WORKING PAPERS
MEASURING THE SIZE AND GROWTH OF CITIES USING NIGHTTIME LIGHT
September 2018
No. 2018/14

MEASURING THE SIZE AND GROWTH OF CITIES USING NIGHTTIME

Ch, R.
Martin, D.
Vargas, J.

DEVELOPMENT BANK OF LATIN AMERICA

Wp

WORKING PAPERS

DEVELOPMENT BANK OF LATIN AMERICA

MEASURING THE SIZE AND GROWTH OF CITIES USING NIGHTTIME LIGHT

Ch, R.
Martin, D.
Vargas, J.

CAF – Working paper No. 2018/14
September 2018

ABSTRACT

This paper uses high-resolution images of nighttime luminosity to estimate a globally comparable measure of the size of metropolitan areas around the world for the years 2000 and 2010. We apply recently-proposed methodologies that correct the known problems of available nighttime luminosity data including blurring, instability of lit pixels overtime and the reduced comparability of night light images across satellites and across time. We then develop a protocol that isolates stable nighttime light pixels that constitute urban footprint, including low luminosity urban settlements such as slums, and excluding confounding phenomena such as highway illumination. When analyzed together with existing geo-referenced population datasets, our measure of urban footprint, can be used to compute city densities for the entire world. After characterizing some basic stylized facts regarding the distribution of urban sprawl, urban population and population density across world regions, we offer an application of our measure to the study of the size distribution of cities, including test of the Zipf's Law and Gibrat's Law.

Wp

WORKING PAPERS

MEDICIÓN DEL TAMAÑO Y EL CRECIMIENTO DE LAS CIUDADES USANDO LUZ NOCTURNA

Ch, R.
Martin, D.
Vargas, J.

CAF - Documento de trabajo N° 2018/14
Septiembre 2018

RESUMEN

Este trabajo usa imágenes de alta resolución de luminosidad nocturna para estimar una medida del tamaño de las áreas metropolitanas de todo el mundo, comparable globalmente, para los años 2000 y 2010. Utilizamos metodologías recientemente propuestas que corrigen los problemas de los datos de luminosidad nocturna, incluyendo borrosidad, inestabilidad de pixeles de luz a lo largo del tiempo y baja compatibilidad de las imágenes de luz nocturna entre satélites y en el tiempo. Después desarrollamos un protocolo que identifica pixeles de luz nocturna estables que constituyen huella urbana, incluyendo asentamientos urbanos de baja luminosidad como asentamientos informales, y excluyendo fenómenos como la iluminación de autopistas interurbanas. Cuando se analiza en conjunto con bases de datos existentes de población geo-referenciada, nuestra medida de huella urbana puede ser usada para calcular densidades urbanas para todo el mundo. Después de caracterizar algunos hechos estilizados básicos sobre la distribución de la huella urbana, de la población urbana y de la densidad poblacional en todas las regiones del mundo, este trabajo muestra una aplicación de nuestra medida al estudio de la distribución de tamaño de las ciudades, incluyendo tests de la Ley de Zip y la Ley de Gibrat.

Measuring the Size and Growth of Cities Using Nighttime Light*

Rafael Ch† Diego A. Martin‡ Juan F. Vargas§

September 20, 2018

Abstract

This paper uses high-resolution images of nighttime luminosity to estimate a globally comparable measure of the size of metropolitan areas around the world for the years 2000 and 2010. We apply recently-proposed methodologies that correct the known problems of available nighttime luminosity data including blurring, instability of lit pixels overtime and the reduced comparability of night light images across satellites and across time. We then develop a protocol that isolates stable nighttime light pixels that constitute urban footprint, including low luminosity urban settlements such as slums, and excluding confounding phenomena such as highway illumination. When analyzed together with existing geo-referenced population datasets, our measure of urban footprint, can be used to compute city densities for the entire world. After characterizing some basic stylized facts regarding the distribution of urban sprawl, urban population and population density across world regions, we offer an application of our measure to the study of the size distribution of cities, including test of the Zipf's Law and Gibrat's Law.

*We are grateful to Gilles Duranton, Cynthia Goytia and Pablo Sanguinetti for useful comments and discussion. We also thank seminar participants at CAF's RED 2017 workshop in Buenos Aires and the First Urban and Regional Economics Workshop at Universidad Javeriana. This paper was possible thanks to generous funding by CAF-Development Bank of Latin America.

†New York University, Department of Politics, 19 West 4th, room 230, NY, NY 10009. Email: rafael.ch@nyu.edu.

‡Princeton University, Woodrow Wilson School of Public and International Affairs, 323 Bendheim Hall, Princeton, NJ 08544. E-mail: diegoam@princeton.edu.

§Corresponding author. Department of Economics, Universidad del Rosario, Calle 12C No. 4-69, Of. 315, Bogotá. E-mail: juan.vargas@urosario.edu.co.

1 INTRODUCTION

The process of economic growth usually involves structural changes, of which urbanization is the most important one (Kuznets 1968). In fact, for many social scientists urbanization is the hallmark of economic development. For example, in the absence of historical information on per capita income, several authors have used historical urbanization rates as a proxy of economic prosperity.[1] This implies that cities, as nuclei of the urbanization process, are the main drivers of economic growth.

But what is a city? Conceptually, cities -or metropolitan areas-[2] are the spatial integration of social and economic activity, and the higher the geographical concentration of households and firms, the greater the potential economic benefits net of congestion costs (Duranton & Puga 2004, Rosenthal & Strange 2004, Duranton & Turner 2011). However, in spite of our relatively consensual theoretical understanding of cities, its empirical measurement has proven more challenging. Empirical definitions usually rely on administrative borders, which generally transcended by the physical space where interactions between economic and social agents occur.

This implies that stylized facts about patterns of urbanization across countries are likely to be biased because of measurement error (Roberts et al. 2016), which may partly explain, for instance, why the available cross-country estimates of the Zipf's Law are not robust.[3] As put by Duranton & Puga (2004) "[T]he empirical validity of Zipf's Law is hotly debated" (p.833), and the cross country evidence (Rosen & Resnick 1980, Soo 2005) has been interpreted both for and against the Zipf's Law. Chauvin et al. (2017), for example, test the existence of the Law in Brazil, China, India and the US, using respectively micro-regions, cities, districts and metropolitan areas, heterogeneous measures whose comparison the authors recognize as "debatable" (p.20).

In this paper, we introduce a novel methodology to translate the concept of city into a standard measurement of urban sprawl that relies on almost no methodological assumption and is independent of country-specific administrative criteria of city limits, allowing for global comparisons.

Specifically, following recent remote sensing and economic development literature -reviewed by Donaldson & Storeygard (2016)-, our measure leverages on the availability of satellite

[1] See, for example, Acemoglu et al. (2002) and Dittmar (2011). The historical urbanization data are estimated by Bairoch (1991).

[2] For simplicity, in this paper we refer to cities and metropolitan areas interchangeably. Our data-driven measure or urban agglomeration is arguably closer to what the latter are.

[3] Country-specific measures of cities or metropolitan areas are based on arbitrary thresholds, many of which depend on the quality of local censuses. Since the latter depends on state and administrative capacity, country-specific measurement error is likely correlated with the level of development, and thus it is potentially non-classical.

images of night lights.[4] We adopt two main steps (see section 2 for details): First, nighttime light data is cleaned and processed. Second, processed light pixels are classified to identify the existence of a city and its extension. The first step recognizes that nighttime light data suffers from several sources of measurement error, which we correct by combining insights from a large body of literature on remote sensing (Imhoff et al. 1997, Elvidge et al. 2009, Liu et al. 2012, Abrahams et al. 2018).[5] Importantly, we do not claim any improvement on the remote sensing methodological literature. Rather, we provide a systematic and replicable algorithm on how to correct several sources of potential bias for the specific purpose of classifying and identifying lit pixels as urban areas or not.

The second step classifies metropolitan areas as spatial conglomerates of 1 km^2 nighttime stable lit pixels.[6] As a result, we end up with a geo-referenced dataset of the location and sprawl of 2,677 (3,072) metropolitan areas with more than 100,000 inhabitants (typically encompassing two or more neighboring cities that are economically integrated) in 2000 (2010) for the entire world (see Table 1).

Based on the Spanish acronym of Metropolitan Areas Extension Database (*Base de Extensión de Áreas Metropolitanas*), we call our resulting data BEAM.[7] BEAM is highly precise, as suggested by multiple cross-validation procedures and comparison with other measures of urban expansion, available for specific cities of the world, such as those based urban built-up and coverage (see Appendix).[8]

Our globally comparable dataset can be further overlaid with existing geo-referenced population datasets in order to estimate urban population and population growth, and compute city-wide densities. In fact, we compare the distribution of urban sprawl, urban population and population density across world regions and document several robust stylized facts (see section 3). In particular, while the distribution of urban population in cities larger than 100,000 people is relatively similar across regions, regardless of their

[4]Nighttime light has been shown to be a reliable proxy for economic activity both nationally and in geographically small areas (Henderson et al. 2011, Bleakley & Lin 2012, Michalopoulos & Papaioannou 2013, Lowe 2014, Storeygard 2012, Weidmann & Schutte 2016, Pinkovskiy 2013, Goldblatt et al. 2018). This implies that it can be used to identify geographically integrated economic markets, thus linking our practical and conceptual definitions.

[5]In particular, we correct censoring problems using a simple clipping procedure, apply a deblurring filter in order to compute an unbiased sprawl for each city/satellite-year, and adjust the instability of lit pixels overtime and between satellites due to differences in satellite-specific crossing times, sensors' degradation and geographic misalignment.

[6]A stable pixel is defined as one that shows up in the satellite-based nighttime images more than 20% of the time. The stability of the pixel is necessary to remove ephemeral lights including cars and other types of spurious bright sources, but takes into account low luminosity areas including peripheral city slums.

[7]Conveniently, beam also means ray of light in English.

[8]An older version of BEAM, which did not include all the correction procedures described in section 2, and that was only available for a subset of regions of the world, was used as an input for the 2017 *Report on Economics and Development*, the flagship report of CAF-Development Bank of Latin America, which funded this research.

level of development, average urban sprawl is larger in developed countries, especially in North America. This implies that population density is higher in developing countries. This is an important fact, as it suggests that developing countries are not fully able to take advantage of agglomeration economies to translate higher density into higher economic growth.

In addition to providing details about the computation on BEAM and describe these basic stylized facts, we illustrate the advantages of this measure by revisiting some facts recently highlighted by Chauvin et al. (2017) about the size distribution of cities in developed and developing countries. We do so both for the same set of countries used in that article -the U.S., Brazil, China and India-, and for the entire world. Zipf's Law predicts that the relationship between the log of cities' population and the log of cities' population rank is equal to -1.[9] Our estimated coefficients of this relationship are consistently smaller in absolute value than those of Chauvin et al. (2017) -with the exception of China- and smaller than Zipf's Law prediction. This implies that, in the four countries analyzed by these authors, as the population rank falls cities are larger than what it is predicted by the Law, a results that is not consistent with the largest city in each country exerting some sort of primacy.[10] This is consistent with our estimates of Gibrat's Law, which relates the initial population of cities with their inter-period growth rate. For every region, we find that smaller cities are growing faster than the largest cities.

