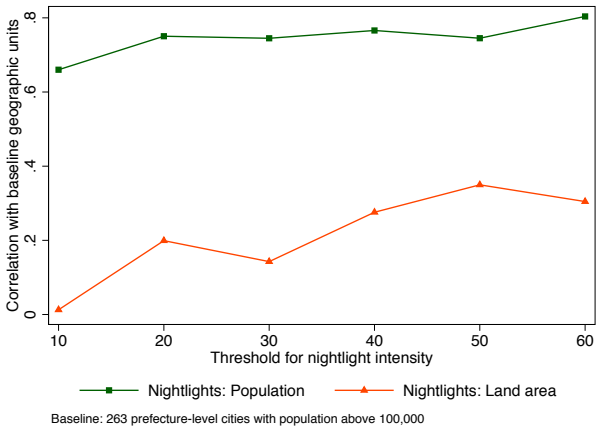
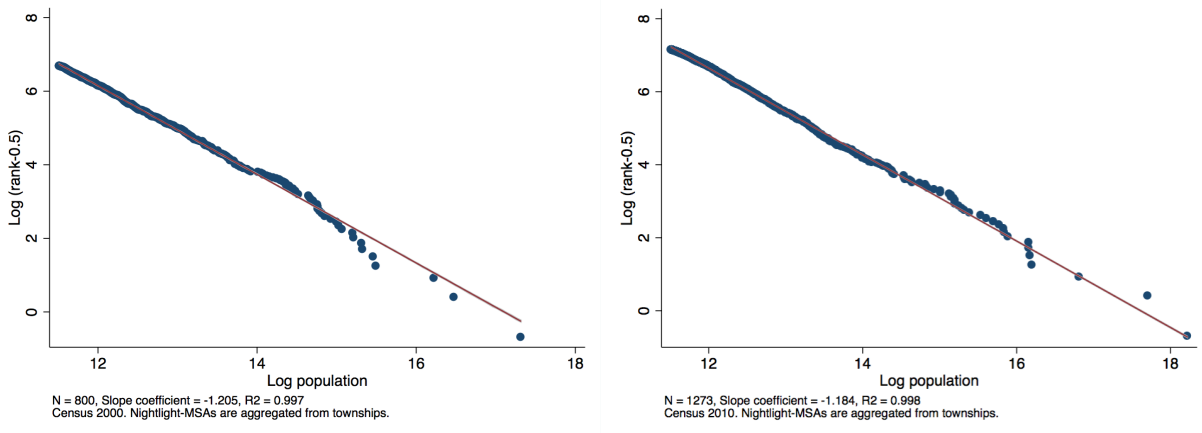



Figure 7: China's city-size distribution with nightlights-based units, 2000 and 2010




NOTES: Sample is Chinese metropolitan areas with population greater than 100,000. Metropolitan areas defined by aggregating townships in areas of contiguous nightlights with intensity greater than 30. Left panel depicts 2000; right panel 2010.

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.

When measured using nightlights-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. Figure 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%. 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. (2016) seems primarily due to their choice of geographic unit.

Table 2: China's city-size distribution with nightlights-based units, 2000 and 2010

| | 2000 | | | 2010 | | |
|---|---|---|---|---|---|---|
| MSA scheme | Zipf $\beta$ | Zipf $R^2$ | $N$ | Zipf $\beta$ | Zipf $R^2$ | $N$ |
| Nightlight intensity 60 | -.856 | .93 | 132 | -.978 | .986 | 418 |
| Nightlight intensity 50 | -1.082 | .988 | 380 | -1.079 | .994 | 852 |
| Nightlight intensity 40 | -1.15 | .994 | 591 | -1.162 | .997 | 1135 |
| Nightlight intensity 30 | -1.205 | .997 | 800 | -1.184 | .998 | 1273 |
| Nightlight intensity 20 | -1.21 | .996 | 959 | -1.159 | .998 | 1334 |
| Nightlight intensity 10 | -1.183 | .997 | 1151 | -1.048 | .995 | 1166 |