The rest of the paper is organized as follows. Section 2 discusses several measures used in the literature to estimate the size and sprawl of metropolitan areas, and describes the methodology, including the data problems, the data cleaning procedure and the pixel classification into urban conglomerates used in this paper. It also includes a brief description of how population counts can be used as an ex-post estimation procedure. Section 3 summarizes world-level stylized facts regarding the distribution of city sprawls, urban population and population density. Section 4 revisits some of the stylized facts reported by Chauvin et al. (2017) using comparable data for the whole world, thus establishing some basic patterns of the size distribution of cities. Section 5 concludes.

2 Measuring the sprawl and size of metropolitan areas

Increasing urbanization rates have highlighted the need to re-define urban areas and explore methodologies to accurately approximate the size of cities and to measure urban sprawl. As cities grow, political administrative boundaries become obsolete to define metropolitan areas in a way that is consistent with modern urban patterns. While the

[9] This implies that a city in the n^{th} rank of population size is expected to have $1/n$ the population of the largest city, thus following a power law.

[10] This is also what Chauvin et al. (2017) find for the U.S. and China.

U.S. introduction of Metropolitan Statistical Areas (MSAs) in the 50s established a general benchmark that was followed by other high income countries, many developing countries still lack institutional definitions for metropolitan areas, and thus rely heavily on administrative boundaries and simple aggregation procedures that are often at odds with patterns of urbanization.

There are at least three broad approaches used to define the extent and size of cities. First, the use of administrative geographic units aggregated through various iterative approaches given different thresholds, for instance minimum population counts or minimum density levels. In some cases, administrative units are agglomerated if they are contiguous and economically integrated, as in the case of Brazil, where contiguous municipalities are integrated into *microregions* conditional on having similar economic features, according to the Brazilian Institute for Geography and Statistics. However, non-comparability arises as countries define "integration" and "contiguity" differently, and use arbitrary thresholds. This is exacerbated by dissimilar historical paths determining how the smaller administrative units came to be in terms of size and shape. The latter is particularly problematic since the smallest administrative units might comprise both rural and urban areas, thus leading to over or underestimation of the size of metropolitan urban areas. In general, as the smallest administrative units increase in area, so does the bias towards including rural areas.

A second approach relies on commuting patterns, understanding metropolitan areas as integrated labor markets. In most developed countries, census data tracks commuting patterns by asking respondents about the location of their residence and that of their job. Most studies that use this approach predefine metropolitan cores using ancillary population data, and then use commuting patterns to determine how to aggregate it. More recently, however, Duranton (2015) showed that there is no need to rely on predefined urban cores by proposing an algorithm to delineate metropolitan areas based solely on commuting thresholds.

However, reliable data on commuting patterns is not available for most countries, especially developing ones. Moreover, while commuting patterns portrait the reach of labor markets, they leave aside input-output linkages which might cover wider areas, or knowledge spillovers which tend to revolve around shorter distances.[11]

Note that both the administrative unit agglomeration and commuting pattern approaches rely on arbitrary thresholds. While some studies show that results hold when modifying thresholds selection, most do not carry out such robustness checks. OECD (2012), for instance, defines urban cores as high-density clusters of contiguous grid cells of 1 km^2 with a density of at least 1,500 inhabitants per km^2, but use a lower density threshold of 1,000

[11]For a discussion on the topic see Duranton (2015).

people per km^2 for Canada and the United States given that, in such countries "several metropolitan areas develop in a less compact manner" (OECD (2013) p. 3). These data processing choices makes cross-country and regional inference non fully comparable.[12]

In response to the weaknesses of the first two approaches, recent work has moved towards a more harmonized definition of cities and rural areas based on the use of geo-referenced population datasets. Uchida & Nelson (2008), for instance, developed the agglomeration method which relies on population size and density thresholds and a travel time radius of a "seizable" settlement, to define a city.[13] In turn, Dijkstra & Poelman (2014) followed a spatial approach and developed a cluster method defining two types of urban areas: *high density clusters*, with threshold values of 1,500 inhabitants per km^2 and 50,000 people, and *urban clusters*, with thresholds of 300 inhabitants per km^2 and 5,000 people.

This suggests that the geo-referenced data approach suffers from similar drawbacks as the other approaches, as it still relies on contiguity criteria as well as on thresholds that are subject to arbitrariness (see Roberts et al. (2016) for a discussion). Another pitfall that is common to all approaches is the perpetration of sources of measurement error along the construction process. For instance, the existing methods take into account national definitions of urban areas which are not comparable across countries and time; geo-referenced population datasets depend on census and therefore rely on each country's census minimum collection unit size, which determine the data's spatial resolution.[14]

More recently, and to overcome the drawbacks of the traditional approaches to measure city size and urban sprawl, researchers have started to use remote-sensing techniques. The next sub-section focuses on the use of the incidence of light at night to delineate the size and shape of cities.

[12]Another not so common approach to define metropolitan areas relies on non-economic criteria including concepts such as "sense of belonging", as captured through survey data.

[13]To that end, Uchida & Nelson (2008) used data from the Center for International Earth Science Information Network's (CIESIN), the Global Rural-Urban Mapping Project (GRUMP) and Landscan geo-referenced population datasets.

[14]One example is the Gridded Population of the World database (GPW), produced by the Socioeconomic Data and Applications Center (SEDAC) from NASA, and hosted by CIESIN. The data rely on an areal-weighting method (known as uniform distribution or proportional allocation method) to disaggregate population from census units into grid cells through the simple assumption that the population of a grid cell is an exclusive function of the land area within that pixel (Doxsey-Whitfield et al. 2015). The first disadvantage of using an areal-weighting disaggregation method is that "the precision and accuracy of a given pixel is a direct function of the size of the input areal unit. Consequently, for countries where the input units are quite large, the precision of population estimates for individual pixels within that unit can be degraded" (Lloyd et al. 2017). For example, the average input unit resolution for developed regions in the GPW data (version 4) is 944 km^2. Contrast that with an input unit resolution of 3,518 (4,700) km^2 for middle-income (developing) regions (de Sherbinin & Adamo 2015). Accuracy is thus proportional to development and this introduces measurement error bias. To tackle this problem, CIESIN developed the Global Rural Urban Mapping Project (GRUMP) which, pretty much in the spirit of BEAM, assigns population over grid cells according to nighttime light data. However, GRUMP does not perform all the necessary corrections to the nighttime light data which we describe in this section, thus introducing bias to the estimates.

2.1 Remote sensing approach and the use of Nighttime Light to measure urban sprawl

Satellite imagery and remote sensing techniques have been widely used to measure physical quantities related to urbanization and urban footprint. For instance, Burchfield et al. (2006) used satellite imagery of land cover and land use in U.S. from 1976 to 1992 to build boundaries of contiguous areas with similar land cover. They find that the extent of land use remained roughly unchanged during this period. More recently, Saiz (2010) used satellite-generated data on terrain elevation and presence of water bodies to estimate the amount of developable land in U.S. metropolitan areas, and found that inelastic housing supply is highly correlated with geographical constraints.

Perhaps the most important contribution to cross-country comparisons of urban footprint using satellite imagery is the *Atlas of Urban Expansion* (AUE), produced by the Marron Institute of Urban Management at NYU, in collaboration with UN-Habitat and the Lincoln Institute of Land Policy. AUE uses high resolution satellite imagery together with disaggregated census data to estimate, for 200 cities in the world, urban footprint, urban built-up area, population density and several indicators that characterize the specificities of the urban outlook of cities. These include including the shares of urban infill, leapfrog, and extension for the years 1990, 2000 and 2014 (Angel et al. 2012).

While valuable and highly precise, the approach followed by the AUE is extremely costly. There are two reasons for this. On the one hand, AUE uses as input various sources of information, including high resolution satellite images from Landsat and Google Earth, disaggregated census data, and specific surveys obtained from local researchers measuring different land uses, property regimes, affordability of housing and attributes of properties available for sale or rent. On the other, producing the detailed portfolio of AUE indicators is computationally demanding. This may partly explain the reduced sample of cities. Furthermore, as with former approaches, census and land use data cannot be pooled to provide suitable comparable observations due to national differences on census methods, accuracies and capture times.

One alternative is to use the Nighttime Light longitudinal data available from the Defense Meteorological Satellite Program (DMSP) of the National Center for Environmental Information (NOAA). DMSP satellites take cloud-free annual composites images to produce digital remotely sensed images of the world's light, that unlike a photograph, can be manipulated to extract information.[15] Each image corresponds to a 30 arc-second pixel (roughly 1 km^2 at the equator) containing a specific *Digital Number* (DN). The DN is assigned according to the pixel luminosity, and varies between 0 to 63.

[15]Satellite images take into account only persistent lighting, thus discarding ephemeral events including fires.

Among other uses, nighttime luminosity can help predicting urban settlements.[16] This has been done at least since Elvidge et al. (1997). Indeed, night lights have been shown to correlate accurately with population density in the U.S. (Sutton et al. 1997), to identify urbanization rates (Imhoff et al. 1997), to estimate global population counts (Sutton et al. 1999), and even to single out low density settlements (Elvidge 2000). Moreover, the use of night-time satellite imagery to derive urban boundaries has been validated both for developed and developing countries (Henderson et al. 2003), and several studies have used this tool to study urbanization changes at different geographical scales (see He et al. (2006), Zhang & Seto (2011) and Liu et al. (2012), among others). Harari (2017) used night-time lights to classify the shape of Indian cities and found that city compactness is correlated with lower wages and higher housing rents.[17]

Nighttime luminosity can be combined with high resolution Landsat images for more accurate predictions of urban construction and population counts. For instance, Landsat images can be used to extract green areas and bodies of water to refine the measures derived from night lights (see the work of Goldblatt et al. (2018) for India, Mexico and the U.S.).

However, nighttime light data suffers from a series of problems that, in the absence of the appropriate corrections, affect its reliability and comparability across space and time (Elvidge et al. 2009, Liu et al. 2012, Zhang & Seto 2011, Abrahams et al. 2018). For instance, factors such as the exact position of satellites at the time of capturing the information, and the satellite-specific degradation of the capturing sensors imply that recorded images can suffer from misalignment within a satellite across time, and across satellites within the same year, making DN values vary for reasons other than the actual luminosity intensity of pixels In addition, nighttime light suffers from blurring, which means that light "overglows" and cities seem to be bigger than they actually are, especially if they are more luminous (because of higher economic activity).