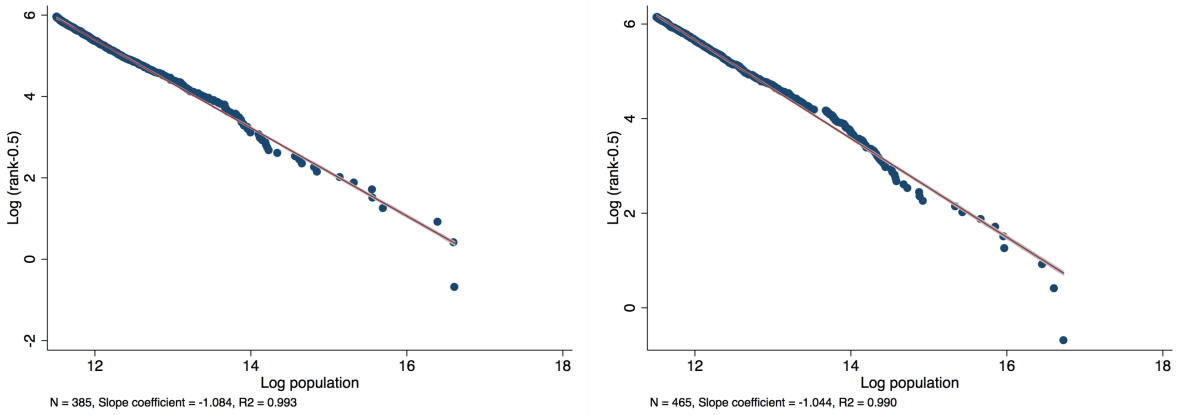
Table 3: India's city-size distribution, subdistrict-nightlights-based metropolitan areas

| | 2001 | | | 2011 | | |
|---|---|---|---|---|---|---|
| MSA scheme | Zipf $\beta$ | Zipf $R^2$ | $N$ | Zipf $\beta$ | Zipf $R^2$ | $N$ |
| Nightlight intensity 60 | -.985 | .938 | 143 | -1.025 | .961 | 250 |
| Nightlight intensity 50 | -1.081 | .984 | 309 | -1.099 | .987 | 441 |
| Nightlight intensity 40 | -1.135 | .993 | 378 | -1.121 | .991 | 491 |
| Nightlight intensity 30 | -1.155 | .994 | 419 | -1.125 | .992 | 506 |
| Nightlight intensity 20 | -1.154 | .994 | 436 | -1.063 | .992 | 472 |
| Nightlight intensity 10 | -1.045 | .992 | 365 | -.947 | .996 | 315 |

Chauvin et al. (2016) 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 nightlights-based procedure corresponds to the Pearl River Delta, the largest urban area in the world (World Bank Group, 2015, p.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, p.36). This multi-jursidictional 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, p.75).

As in China, the Indian city-size distribution looks different when we use metropolitan areas rather than administrative units. The distribution in Figure 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. Figure 8 depicts this relationship using urban agglomerations as the geographic units. This distribution is much closer to the expected power-law relationship, 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 nightlights 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 nightlight intensity thresholds used to define the metropolitan areas.

Figure 8: India's city-size distribution, urban agglomerations, 2001 and 2011




In all three developing economics 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 nightlights-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 nightlights-based methodology to build metropolitan areas. Having built these metropolitan areas, we now turn to examining the distribution of skills and sectors across these metropolitan areas.

# 3 Empirical approach

Davis and Dingel (2014) introduce a general-equilibrium, spatial-equilibrium theory in which the comparative advantage of cities is jointly governed by the comparative advantage of individuals and their locational choices. The model has a few key ingredients. First, locations within cities exhibit heterogeneity in their desirability. Agglomeration causes larger cities to have higher TFP, which makes a location within a larger city more attractive than a location of the same innate desirability within a smaller city. Second, there is a complementarity between individual income and locational attractiveness, so more skilled individuals are more willing to pay for more attractive locations and occupy these locations in equilibrium. Third, comparative advantage causes more skilled individuals to work in more skill-intensive sectors. Under modest assumptions about the primitive distribution of locational desirability within each city, Davis and Dingel (2014) obtain the prediction that larger cities are skill-abundant and specialize in skill-intensive economic activities, in that the skill and sectoral employment distributions exhibit monotone likelihood ratio dominance when ordered by city size. Under slightly stronger assumptions, larger cities will be absolutely larger in all sectors. These predictions distinguish the theory from prior theories in which cities are either completely specialized "industry towns" or perfectly diversified hosts of all economic activities.

Davis and Dingel (2014) introduce two means of taking these predictions to the data. First, linear regressions of each skill or sector's log employment in a city on that city's log

population yields "population elasticities". The theory predicts the order of these population elasticities, and its absolute-size prediction says that all these elasticities are positive. Second, non-parametric pairwise comparisons of relative employment levels of any two skills and any two cities are predicted by the theory: the more skilled group (or more skill-intensive sector) should be relatively larger in the more populous city. Both of these methods shows that the theory is an apt description of the pattern of specialization across US metropolitan areas in 2000.

Our aim in this section is twofold. First, we summarize the empirical methods that will be used to assess these predictions for Brazil, China, and India. Second, we describe the data on skills and sectoral employment for each country that employ to implement the empirical approach.

## 3.1 Methods

Let $f(\nu, c)$ denote the number of individuals in city $c$ of skill $\nu$ or employed in sector $\nu$. If we order $\nu$ and $c$ such that $\nu$ is increasing in skill level or skill intensity and $c$ is increasing with city population size, the prediction of Davis and Dingel (2014) is that $f(\nu, c)$ is a log-supermodular function. If $f(\nu, c)$ is log-supermodular, then

1. a linear regression $\ln f(\nu, c) = \alpha_\nu + \beta_\nu \ln L(c) + \epsilon_{\nu,c}$ in which $\alpha_\nu$ are fixed effects and $L(c)$ is city population yields $\beta_\nu \geq \beta_{\nu'} \iff \nu \geq \nu'$;

2. if $\mathcal{C}$ and $\mathcal{C}'$ are distinct sets and $\mathcal{C}$ is greater than $\mathcal{C}'$ ($\inf_{c \in \mathcal{C}} L(c) > \sup_{c' \in \mathcal{C}'} L(c')$) and $n_{\mathcal{C}}$ ($n_{\mathcal{C}'}$) is the number of elements in $\mathcal{C}$ ($\mathcal{C}'$),

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

Davis and Dingel (2014) refer to the first implication as a "population elasticity" comparison and the second implication as a "pairwise comparison". Standard econometric tests are available to assess whether estimated population elasticities exhibit the property that $\nu \geq \nu' \Rightarrow \beta_\nu \geq \beta_{\nu'}$. To summarize the pairwise comparisons, Davis and Dingel (2014) report the fraction of pairwise inequalities matching the predicted sign, weighted by the product of the two cities' difference in log population and two sectors' difference in skill intensity, and compare this observed success rate to the null hypothesis that skills and sectors are uniformly distributed across cities.[13] These two empirical tests are not independent, since they are both implied by log-supermodularity. Success of one comparison implies success of the other, to the extent that the first-order approximations of $\ln f(\nu, c)$ fit the data well.

Implementing these empirical comparisons requires suitable observations of $f(\nu, c)$. This involves defining metropolitan areas $\{c\}$ and skill groups and sectors $\{\nu\}$in such a manner that publicly available data from Brazil, China, and India provide comprehensive coverage. The next subsection describes the data we employ in each country.

[13] Davis and Dingel (2014) show that, in the presence of additive random errors to $\ln f(\nu, c)$, the likelihood of a successful pairwise comparison increases with the difference in population size, the difference in skill (intensity), and the number of cities assigned to each bin.

### 3.2 Data

For Brazil, we use 2010 Census microdata. These microdata describe anonymized individuals in terms of many characteristics. In addition to the commuting information that we use to construct metropolitan areas, we use the individual-level information on age, level of education, industry, occupation, and municipio. We construct population counts for educational categories and employment counts for industries and occupations for the metropolitan areas defined in Section 2.3 by aggregating individual observations (with appropriate sampling weights). We construct industries' and occupations' skill intensities by computing the average years of schooling of those individual employed in them.