The literature has proposed several ways to address each of these issues, albeit in a decentralized fashion. For instance, to address the overglow problem, Goldblatt et al. (2018) utilize Landsat satellite imagery to cutoff the excess night light glowing beyond urban built-up. However, this technique cannot fix the existent spatial correlation that blurring causes among contiguous light pixels, and that is increasing with luminosity. This issue is addressed by Abrahams et al. (2018), who develop a de-blurring procedure

[16]High resolution luminosity satellite data has been used as a proxy for economic activity across countries, at the sub-national level and for very small areas (Henderson et al. 2011, Bleakley & Lin 2012, Michalopoulos & Papaioannou 2013, Martinez 2018, Lowe 2014, Storeygard 2012, Weidmann & Schutte 2016, Pinkovskiy 2013). The literature on night lights as a proxy for development is summarized by Donaldson & Storeygard (2016). These data have also been used, in conjunction with high-resolution daytime satellite images and machine learning tools, to predict sub-national poverty (Jean et al. 2016).

[17]For a thorough discussion on the use of nighttime light data to evaluate urban coverage and change see Goldblatt et al. (2018).

that does not rely on ancillary data such as Landsat.[18]

An important contribution of this paper is to propose general and replicable procedure to apply the known correction techniques while relying on minimal cleaning and processing assumptions to estimate the urban sprawl of cities in the entire world. The next subsection describes the data sources as well as the cleaning and processing procedure.

2.2 Data

Following the conceptual definition of cities as the spatial integration of social and economic activity, as well as the growing literature on how night lights are strongly associated with economic performance and development, this paper uses high-resolution images of luminosity at night to compute the size of metropolitan areas for the entire world, in 2000 and 2010.

To this end, we rely on average visible nighttime light raster files and frequency images provided by NOAA.[19] Our methodology can be divided in three steps: i) the cleaning and processing of nighttime light data; ii) the classification of pixels to define the existence of a city and its extension; iii) The estimation of population zonal statistics.

2.2.1 Cleaning and processing nighttime light

DMSP nighttime light data suffers from blurring. Light spreads far beyond urban built-up coverage, and brighter pixels generate larger blurs. We rely on (Abrahams et al. 2018)'s two-step deblurring filter, with a threshold of 20% minimum cloud-free nights.[20] We apply this filter to each city/satellite-year and hence do not use a single deblurring correction parameter.[21]

After deblurring the night light images, we apply a number of additional correction

[18] A different approach to correct the blurring problem when measuring cities with night lights is to use luminosity thresholds in an attempt to screen rural areas out. However, as discussed, thresholds are arbitrary and data are not comparable when different thresholds are used for different regions, as in the case of Liu et al. (2012).

[19] The raster images of interest are those from the F14-2000, F15-2000, F18-2010 satellites-years. We do not rely on radiance-calibrated images given that, since we do not perform ant within-city inference, we are not concerned about top-coding (the fact that high light values end up at the top of the 0-63 scale), but only about the emission of light above certain frequency threshold. Hsu et al. (2015) discuss how radiance-calibrated nighttime lights deal with top coding in high luminosity areas, such as city centers.

[20] This threshold allows us to capture low luminous areas such as slums and other informal settlements, but leaves out transitory lights from fires or road traffic, for example.

[21] In order to identify the areas where to apply the deblurring filter (thus *potential* cities), we clip cities using the non-deblurred stable contiguous lit pixels of any given area (the procedures to obtain stable pixels is explained below). While non-deblurred areas are, by definition, larger than the deblurred ones, in general they follow the city's shape.

procedures to ensure a reliable estimation of city sizes. First, because images recorded by the same satellite may not be comparable across years, we apply Wu et al. (2013)'s inter-calibration procedure. Second, as noted by Zhao et al. (2015), nighttime light rasters of different satellites (and of different years for the same satellite) tend to be shifted a couple of pixels, reducing comparability across sources and over time. Following these authors, we use a reference image (specifically that of satellite/year F142001) and shift all the other nighttime lights rasters to fit it.[22] Lastly, comparing the information recorded by two different satellites (F14 and F15) within a given year allows us to remove any intra-annual unstable lit pixels.[23] In addition, we assign to every stable lit pixel the average DM value of the two satellites, hence producing an intra-annual stable composite for each sample year.[24]

2.2.2 Pixel classification

We classify pixels into urban polygons according to the following process. First, we drop all pixels with DN equals to zero (i.e., with no recorded luminosity); second, we classify as a city any conglomerate of at least one km^2 of stable light pixels; third, we merge the country shapefiles of Natural Earth and DIVA-GIS to assign cities to existing countries and use the universe of cites identified by the AUE for year 2000 to assign city names to our conglomerates;[25] fourth, we clip water bodies (including both the ocean and inland lakes).[26] Once we have the city polygon we can estimate city-level area and different zonal statistics, including a calculation of the total population count in a given city.

2.2.3 Population zonal statistics

The urban sprawl polygon shapefile resulting from the pixel classification explained in the previous section allows us to estimate a wide array of statistics within the boundaries of

[22] As a result, Satellite/year F142000 is unchanged, F152000 is moved left 1 pixel and up 1 pixel, and F182010 up 1 pixel.

[23] A lit pixel is defined as an intra-annual unstable if it was detected with a positive DM by only one satellite.

[24] Using the same procedure for inter-annual corrections assumes, as Liu et al. (2012), that cities cannot *reduce* their size overtime. We consider this an unrealistic assumption that is not consistent with experiences such as that of Detroit is the U.S., and that would then tend to overestimate the size of cities.

[25] In several cases, estimated metropolitan areas are big enough to encompass two or more existing cities. This is the case, for instance, of the metropolitan northeast corridor of the U.S., from New York City to Washington D.C. We split this mega-metropolitan areas using the only processing assumption of the paper (besides the use of a minimum percentage of cloud-free nights, see section 2.2.1): cities that are adjacent are split if the width of the connecting processed light is less than 5 Km. Of course, a specific researcher may want to avoid this arbitrary separation and work with the larger estimated metropolitan area.

[26] More ancillary data could be use in the future to further define city shapes as done by Goldblatt et al. (2018).

the estimated metropolitan areas. Of those, population counts are particularly important. Several population raster files exist to estimate population counts for a given city across time.[27] In particular, we generate zonal statistics using the Landscan geo-referenced population gridded raster files, for the years 2000 and 2010.[28]

As noted above, all geo-referenced population datasets rely on census or population data, and thus entail comparability difficulties given differences in census minimum collection unit size (spatial resolution) across countries. While Landscan cannot rule out completely this potential source of bias, Stevens et al. (2015) and Sorichetta et al. (2015) show that Landscan's mapping accuracy outperforms other geo-referenced population datasets, particularly those that rely on the areal-weighting method explained above.[29]

Using Landscan, we estimate population counts and the average population density by city. It is worth noting that we restrict the dataset to cities with more than 100,000 after estimating zonal statistics for two reasons. First, this ensures we exclude other light emission zones such as oil refineries. Second, we are interested in revisiting stylized facts about urbanization and the size distribution of cities (including Zipf's Law -a right-tail distribution law- and Gibrat's Law) that have been recently studied using cities defined with the same threshold (see, for instance, Chauvin et al. (2017)).

The result of this three-step procedure is the Metropolitan Areas Extension Database, called BEAM based on the Spanish acronym.

2.3 Benefits of BEAM

This urban sprawl measure based on nighttime light data has several benefits in terms of accuracy and empirical-computational practicality. First, it allows a comparable analysis of stylized facts about urbanization across different regions of the world, with potentially non-homogeneous administrative definitions of cities. Second, it does not rely on assumptions about people's mobility (like the measures based on commuting patters). Third, the methodology does not assume an arbitrary population density threshold, or a threshold expressed in terms of time and number of travels.[30] Fourth, while the computational de-

[27]The on line appendix of this link provides a description of the main geo-referenced population datasets.

[28]As a robustness test we also utilize the Global Human Settlement geo-referenced dataset. Results are available upon request.

[29]This is because Landscan uses spatial data on land cover, roads, slopes, urban built up and village locations to run a multi-variable model used to disaggregate census population counts within certain administrative boundaries in a non-uniform way.

[30]For practical reasons, we do have a contiguity assumption when merging pixels. The other two assumptions are the lit pixel stability threshold and the 5km threshold to split mega regions. While the former is necessary to screen out ephemeral lights such as inter-urban highways that might not be considered part of urban polygon, the latter can be lifted depending on the specific interests of the researcher.

mand of BEAM is still high, it differs substantially from that of other methods, including the use of construction coverage. Fifth, our methodology also allows urban area identification without the need of administrative census data. Sixth, it does not establish a population threshold ex-ante. Rather, the population count of an area is estimated *after* the creation of the city's urban extension. This guards BEAM from potential political manipulation, a concern that arises when utilizing pre-defined urban cores, as noted by Duranton (2015).

Lastly, and most importantly, luminosity has been extensively proven to be a reliable proxy of economic activity, and so it allows us to identify clearly integrated urban markets. As noted for the case of Bogota by Duranton (2015), the city is not officially constituted as a metropolitan area even though there is no discontinuity with several nearby municipalities, including large towns such as Soacha (in the south) and Chia (in the north). The needed administrative integration does not, however, affect the economically relevant estimates of BEAM.

It should also be noted that the cleaning and pixel processing algorithm is transparent and simple, and it is flexible enough to allow for changes that any researcher may consider relevant. According to Duranton (2015), transparency, clarity and flexibility are desirable characteristics of these type of measures. Privileging computational practicality and comparability across regions despite an arguably small cost in terms of accuracy, and in order to being able to identify integrated labor markets instead of built-up coverage, we chose not to complicate our measure by, for instance, subtracting green areas and non-built up areas, unless they do not form part of the conglomerate of light pixels.

3 Urban Population, City Size and Population Density

Table 1 shows the number of metropolitan areas of at least 100,000 people estimated by BEAM in 2000 and 2010, in each of the world's regions, as defined by the World Bank.[31] The total number of cities went from 2,677 in 2000 to 3,072 in 2010. Except for Europe and Central Asia (ECA) the estimated number of cities increases over the decade-long period.[32] The highest growth occurs in Sub-Saharan Africa (SSA, where the number of metropolitan areas increases by almost 60%), followed by Middle East and North Africa (MENA) and by South Asia (SA, in both regions the number of metropolitan areas increases by 25%), East Asia and Pacific (EAP, with a 20% increase), Latin America and

[31] Regions in the Table are listed alphabetically.

[32] This is driven by Western Europe, where the estimated number of metropolitan areas decreases due to the merger (in terms of nighttime light-pixel clusters) of several cities. This happens mainly in the metropolitan area of London, in Belgium and Holland, in the Rhine area, and in northern Italy.

the Caribbean (LAC, 18%) and North America (NA, 8%).[33] The region with the largest share of metropolitan areas above 100,000 people in 2000 was ECA (28% of the total number of estimated metropolitan areas). In 2010 it was EAP (27%, with ECA in the second place at 23%).