For China, we use 2000 Census data, since China has not yet made 2010 Census data available at the level of townships. There are township-level tabulations that enumerate a townships' population in terms of educational level, industrial employment, and occupational employment. We construct these population counts for the metropolitan areas described in Section 2.4 by summing over the constituent townships. To obtain industries' and occupations' skill intensities, we use 2000 Census microdata that describe anonymized individuals in terms of many characteristics but only identify their geographic location at the level of prefecture. We construct industries' and occupations' skill intensities by computing the average years of schooling of those individual employed in them.

For India, we use 2001 Census data, since India has not yet released 2011 Census data describing industrial and occupational employment in geographic detail. There are town-level tabulations that describe a town's main workers in terms of educational level, industrial categories, and occupational classification. We construct these population counts for the metropolitan areas described in Section 2.5 by summing over the constituent towns. To obtain industries' and occupations' skill intensities, we compute the average years of schooling of those individual employed in them, controlling for state fixed effects.

## 4 Skill distributions

In each country, we characterize the distribution of skill using four categories of educational attainment of (very) roughly equal size. Where possible, we report population elasticities for a variety of metropolitan-area definitions to assess the robustness of our results.

### 4.1 Brazil: Population elasticities

For Brazil, the four educational categories are "no schooling", "high school dropout", "college dropout", 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 4: Brazil: Population shares for educational categories, 2010

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

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

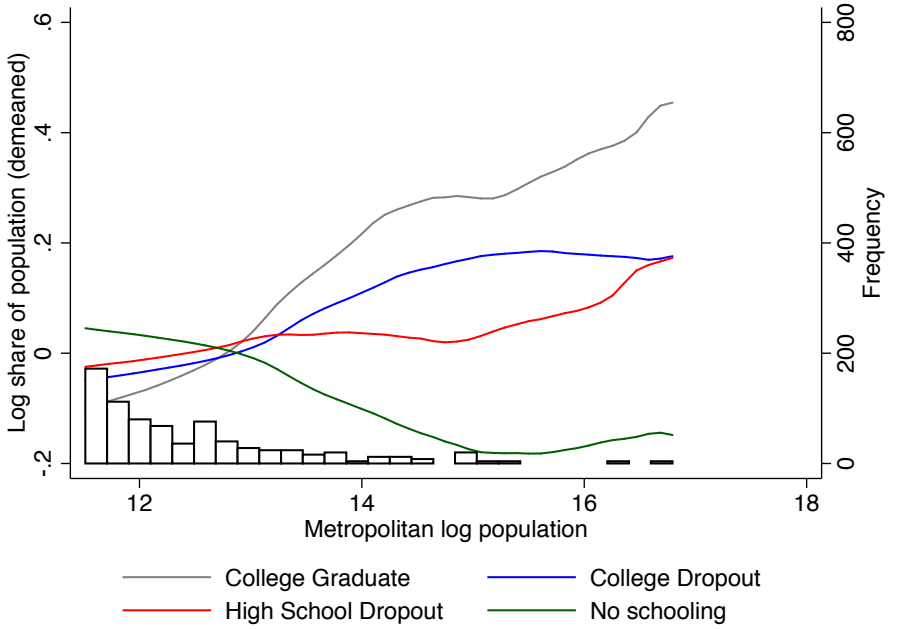
| | Commuting | | | | Nightlights | | | Microregions |
|---|---|---|---|---|---|---|---|---|
| Threshold | 05 | 10 | 15 | 25 | 10 | 30 | 50 | NA |
| No schooling | 0.912 | 0.919 | 0.921 | 0.914 | 0.929 | 0.916 | 0.912 | 0.858 |
| × log population | (0.00942) | (0.00963) | (0.00975) | (0.0104) | (0.0109) | (0.0110) | (0.0113) | (0.00896) |
| High School Dropout | 1.041 | 1.035 | 1.033 | 1.027 | 1.044 | 1.033 | 1.025 | 1.115 |
| × log population | (0.0104) | (0.0105) | (0.0105) | (0.0110) | (0.0121) | (0.0110) | (0.0108) | (0.0132) |
| College Dropout | 1.102 | 1.086 | 1.087 | 1.086 | 1.091 | 1.096 | 1.095 | 1.217 |
| × log population | (0.0123) | (0.0118) | (0.0122) | (0.0129) | (0.0136) | (0.0133) | (0.0134) | (0.0159) |
| College Graduate | 1.178 | 1.159 | 1.168 | 1.179 | 1.155 | 1.173 | 1.174 | 1.302 |
| × log population | (0.0236) | (0.0233) | (0.0247) | (0.0263) | (0.0270) | (0.0258) | (0.0257) | (0.0273) |
| Observations | 816 | 768 | 768 | 880 | 620 | 676 | 724 | 1,672 |
| Geographic units | 204 | 192 | 192 | 220 | 155 | 169 | 181 | 418 |

NOTES: Sample is geographic units with population greater than 100,000.

Table 5 reports population elasticities for these four skill groups for eight different definitions of metropolitan areas. The first four columns use commuting-based metropolitan areas, the next three use nightlights-based metropolitan areas, and the final column reports results for microregions, the geographic unit commonly employed in prior studies of Brazil. A few patterns are immediately evident. First, within any column, the order of the population elasticities conforms to the prediction of the model in Davis and Dingel (2014): more skilled groups exhibit higher population elasticities. Comparing across the commuting-based and nightlights-based columns, the estimated elasticities are quite stable. As suggested by the comparisons in Figure 3, the patterns of economic activity are not sensitive to the threshold employed in defining the metropolitan areas and the two different methodologies yield metropolitan areas that exhibit similar patterns. The results for microregions exhibit notable contrast to 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.

Figure 9 relaxes the first-order approximation employed in Table 5 by plotting a local

Figure 9: Brazil: Non-parametric population elasticities for educational categories, 2010




NOTES: Each series plots a local mean smoother using an Epanechnikov kernel. Metropolitan areas defined by commuting ties between municipios, using Duranton (2015) algorithm with 10% threshold.

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-liner 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 (2014) found for US metropolitan areas.

In sum, we find that in Brazil larger cities are skill-abundant when employing a high-dimensional notion of skill.

## 4.2 China: Population elasticities

We compute population elasticities for educational categories in China for the year 2000. While year-2010 township-level population counts are available, year-2010 data describing educational attainment and sectoral employment counts is only available at the county level. As we show below, characterizations of the spatial distribution of skills are sensitive to

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 |

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.[14]

Table 7 reports population elasticities for these four skill groups for four different definitions of metropolitan areas. The first three columns use metropolitan areas obtained by aggregating townships on the basis of nightlights, 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 nightlight 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 in China larger cities are skill-abundant when employing a high-dimensional notion of skill. 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.

[14]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 7: China: Population elasticities for educational categories, 2000

| | Township-based | | | County-based |
|---|---|---|---|---|
| Nightlight intensity threshold: | 10 | 30 | 50 | 30 |
| Primary school or less | 0.911 | 0.900 | 0.907 | 0.968 |
| × log population | (0.00765) | (0.00880) | (0.0104) | (0.0121) |
| Middle school | 1.003 | 0.985 | 0.973 | 1.018 |
| × log population | (0.00469) | (0.00591) | (0.00889) | (0.00696) |
| High school | 1.129 | 1.092 | 1.069 | 0.995 |
| × log population | (0.0123) | (0.0123) | (0.0135) | (0.0206) |
| College or university | 1.361 | 1.320 | 1.308 | 1.093 |
| × log population | (0.0246) | (0.0266) | (0.0321) | (0.0389) |
| Observations | 4,604 | 3,200 | 1,520 | 4,572 |
| Number of geographic units | 1151 | 800 | 380 | 1143 |

NOTES: Sample is geographic units with population greater than 100,000. DATA SOURCES: Census of Population

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 |

## 4.3 India: Population elasticities

For India, 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.