Based on BEAM, in this section we report some basic stylized facts about urbanization. This is made in relation to three dimensions: the spatial extent of cities or metropolitan areas, urban population and population density, which is the ratio of population to area. These are summarized in Figure 1. Specifically, Panel A reports, through a box plot, the distribution of the log of the size of metropolitan areas in each region, as estimated by BEAM.[34] Regions are organized ascendantly according to the median size of metropolitan areas in 2000.

EAP, SA and SSA were the regions with the lowest average city size in 2000 (and very similar distributions with the exception of a few large outliers in EAP). MENA, ECA and LAC had somewhat larger cities, and NA had significantly larger cities, compared to any region. By 2010, the average size of EAP cities grew relative to SA and SSA and became comparable to that of MENA and LAC. The average size of cities in ECA also grew but NA cities continued to be the largest of the world.

As described in section 2, we overlay BEAM with the Landscan geo-referenced population datasets for the years 2000 and 2010. This allows us to compute urban population levels for the estimated metropolitan areas in every region. Interestingly, Panel B of Figure 1 shows that the distributions of urban population were remarkably similar across regions, both in 2000 and 2010. Panel C, in turn, reports population density levels, which by and large are the mirror image of urban sprawls, precisely because of the similarities in urban population levels across regions: both in 2000 and 2010 EAP, SA and SSA were the regions with the highest population density. MENA, ECA and LAC showed lower values on average and NA was, by and large, the region with the lowest density.

The evidence presented in panel C of Figure 1 is consistent with the existence of a negative association between population density and degree of development. This can have two complementary explanations. On the one hand, increases in the average income levels are associated with declines in urban density. This is explained by a positive relationship between income and the demand for residential space. On the other hand, the relatively high levels of density in developing regions may be a reflection of the recent, accelerated and informal urbanization process, experienced by these regions in the mid-

[33]The World Bank classifies Mexico in LAC, and hence NA is constituted by the U.S. and Canada only.

[34]The quantities of interest included in the box plot are: the median of the distribution (horizontal line inside the box); the 25^{th} and 75^{th} percentiles (box limits); upper and lower adjacent value (box whiskers, defined as the box upper (lower) limit + (-) 3/2 of the 75^{th} to 25^{th} percentiles gap); outliers (values outside the whiskers).

twentieth century. In fact, much of the high density of developing cities is explained by the incidence of informal settlements (Jedwab & Vollrath 2017).

In the next section we use BEAM to study the size distribution of metropolitan areas across world's regions.

4 City size distribution and growth

In this section, we use BEAM to study the size distribution of cities across different regions of the world using a homogeneous, comparable measure. A commonly used tool for this purpose is the so-called Zipf's Law. In essence, the law states that, if we consider the N cities from a country and sort them according to their size from 1 to N, the city in position i will have a population equal to that of the largest city divided by i. So, the second city will have half the population of the first, the third, one third, and so on (Gabaix 1999a).[35]

Compliance with Zipf's Law implies that cities' growth process is independent of their own size and is due to exogenous productivity shocks (Duranton & Puga 2014). If this is true, then Gibrat's Law -which states that the growth of urban areas is independent from its size- also holds.[36]

Zipf's Law and Gibrat's Law constitute an interesting reference point for a descriptive analysis of the potential primacy of the largest cities within a country or region, as well as the relationship between the size of cities and their growth rate. However, comparisons across countries or regions are limited by the lack of homogeneous data (Duranton & Puga 2014). For instance, in a recent paper, Chauvin et al. (2017) explore whether the known facts about urbanization in the U.S. also hold in developing countries, specifically Brazil, China and India. To this end, the authors include tests of Zipf's Law and Gibrat's Law, comparing the four countries using different definitions of city's extension, a limitation the authors are aware of and upfront about when recognizing that the definitions used are "debatable", although probably the best available at the time of writing the paper for the goal of creating a standard definition across different countries.[37]

[35]This is equivalent to saying that the size distribution of cities follows a power law, with a power parameter equal to 1.

[36]Gabaix (1999b) shows that if Gibrat's holds then distribution of city sizes, in equilibrium, will be consistent with Zipf's Law.

[37]The data sources used by Chauvin et al. (2017) are: For the United States, the 2010 *Consolidated Metropolitan Statistical Areas*, defined by the U.S. Census. For Brazil, *microregions*, composed by agglomerations of contiguous and economically integrated municipalities that have similar economic features, and defined by the Brazilian Institute for Geography and Statistics. For China, administrative *cities*, including provincial-level and prefecture-level areas, which typically comprise both urban and rural territories. For India, *districts*, the second-level administrative division of the country after states and union territories. In the four countries, Chauvin et al. (2017) restrict the samples to urban areas with 100,000 urban dwellers or more.

In this section we review the descriptive analysis of Chauvin et al. (2017) for the case of Zipf's Law and Gibrat Law. However, in order to better understand the potential differences of BEAM with the data sources using by these authors, we start by reporting the distribution of population across cities of different population size, both for the countries studied by Chauvin et al. (2017) (Table 2.1) and for all the regions of the world (Table 2.2).[38]

Table 2.1 is equivalent to Table 1 of Chauvin et al. (2017). The distribution of people living in the metropolitan urban areas of different sizes estimated by BEAM for the U.S., both in 2000 and 2010 is remarkably similar to what Chauvin et al. (2017) report for the U.S. Metropolitan Statistical Areas (MSAs), as defined by the United States Office of Management and Budget. This is because MSAs are conceptually much closer to the object that BEAM measures using contiguous conglomerates of night light pixels, relative to the definitions used by these authors for other countries.[39] Instead, the distribution of people living in BEAM-estimated metropolitan urban areas of different sizes differs to what Chauvin et al. (2017) report for Brazil's "microregions" and China's "cities", and to a lesser extent to what they report using India's "districts". These differences highlight the importance of using a unified metric such as the proposed in this paper.

Table 2.2 goes beyond the four countries analyzed by Chauvin et al. (2017) and reports the distribution of people across metropolitan areas for all the regions of the world, using the classification of the World Bank. NA is the region with the largest concentration of people in the largest city size bin (cities with population larger than 1.5 million). In 2000 (2010), 47% (52%) of NA residents lived in these cities, followed by MENA and LAC with 44% (48%) and 29% (34%), respectively. In turn, SSA is the region with the smallest concentration of people in the largest city size bin, with 8% (10%) in 2000 (2010).

If we aggregate the population living of cities smaller that 1.5 million people (in all bin sizes starting from 100,000), the region with the largest share of people living in small and medium cities is EAP, with 28% (27%) of the total population in 2000 (2010). At the other end, the region with the smallest share of people in this type of cities is SSA with 7% (8%) of the total population in 2000 (2010).

We now use BEAM to compute Zipf's Law and Gibrat's Law, both for the set of countries studied by Chauvin et al. (2017) and for all the regions (and countries) of the world.

[38] As described in section 2, we overlay BEAM with the Landscan geo-referenced population datasets to compute urban population levels for the metropolitan areas estimated by BEAM.

[39] In the Appendix, we examine how BEAM compares with other measures, including MSAs. The level of coincidence is 94%. The 6% discrepancy is mostly explained by the existence of MSAs with metropolitan areas with less than 100,000 inhabitants, the threshold established by BEAM (see Figure A-4).

4.1 Zipf's Law

We test Zipf's Law by estimating the relationship between the log of population of metropolitan areas and the log of population rank.[40] The law implies an estimated coefficient equals to -1. A coefficient greater than -1 suggests that cities ranked lower than the biggest are bigger than Zipf's prediction and hence there is no evidence of primacy of the main city. On the other hand, a coefficient less than -1 suggests that cities ranked lower than the biggest are smaller relative to Zipf's prediction, and there is evidence of primacy of the main city in the city-size distribution within a country or region.

The results are reported in Figures 2.1 and 2.2, where we plot the cumulative distribution of city sizes for cities larger than 100,000 people, together with the fitted lines of the estimation procedure, applied to the 2010 BEAM sample. The first figure focuses on the sample of countries of Chauvin et al. (2017) and the second looks at each one of the seven regions defined by the World Bank. Below each sub-figure we report the coefficients from the estimated equation model. The slope coefficient represents the Zipf's Law test.

A comparison of Figure 2.1 with the equivalent Figure 2 of Chauvin et al. (2017) suggests that, in the two exercises, the rank-size relationship is very similar for the cases of U.S. and Brazil. But this is not true for China and India. In these countries, the rank-size relationship reported by Chauvin et al. (2017) displays a strong non-linearity which is not present using BEAM. This makes the fitted line of the Zipf's Law test reported in our paper a better representation of the true rank-size relationship.[41]

The estimated slope coefficients of the Zipf's Law test applied, using BEAM, to the four countries studied by Chauvin et al. (2017) in 2000 and 2010, are reported in Table 3.1. We also report the number of metropolitan areas larger than 100,000 people included in each case. The 2010 coefficients are smaller than those estimated by Chauvin et al. (2017) for each country except China, and all are larger than -1 (smaller in absolute value). This suggest that, in the four countries, as the population rank falls the population is larger than what the Zipf's Law predicts. Irrespective of the actual values, this is also the conclusion of Chauvin et al. (2017) with the exception of Brazil, for which their estimated coefficient is -1.18, slightly smaller (larger in absolute value) than -1. Rather, our estimated coefficient for this country is -0.91.[42]

Because they imply that population rises too quickly as rank falls, our estimated coeffi-

[40] For comparability with Chauvin et al. (2017), we follow Gabaix & Ibragimov (2011) and subtract 1/2 from the log of population rank. Gabaix & Ibragimov (2011) use simulations to show that this is a better estimate of the coefficient of the power law distribution of city size relative to the traditional estimate.

[41] The rank-size relationship displayed in Figure 2.2 for each one of the world's seven regions is also well fitted by the straight line of the Zipf's Law test, with no apparent non-linearities.

[42] And that estimated for 2000 is -0.96, very similar to the estimate, for the same year of Soo (2014).

cients are inconsistent with a potential primacy of the main city of these four countries. Our estimates of Gibrat's Law, reported in the next subsection, are consistent with this finding.

This is also the case when we test Zipf's Law at the region level (see Table 3.2). Out of the seven world regions classified by the World Bank, only ECA (both in 2000 and 2010) has an estimated coefficient roughly equal to -1. In fact, our results using BEAM suggest that, with coefficients equal to -1.06 in 2000 and -1.03 in 2010, ECA meets the Zipfs's Law.

The region-level estimates, however, mask some important variation across countries. This point is apparent when examining Figure 2.3, which plots over the same support the estimated slope coefficients of the rank-size relationship of each country, together with the 95% confidence intervals.[43]

A few regions have all or most their countries with coefficients consistently larger than -1. This is the case, for example of EAP, NA, SA and SSA. Most countries in MENA meet Zipf's Law, except Saudi Arabia (SAU) where the coefficient is significantly larger than -1, and Algeria (DZA), where the coefficient is significantly smaller (and thus the population rises too slowly as rank falls, which is consistent with the primacy of Algiers). In LAC countries, as the population rank falls the population is larger than what the Zipf's Law predicts, with the exception of Ecuador and Chile, which meet Zipf's Law.