In the present draft, 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 nightlights. We therefore use the definition of metropolitan areas that is the union of urban agglomerations (aggregated across state borders on the basis of nightlights) and census towns of sufficient population size. The data source that we employ describes educational attainment 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: India: Population elasticities for educational categories, 2001

| | Inclusion threshold | | |
|---|---|---|---|
| | None | 0.8 | 0.95 |
| No education | 0.962 | 0.983 | 0.974 |
| × log population | (0.0264) | (0.0177) | (0.0267) |
| Primary | 0.975 | 0.989 | 1.006 |
| × log population | (0.0218) | (0.0200) | (0.0286) |
| Secondary | 1.017 | 1.037 | 1.052 |
| × log population | (0.0174) | (0.0149) | (0.0217) |
| College graduate | 1.029 | 1.064 | 1.070 |
| × log population | (0.0216) | (0.0178) | (0.0284) |
| Observations | 1,320 | 1,160 | 820 |
| Number of geographic units | 330 | 290 | 205 |

NOTES: Sample is 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.

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 are increasing in the level of educational attainment. 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.

As in the cases of Brazil and China, India's metropolitan areas that are more populous are more skill-abundant. This finding is robust across various samples that we consider in order to address limitations of the underlying data sources.

## 4.4 Pairwise comparisons

We now characterize the spatial distribution of skills in the three economies by means of the pairwise comparisons procedure introduced in Section 3.1. In the interest of brevity, we only report results for one definition of metropolitan areas for each economy. Table 10 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 weighted 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, in a

Table 10: 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 | 0.87 | 0.88 | 5 | 60 | 0.77 | 0.84 |
| 16 | 720 | 0.82 | 0.95 | 10 | 270 | 0.81 | 0.82 | 11 | 330 | 0.65 | 0.76 |
| 64 | 12096 | 0.72 | 0.88 | 50 | 7350 | 0.73 | 0.79 | 33 | 3168 | 0.58 | 0.63 |
| 96 | 27360 | 0.68 | 0.83 | 150 | 67050 | 0.67 | 0.69 | 110 | 35970 | 0.53 | 0.57 |
| 192 | 110016 | 0.63 | 0.77 | 800 | 1917600 | 0.59 | 0.60 | 330 | 323736 | 0.52 | 0.54 |
| Weighted | | | √ | | | | √ | | | | √ |

NOTES: Samples are geographic units with population greater than 100,000: 192 Brazilian metropolitan areas defined by commuting with 10% threshold, 800 Chinese metropolitan areas defined by nightlights 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.

high-dimensional sense as captured by four educational-attainment categories.

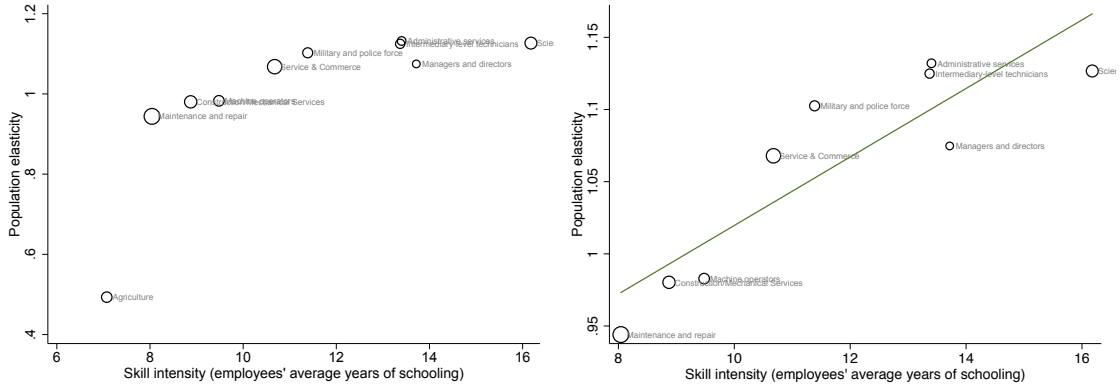
If we are willing to assume that the comparison of any two educational categories is equally informative across the three economies, we can also compare the success rates across countries to gauge the degree to which larger cities are more skill abundant. While these comparisons are also complicated by the fact that the number of metropolitan areas with population greater than 100,000 differs significantly across Brazil, China, and India, 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 (2014) is that of Brazil, followed by China and then India. This is similar to the finding of Chauvin et al. (2016), 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.

In sum, using two different empirical methods to assess the degree to which the population distribution is log-supermodular in skill and metropolitan population, we find broad evidence that larger cities are skill-abundant in Brazil, China, and India. These findings are, in some cases, sensitive to using metropolitan areas defined by contiguity of nightlights 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.

# 5 Sectoral distributions

In the theory of Davis and Dingel (2014), larger cities are relatively more skilled, cities' equilibrium productivity differences are Hicks-neutral, and sectors can be ordered by their

Figure 10: 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 defined by commuting ties between municipios, using Duranton (2015) algorithm with 10% threshold.

skill intensity, so larger cities employ relatively more labor in skill-intensive sectors. The results of Section 4 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.

In this preliminary draft, we restrict attention to Brazil.

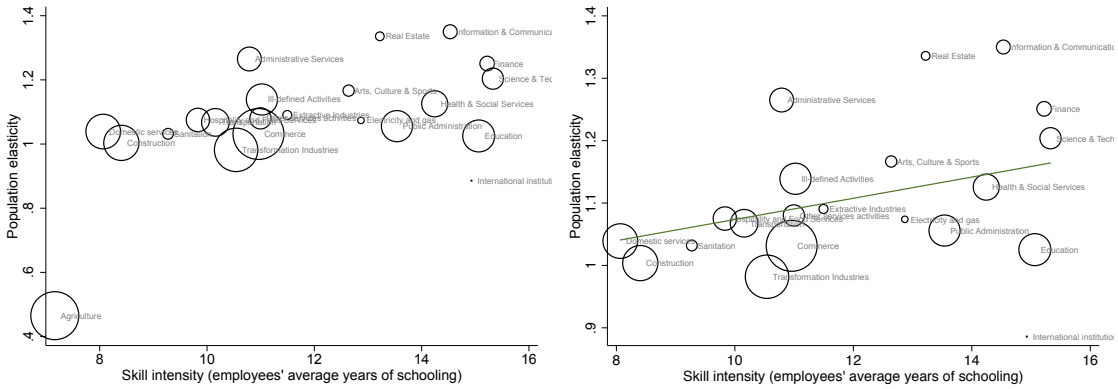
## 5.1 Brazil: Population elasticities

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.

Figure 10 depicts the results of using 10 occupational categories to define sectors. In the left panel, the very low population elasticity of agricultural 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 (2014) predicts that the population elasticity of occupational employment should rise with skill intensity and indeed we see a clear positive relationship in Figure 10.

Figure 11 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 are increasing with skill intensity in general. 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

Figure 11: 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.

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.

# 6 Conclusion

[incomplete]

# References


Bleakley, Hoyt and Jeffrey Lin, "Portage and Path Dependence," *The Quarterly Journal of Economics*, 2012, *127* (2), 587–644.

Bustos, Paula, Bruno Caprettini, and Jacopo Ponticelli, "Agricultural Productivity and Structural Transformation: Evidence from Brazil," *American Economic Review*, June 2016, *106* (6), 1320–65.

Cavalcanti, Tiago, Daniel Da Mata, and Frederik G Toscani, "Winning the Oil Lottery; The Impact of Natural Resource Extraction on Growth," IMF Working Papers 16/61, International Monetary Fund March 2016.

Chauvin, Juan Pablo, Edward Glaeser, Yueran Ma, and Kristina Tobio, "What is different about urbanization in rich and poor countries? Cities in Brazil, China, India and the United States," *Journal of Urban Economics*, 2016, pp. –.

Costa, Francisco, Jason Garred, and João Paulo Pessoa, "Winners and losers from a commodities-for-manufactures trade boom," *Journal of International Economics*, 2016, *102*, 50 – 69.

Davis, Donald R. and Jonathan I. Dingel, "The Comparative Advantage of Cities," NBER Working Paper 20602 2014.

Dix-Carneiro, Rafael and Brian K. Kovak, "Trade Reform and Regional Dynamics: Evidence From 25 Years of Brazilian Matched Employer-Employee Data," Working Paper 20908, National Bureau of Economic Research January 2015.

Donaldson, Dave and Adam Storeygard, "The View from Above: Applications of Satellite Data in Economics," *Journal of Economic Perspectives*, November 2016, *30* (4), 171–98.