The most interesting case, however, is that of ECA. Recall from the region-wide test that, with a coefficient equal to -1.03 in 2010, ECA was the only region with a size distribution of cities consistent with Zipf's Law. Figure 2.3, however, shows that this is the result o a large variation across the region's countries, with many of them having estimated coefficients larger than -1, many other having coefficients smaller than -1, and just a couple (Poland and Uzbekistan) actually being consistent with Zipf's Law.

4.2 Gibrat's Law

Gibrat's Law states that the growth of urban areas is independent from its size. If bigger cities grew faster then they would tend to concentrate a larger share of population and economic activity overtime, becoming a sort of urban black holes. Gibrat's Law is usually tested by estimating, at the city level, the relationship between the change in population over a specific period and the initial population level. If the estimated correlation coefficient is not statistically distinguishable from zero this would be consistent

[43] We drop countries with less than 10 cities larger than 100,000 to avoid estimated relationships that are too imprecise. However, below each sub-figure we report the mean and standard deviation of the within region estimated slope coefficient, both for the plotted subsample as well ad for the entire sample of countries.

with Gibrat's Law.

Using BEAM, Table 4.1 reports the estimates of this relationship for the four countries analyzed by Chauvin et al. (2017). Specifically we report the correlation between the (log) population in the first BEAM year (2000) and the population growth between 2000 and 2010.[44] In all cases, except Brazil, the estimated coefficient is negative, suggesting that smaller cities grew faster in population during the decade-long period. The coefficient is statistically significant for China and India.

Table 4.2 reports equivalent estimation results for each one of the seven world regions. Consistent with the findings pertaining Zipf's Law, which precluded an interpretation favoring the primacy of large cities, the estimated coefficient is, in all cases, negative (and statistically significant in all cases except ECA and LAC).

5 Conclusion

Data-driven spatial definitions of the size and growth of urban areas face a trade-off between the precision of estimates and the empirical practicality of the estimation process, mainly given by computational demand. In general, highly precise measures imply low empirical practicality. The measure and dataset introduced in this paper offer a very precise estimate (at least when compared to other existing databases) that is also computationally tractable, as well as transparent and replicable.

We use nighttime light satellite data, that we pre-process using various existing correction procedures, to estimate the size of urban areas in the entire world. We also merge our data with existing geo-referenced population datasets to compute various comparable measures at the city level, including the size of cities, urban population and population density.

The result is a comprehensive dataset of global metropolitan areas greater than 100,000 people in two time periods, 2000 and 2010. Because our measures are comparable for cities of all the world's regions, and are thus independent of the level of development and of administrative definitions, the dataset allows us to revisit some of the basic stylized facts about urbanization in the world. For instance, we compute the statistics necessary to describe the size distribution of cities across the different regions of the world, and test Zipf's Law and Gibrat's Law.

We find that, in most of the world's regions (and regardless of the level of development) cities ranked after the largest urban areas have populations greater than what it is pre-

[44] To that end, we keep a balanced sample of cities, dropping from the estimation sample cities that were smaller than 100,000 in 2000 (and thus were not included in the dataset) but larger than the threshold in 2010 (thus making it into the dataset).

dicted by Zipf's Law. Consistent with this, when testing Gibrat's Law we find a negative association coefficient between the initial population in 2000 and the population growth from 2000 to 2010 at the city level, implying that smaller cities grew faster in population during the first decade of the XXI century.

The future of remote sensing tools and nighttime light appears to be promising as an increasing number of satellites become available to follow human migration and urban built-up patterns. DMSP nighttime light data comes with several measurement pitfalls, but recently launched satellites with new sensors are incorporating new technology, including multi-spectral bands to capture land use, radiometrically calibrated images to solve light saturation problems to analyze highly luminous areas in a reliable way, and higher levels of spatial resolution to increase the accuracy of delineating urban expansion, size and shape. The urban economics literature is already moving towards exploiting these type of datasets, that permit better cross and within countries comparable analyses. BEAM is an example of such datasets.

References

Abrahams, A., Lozano-Gracia, N. & Oram, C. (2018), 'Deblurring dmsp nighttime lights: A new method using gaussian filters and frequencies of illumination', *Remote Sensing of Environment* **210**(1), 242–258.

Acemoglu, D., Johnson, S. & Robinson, J. A. (2002), 'Reversal of fortune: Geography and institutions in the making of the modern world income distribution', *The Quarterly journal of economics* **117**(4), 1231–1294.

Angel, S., Blei, A. M., Civco, D. L. & Parent, J. (2012), *Atlas of urban expansion*, Lincoln Institute of Land Policy Cambridge, MA.

Bairoch, P. (1991), *Cities and economic development: from the dawn of history to the present*, University of Chicago Press.

Bleakley, H. & Lin, J. (2012), 'Portage and path dependence', *Quarterly Journal of Economics* **127**, 587–644.

Burchfield, M., Overman, H. G., Puga, D. & Turner, M. A. (2006), 'Causes of sprawl: A portrait from space', *The Quarterly Journal of Economics* **121**(2), 587–633.

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', *Journal of Urban Economics* **98**, 17–49.

de Sherbinin, A. & Adamo, S. (2015), 'Ciesin's experience in mapping population and poverty', *United Nations Expert Group Meeting on Strengthening the Demographic Evidence Base for the Post-2015 Development Agenda, Population Division, Department of Economic and Social Affairs* .

Dijkstra, L. & Poelman, H. (2014), 'A harmonised definition of cities and rural areas: the new degree of urbanization', *Regional Working Paper, European Commission, Directorate-General for Regional and Urban Policy* .

Dittmar, J. E. (2011), 'Information technology and economic change: the impact of the printing press', *The Quarterly Journal of Economics* **126**(3), 1133–1172.

Donaldson, D. & Storeygard, A. (2016), 'The view from above: Applications of satellite data in economics', *Journal of Economic Perspectives* **30**(4), 171–198.

Doxsey-Whitfield, E., MacManus, K., Adamo, S., Pistolesi, L., Squires, J., Borkovska, O. & Baptista, S. (2015), 'Taking advantage of the improved availability of census data: A first look at the gridded population of the world, version 4 (gpwv4).', *Papers in Applied Geography* **1**(3), 226–234.

Duranton, G. (2015), 'A proposal to delineate metropolitan areas in colombia', *Revista Desarrollo y Sociedad* (75), 223–264.

Duranton, G. & Puga, D. (2004), Micro-foundations of urban agglomeration economies, *in* 'Handbook of regional and urban economics', Vol. 4, Elsevier, pp. 2063–2117.

Duranton, G. & Puga, D. (2014), The growth of cities, *in* 'Handbook of Economic Growth', Vol. 2, Elsevier, pp. 781–853.

Duranton, G. & Turner, M. A. (2011), 'The fundamental law of road congestion: Evidence from us cities', *The American Economic Review* **101**(6), 2616–2652.

Elvidge, C., Baugh, K., Kihn, E., Kroehl, H. & Davis, E. (1997), 'Mapping city lights with night-time data from the dmsp operational linescan system', *Photogrammetric Engineering and Remote Sensing* **63**, 727–734.

Elvidge, C., Baugh, K., Tilottama, G. & Zhizhin, M. (2009), 'A fifteen year record of global natural gas flaring derived from satellite data', *Energies* **2**, 595–622.

Elvidge, C. D. (2000), 'Radiance calibration of dmsp-ols low-light imaging data of human settlements (cd-rom).', *US Department of Commerce, National Oceanographic and Atmospheric Administration.* .

Gabaix, X. (1999*a*), 'Zipf's law and the growth of cities', *American Economic Review* **89**(2), 129–32.

Gabaix, X. (1999*b*), 'Zipf's law for cities: An explanation', *Quarterly Journal of Economics* **114**(3), 739–67.

Gabaix, X. & Ibragimov, R. (2011), 'Rank- 1/2: a simple way to improve the ols estimation of tail exponents', *Journal of Business & Economic Statistics* **29**(1), 24–39.

Goldblatt, Ran, S. M. F., Tellman, B., Clinton, N., Hanson, G., Georgescu, M., Wang, C., Serrano-Candela, F., Khandelwal, A., Cheng, W.-H. & Balling Jr, R. C. (2018), 'Using landsat and nighttime lights for supervised pixel-based image classification of urban land cover', *Remote Sensing of Environment* **205**(253-275).

Harari, M. (2017), 'Cities in bad shape: Urban geometry in india', *Working paper, The Wharton School* .

He, C., Shi, P., Li, J., Chen, J., Pan, Y., Li, J., Zhuo, L. & Ichinose, T. (2006), 'Restoring urbanization process in china in the 1990s by using non-radiance-calibrated dmsp/ols nighttime light imagery and statistical data', *Chinese Science Bulletin* **51**(13), 1614–1620.

Henderson, M., Yeh, E., Gong, P., Elvidge, C. D. & Baugh, K. (2003), 'Validation of urban boundaries derived from global night-time satellite imagery', *International Journal of Remote Sensing* **24**(3), 595–609.

Henderson, V., Storeygard, A. & Weil, D. N. (2011), 'A bright idea for measuring economic growth', *American Economic Review* **101**(3), 194–199.

Hsu, F.-C., Baugh, K., Ghosh, T. & Zhizhin, M. (2015), 'Dmsp-ols radiance calibrated nighttime lights time series with intercalibration', *Remote Sensing of Environment* **7**, 1855–1876.

Imhoff, M., Lawrence, W., Stutzer, D. & Elvidge, C. (1997), 'A technique for using composite dmsp/ols "city lights" satellite data to map urban area', *Remote Sesing of Environment* **61**, 361–370.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. & Ermon, S. (2016), 'Combining satellite imagery and machine learning to predict poverty', *Science* **353**(6301), 790–794.

Jedwab, R. & Vollrath, D. (2017), 'The urban mortality transition and poor country urbanization'.

Kuznets, S. (1968), *Toward a Theory of Economic Growth*, W.W. Norton & Company.

Liu, Z., He, C., Zhang, Q. & Yang, Y. (2012), 'Extracting the dynamics of urban expansion in china usin dmsp-ols nighttime light data from 1992 to 2008', *Landscape and Urban Planning* **106**, 62–72.

Lloyd, C., Sorichetta, A. & Tatem, A. (2017), 'High resolution global gridded data for use in population studies', *Scientific Data* **4**(170001).

Lowe, M. (2014), 'The privatization of african rail', *Working paper* .

Martinez, L. R. (2018), 'How much should we trust the dictator's gdp estimates?', *Working paper, Harris School of Public Policy* .

Michalopoulos, S. & Papaioannou, E. (2013), 'Pre-colonial ethnic institutions and contemporary african development', *Econometrica* **81**(1), 113–152.

OECD (2012), *Redefining "urban": A new way to measure metropolitan areas.*, Paris: OECD.

OECD (2013), 'Definition of functional urban areas (fua) for the oecd metropolitan database', *OECD Working paper* .
URL: *http://www.oecd.org/regional/regional-policy/Definition-of-Functional-Urban-Areas-for-the-OECD-metropolitan-database.pdf*

Pinkovskiy, M. (2013), 'Economic discontinuities at borders: Evidence from satellite data on lights at night', *Working paper* .

Roberts, M., Blankespoor, B., Deuskar, C. & Stewart, B. (2016), 'Urbanization and development: Is latin america and the caribbean different from the rest of the world?', *Working paper, The World Bank* .

Rosen, K. T. & Resnick, M. (1980), 'The size distribution of cities: an examination of the pareto law and primacy', *Journal of Urban Economics* **8**(2), 165–186.

Rosenthal, S. S. & Strange, W. C. (2004), 'Evidence on the nature and sources of agglomeration economies', *Handbook of regional and urban economics* **4**, 2119–2171.

Saiz, A. (2010), 'The geographic determinants of housing supply', *The Quarterly Journal of Economics* **125**(3), 1253–1296.

Soo, K. T. (2005), 'Zipf's law for cities: a cross-country investigation', *Regional science and urban Economics* **35**(3), 239–263.

Soo, K. T. (2014), 'Zipf, gibrat and geography: Evidence from china, india and brazil', *Papers in Regional Science* **91**(1), 159–182.

Sorichetta, A., Hornby, G., Stevens, F. R., Gaughan, A. E., Linard, C. & Tatem, A. (2015), 'High-resolution gridded population datasets for latin america and the caribbean in 2010, 2015, and 2020', *Scientific Data* **2**(150045).

Stevens, F. R., Gaughan, A. E., Linard, C. & Tatem, A. J. (2015), 'Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data', *PloS ONE* **10**(e0107042).

Storeygard, A. (2012), 'Farther on down the road: transport costs, trade and urban growth in sub-saharan africa', *JMP* .

Sutton, P., Roberts, D. & Elvidge, C. D. (1997), 'A comparison of night- time satellite imagery and population density for the continental united states.', *Photogrammetric Engineering and Remote Sensing* **63**, 1303–1313.

Sutton, P., Roberts, D., Elvidge, C. & Miej, H. (1999), 'Census from heaven: an estimate of the global human population using night-time satellite imagery.', *Paper presented at the Western Regional Science Association annual meeting, Ojai, California, USA,* .

Uchida, H. & Nelson, A. (2008), 'Agglomeration index: Towards a new measure of urban concentration', *United Nations, Working Paper* .

Weidmann, N. B. & Schutte, S. (2016), 'Using night light emissions for the prediction of local wealth', *Journal of Peace Research* **54**(2), 1–16.

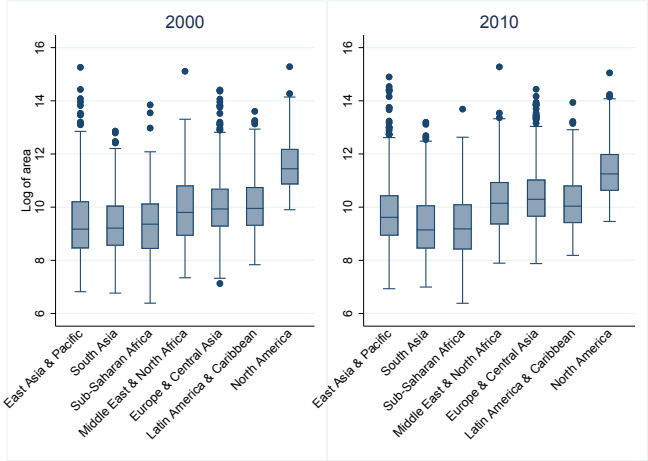
Wu, J., He, S., Peng, J., Li, W. & Zhong, X. (2013), 'Intercalibration of dmsp-ols night-time light data by the invariant region method', *International Journal of Remote Sensing* **34**(20), 7356–7368.

Zhang, Q. & Seto, K. (2011), 'Mapping urbanization dynamics at regional and global scales using multip-temporal dmsp/ols nighttime light data', *Remote Sesing of Environment* **115**(9), 2320–2329.

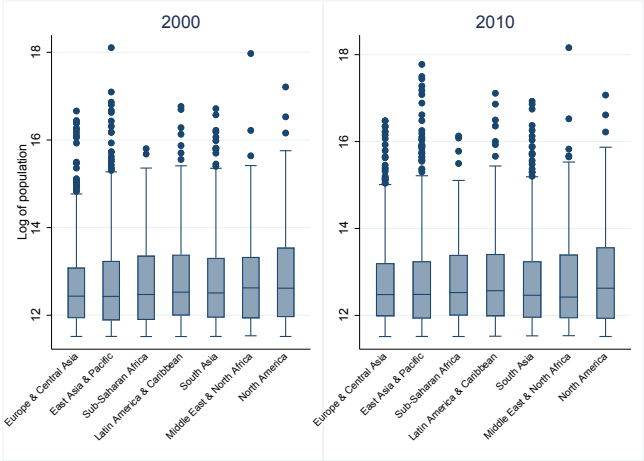
Zhao, N., Zhou, Y. & Samson, E. L. (2015), 'Correcting incompatible dn values and geometric errors in nighttime lights time-series images', *IEEE Transactions on Geoscinece and Remote Sensing* **53**(4), 2039–2049.