Duranton, Gilles, "Delineating Metropolitan Areas: Measuring Spatial Labour Market Networks Through Commuting Patterns," in Tsutomu Watanabe, Iichiro Uesugi, and Arito Ono, eds., *The Economics of Interfirm Networks*, Tokyo: Springer Japan, 2015, pp. 107–133.

Henderson, J. Vernon, "The Sizes and Types of Cities," *American Economic Review*, September 1974, *64* (4), 640–56.

_ , *Urban Development: Theory, Fact, and Illusion*, Oxford University Press, March 1991.

_ , "Urbanization and the Geography of Development," Technical Report 6877, World Bank Policy Research Working Paper 2014.

_ , Adam Storeygard, and David N. Weil, "Measuring Economic Growth from Outer Space," *American Economic Review*, April 2012, *102* (2), 994–1028.

Hoffmann, Rodolfo, "Inequality in Brazil: the contribution of pensions," *Revista Brasileira de Economia*, 12 2003, *57*, 755 – 773.

Hu, Shiwei, Steven Brakman, and Charles van Marrewijk, "Smart Cities are Big Cities - Comparative Advantage in Chinese Cities," CESifo Working Paper Series 5028, CESifo Group Munich 2014.

Kovak, Brian K., "Regional Effects of Trade Reform: What is the Correct Measure of Liberalization?," *The American Economic Review*, 2013, *103* (5), 1960–1976.

Office of Management and Budget, "2010 Standards for Delineating Metropolitan and Micropolitan Statistical Areas," in "Federal Register," Vol. 75 June 2010, pp. 37246–37252.

Rozenfeld, Hernán D., Diego Rybski, Xavier Gabaix, and Hernán A. Makse, "The Area and Population of Cities: New Insights from a Different Perspective on Cities," *American Economic Review*, August 2011, *101* (5), 2205–25.

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

# A Defining metropolitan areas

## A.1 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 nightlights raster layer and convert those contour lines into polygons. Occasionally this procedure creates polygons-within-polygons, which can lead to erroneous assignments in subsequent steps. This happens, for instance, when the nightlights reveal a sufficiently large (dark) park or lake entirely surrounded by a metropolitan area. We obtain contiguous areas by dissolving these smaller polygons with 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 merge. This yields a many-to-one assignment of spatial units to metropolitan areas.

## A.2 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.

# B Data description

## B.1 Satellite image data

Nightlights 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 and 2010, F152000 and F182010, from the "average visible, stable lights" series.[15]

## B.2 Brazil

### B.2.1 Geography

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

[15] Filenames of the form: `F1?YYYY_v4b_stable_lights.avg_vis.tif`

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 nightlights, we use the nightlights raster data described above and shapefiles for Brazilian municipios.

### B.2.2 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.

## B.3 China

### B.3.1 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. To implement our nightlights-based methodology, we use the nightlights raster data for 2000 and 2010 described above and apply light-intensity thresholds ranging from 10 to 60 in increments of 10.

### B.3.2 Skills and Sectors

Data on township-level and county-level employment by educational level, industrial categorial, and occupational classification 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. Data on workers' educational attainment by industry and occupation comes from the 2000 Population Census and the 2010 China Family Panel Studies.

## B.4 India

### B.4.1 Geography

In this preliminary draft, 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 nightlights-based methodology. 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.

### B.4.2 Skills and Sectors

Tables from the 2001 Census and 2011 Census are available via the Government of India's website. Data on town-level employment by educational level, industrial categorial, and occupational classification come from Tables B-9, B-4, and B-24, respectively. These are "Main Workers by Educational Level, Age and Sex", "Main Workers Classified by Age, Industrial Category and Sex", and "Occupational Classification of Main Workers in Non-Households Industry, Trade, Business, Profession or Service by Class of Worker and Sex", respectively. Data on workers' educational attainment by industry and occupation comes from Tables B-7 and B-27 of the 2001 Census, "Main Workers Classified By Industrial Category, Educational Level and Sex" and "Occupational Classification of Main Workers and Marginal Workers other than Cultivators and Agricultural Labourers by Sex and Educational Level". We restrict attention to individuals between the ages of 25 and 59.