Figure 1: City size distribution by region

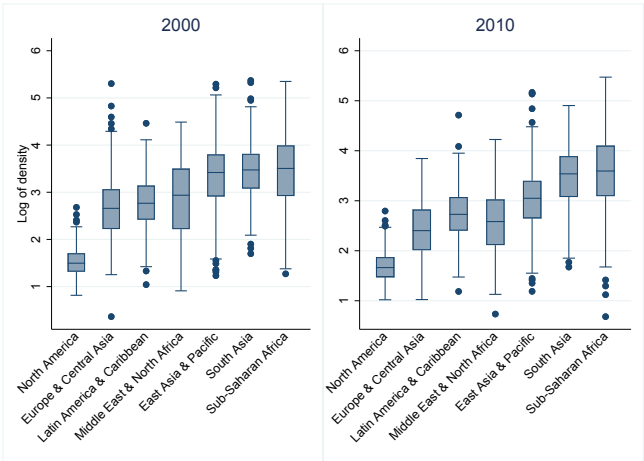
Panel A: Area



Panel B: Population

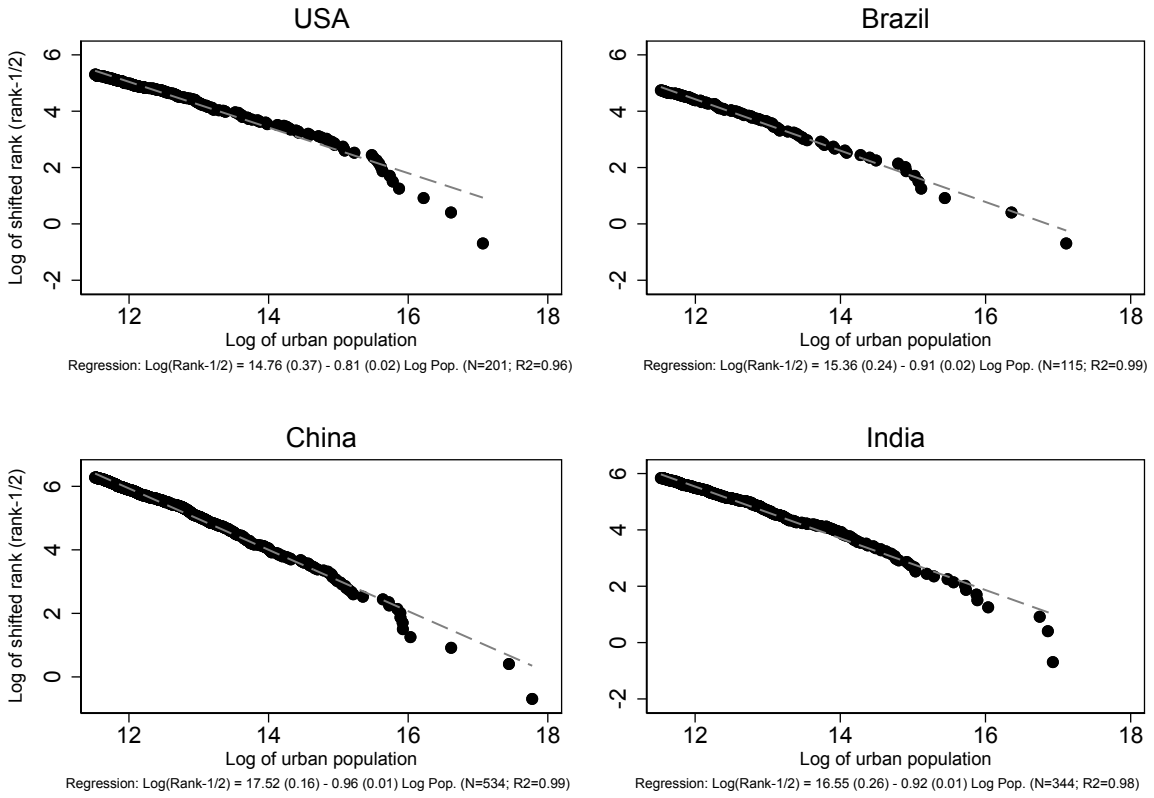


Panel C: Density



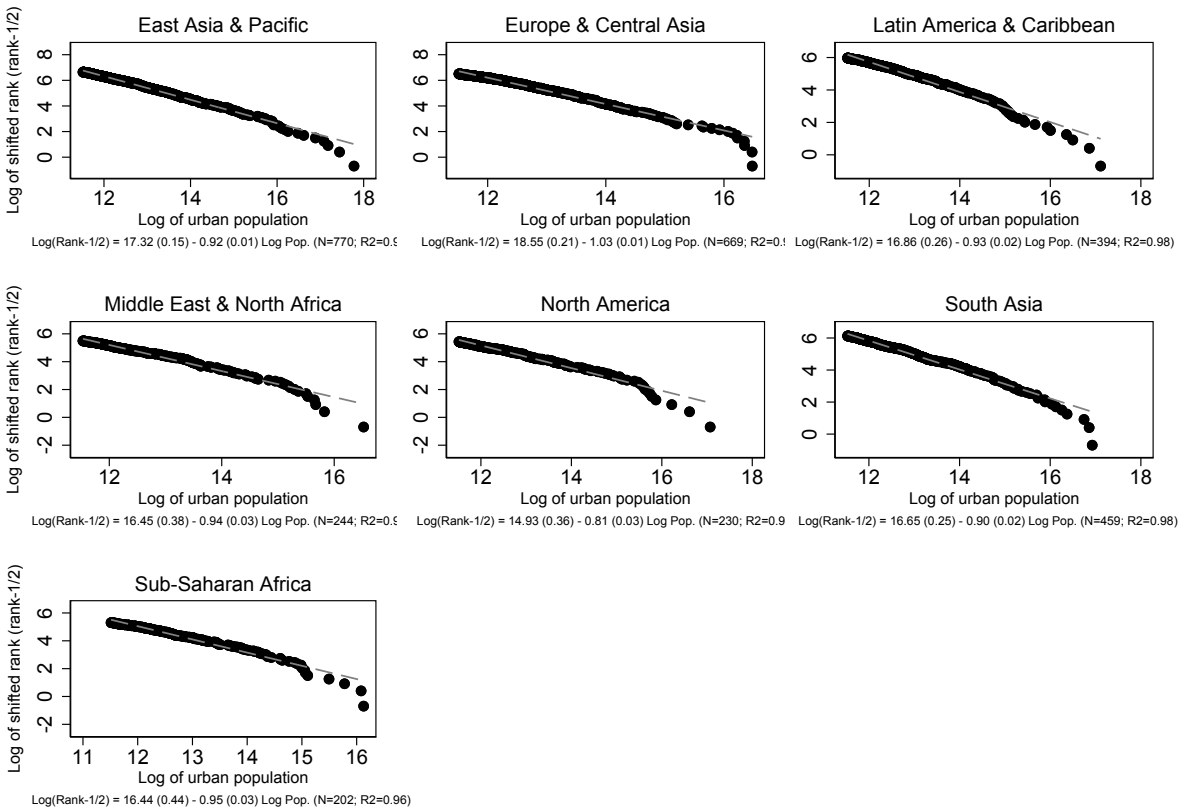
Note. The box of the box plot shows the 25, 50 and 75 percentiles. The vertical lines show the minimum and maximum values, excluding outliers. Outliers are those point values 1.5 times larger (or smaller) than the maximum (or minimum).

Figure 2.1: Zipf's Law by country, 2010



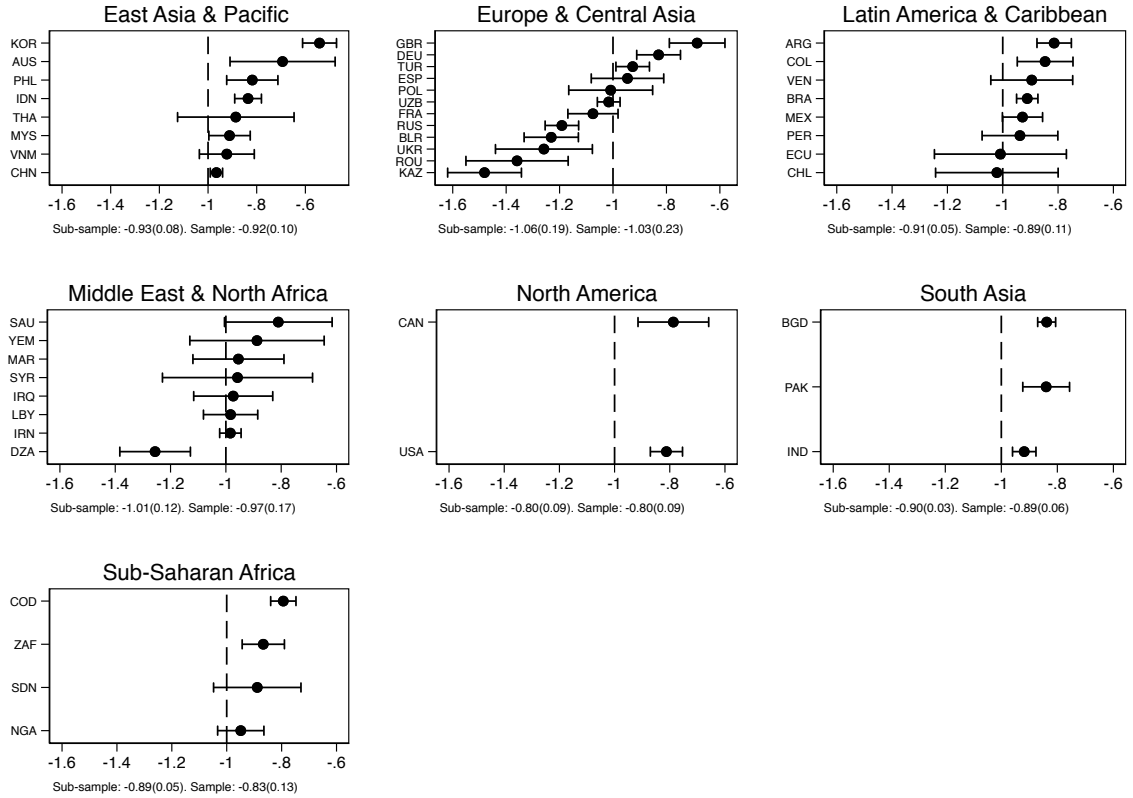
Note. (log) Urban population and (log) rank in 2010. Regression point estimates and standard errors based on (Gabaix & Ibragimov 2011) following a shifted rank, i.e. rank - 1/2. For each country, Zipf's coefficients are statistically significant to the 1% level.

Figure 2.2: Zipf's Law by region, 2010



Note. (log) Urban population and (log) rank in 2010. Regression point estimates and standard errors based on (Gabaix & Ibragimov 2011) following a shifted rank, i.e. rank - 1/2. For each region, Zipf's coefficients are statistically significant to the 1% level. Countries classified following World Bank's regional classification.

Figure 2.3: Zipf's Law β estimator by world regions, 2010



Note. The graph reports regression point estimates and standard errors based on (Gabaix & Ibragimov 2011) following a shifted rank, i.e. rank - 1/2 for each country. The dots show Zipf's coefficient while the horizontal lines represent 95% confidence intervals. We subset countries with at least 10 cities with more than 100,000 inhabitants in 2010, leaving us with 45 countries.The mean and standard deviation (in parenthesis) are included for both the subsample of 45 countries and the full sample. Countries classified following World Bank's regional classification.

Table 1: Estimated number of cities by region

	2000			2010		
	Cities	Percent	Cum.	Cities	Percent	Cum.
	(1)	(2)	(3)	(4)	(5)	(6)
East Asia & Pacific	691	25.8	25.8	831	27.1	27.1
Europe & Central Asia	747	27.9	53.7	707	23.0	50.1
Latin America & Caribbean	334	12.5	66.2	395	12.9	63.0
Middle East & North Africa	198	7.4	73.6	248	8.1	71.1
North America	213	8.0	81.6	230	7.5	78.6
South Asia	367	13.7	95.3	459	14.9	93.5
Sub-Saharan Africa	127	4.7	100	202	6.5	100
Total	2,677	100		3,072	100	

Notes. Cities with more than 100,000 inhabitants or larger. Countries classified following World Bank's regional classification.

Table 2.1: Share of people living in urban areas of different size, by country

	Area of						
	100 k - 250 k	250 k - 500 k	500 k - 1 M	1 M - 1.5 M	1.5 M +	Population in areas 100 K+	Total population
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
2000							
United States	5	6	6	4	47	191	280
Brazil	4	6	6	2	27	76.3	171
China	3	3	3	2	11	273	1260
India	2	2	2	2	13	217	1010
2010							
United States	4	6	6	3	53	218	305
Brazil	5	5	5	2	35	103	199
China	3	3	4	2	20	426	1320
India	2	2	2	2	15	279	1170

Notes. Columns (1) to (5) expressed as a percent of the total population of a country in column (7). Column (6) represents the total urban population in cities with at least 100,000 inhabitants. Population estimated using *BEAM* and Landscan georeferenced population dataset.

Table 2.2: Share of people living in urban areas of different size, by region

	Area of					Population	Total
	100 k - 250 k	250 k - 500 k	500 k - 1 M	1 M - 1.5 M	1.5 M +	in areas 100 K+	population
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
2000							
East Asia & Pacific	3	3	3	2	19	600	2020
Europe & Central Asia	7	8	8	5	26	462	848
Latin America & Caribbean	5	6	7	3	29	255	525
Middle East & North Africa	4	6	8	3	44	197	308
North America	5	6	7	4	47	213	310
South Asia	2	2	2	2	13	301	1350
Sub-Saharan Africa	1	2	2	2	8	87.1	617
2010							
East Asia & Pacific	3	3	4	2	26	819	2170
Europe & Central Asia	6	7	8	6	28	472	881
Latin America & Caribbean	5	6	6	4	35	329	580
Middle East & North Africa	5	5	8	4	48	254	367
North America	4	6	6	4	52	243	337
South Asia	2	2	2	2	16	397	1570
Sub-Saharan Africa	2	2	2	2	10	143	835

Notes. Columns (1) to (5) expressed as a percent of the total population of a country in column (7). Column (6) represents the total urban population in cities with at least 100,000 inhabitants. Population estimated using *BEAM* and Landscan georeferenced population dataset. Countries classified following World Bank's regional classification.

Table 3.1: Zipf's Law by country

	2000		2010	
	Zipf's Law	Cities	Zipf's Law	Cities
	(1)	(2)	(3)	(4)
United States	-0.85	188	-0.81	201
Brazil	-0.96	97	-0.91	115
China	-1.03	460	-0.96	534
India	-0.91	271	-0.92	344

Notes. (log) Urban population and (log) rank in 2010. Regression point estimates and standard errors based on (Gabaix & Ibragimov 2011) following a shifted rank, i.e. rank - 1/2. For each country, Zipf's coefficients are statistically significant to the 1% level.

Table 3.2: Zipf's Law by region

	2000		2010	
	Zipf's Law	Cities	Zipf's Law	Cities
	(1)	(2)	(3)	(4)
East Asia & Pacific	-0.96	691	-0.92	831
Europe & Central Asia	-1.06	747	-1.03	707
Latin America & Caribbean	-0.95	334	-0.93	395
Middle East & North Africa	-0.96	198	-0.94	248
North America	-0.85	213	-0.81	230
South Asia	-0.91	367	-0.90	459
Sub-Saharan Africa	-0.92	127	-0.95	202

Notes. (log) Urban population and (log) rank in 2010. Regression point estimates and standard errors based on (Gabaix & Ibragimov 2011) following a shifted rank, i.e. rank - 1/2. For each country, Zipf's coefficients are statistically significant to the 1% level. Countries classified following World Bank's regional classification.

Table 4.1: Gibrat's Law by country

	USA	BRA	CHI	IND
	(1)	(2)	(3)	(4)
2000-2010	-0.025	0.183	-0.163*	-0.089***
	(0.024)	(0.217)	(0.090)	(0.027)
Constant	0.446	-1.996	2.598**	1.603***
	(0.299)	(2.655)	(1.227)	(0.352)
R-squared	0.014	0.026	0.008	0.027
No of observations	166	90	372	239

Notes. All columns report the point estimate on the regression of the (log) population in 2000 on the (log) change in urban population from 2000 to 2010 of cities with at least 100,000 inhabitants. Robust standard errors are shown in parentheses. *** is significant at the 1% level, ** is significant at the 5% level, * is significant at the 10% level.

Table 4.2: Gibrat's Law by region

	East Asia & Pacific	Europe & Central Asia	Latin America Caribbean	Middle East & North Africa	North America	South Asia	Sub-Saharan Africa
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
2000-2010	-0.310**	-0.124	-0.002	-0.101***	-0.039	-0.074**	-0.278**
	(0.156)	(0.113)	(0.062)	(0.029)	(0.024)	(0.037)	(0.111)
Constant	4.841**	1.666	0.277	1.584***	0.590*	1.313***	4.247***
	(2.230)	(1.537)	(0.764)	(0.380)	(0.306)	(0.465)	(1.507)
R-squared	0.003	0.001	0.000	0.058	0.027	0.010	0.049
No of observations	535	649	312	172	191	328	124

Notes. All columns report the point estimate on the regression of the (log) population in 2000 on the (log) change in urban population from 2000 to 2010 of cities with at least 100,000 inhabitants. Countries classified following World Bank's regional classification. Robust standard errors are shown in parentheses. *** is significant at the 1% level, ** is significant at the 5% level, * is significant at the 10% level.

A Appendix: Cross-validation of BEAM's city measurement

How does BEAM's city size estimates compare to other measures? Figure A-1 shows the BEAM-estimated urban sprawl of Chicago in the U.S., Sao Paulo in Brazil, Wuhan in China, and Pune in India in 2010, and compares them to the sprawl identified by the AUE (described in section 2.1, see (Angel et al. 2012)). Figure A-2 further adds the 2018 Google Earth high-resolution satellite imagery to the comparison.

In 2010, BEAM's urban extension exceeded that from the AUE, except for the case of Wuhan. Two reasons explain the difference. First, by focusing on continuous built up area the AUE seems not to consider the formation of metropolitan areas in which the contiguity of two cities contains suburban areas of at least one of them. Sao Paulo, for example, shows a night light continuum with Campinas, Idaiatuba and Itu (to the northwest of the city of Sao Paulo), but AUE only takes into account the main metropolitan area. Similarly, Chicago and Milwaukee reflect an interconnected market measured by city night light continuity: the northern suburbs of Chicago constitute a continuum with the southern suburbs of Milwaukee. Indeed, many workers in south Milwaukee work in northern Chicago.

From the point of view of the proposed conceptual definition of cities, namely the spatial integration of social and economic activity, these unions make sense to the extent that, over time, Sao Paulo, Chicago, and many other cities in the world have expanded to the point of sharing markets with their neighboring towns.

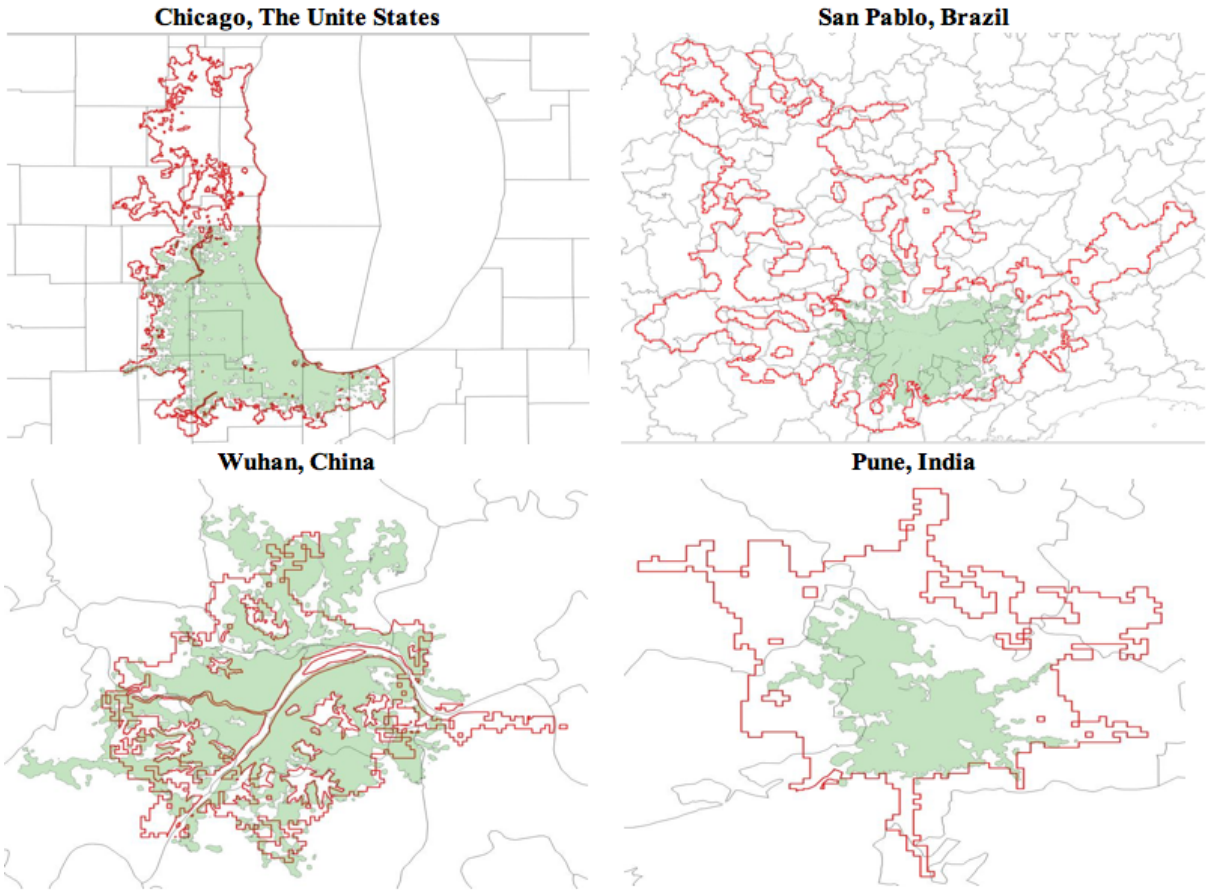
For the case of Wuhan, when comparing with Google Earth we notice in Figure A-2 that BEAM follows the continuous construction pattern even more tightly than the urban footprint defined by AUE. As exemplified by Wuhan (and also by Chicago), this implies that the sacrifice in terms of accuracy of using nighttime luminosity does not seem significant, as it follows closely the urban footprint patterns suggested by Google. In other cases (such as Sao Paulo and Pune), BEAM is less accurate by extending the main city to encompass neighboring regions. However, one has to weight this with the computational efficiency gain of BEAM.

Beyond these concrete examples, it is possible to make a systematic comparison between BEAM's full sample of cities and the universe of cities from AUE in the year 2000. In particular, Figure A-3 presents a *Spearman* correlation coefficient to get a sense of the strength and direction of the association between the city size rank of the two measures. The resulting correlation is very large (0.74), and we also reject the null hypothesis that the two city size ranks come from independent distributions (significant at the 1% level). Thus, despite of the cardinal differences, the two sources generate similar ordinal results. In other words, while BEAM's sprawl usually exceeds that of AUE because of

the consideration of suburban areas and peripheral urban slums, the city size ranking is, by and large, maintained.

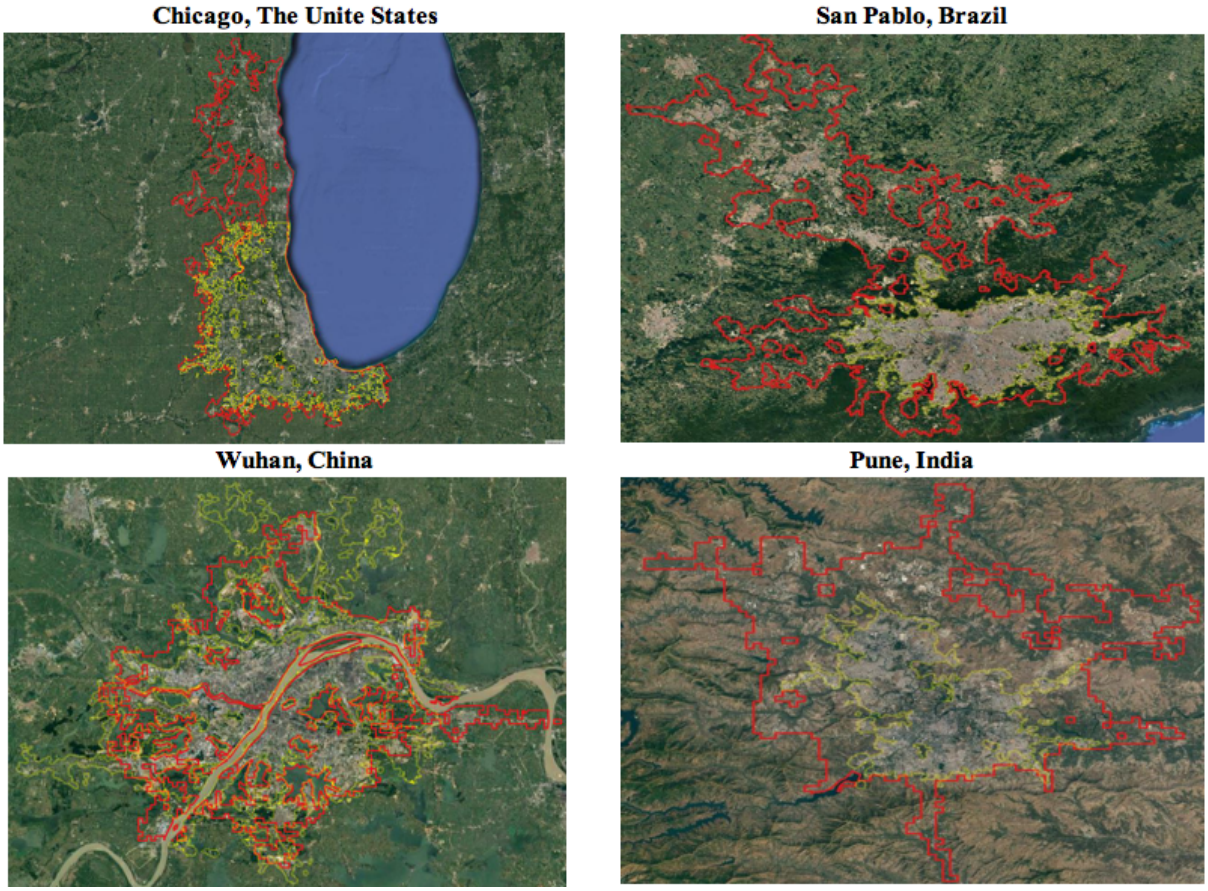
Finally, we compare the BEAM-estimated U.S. metropolitan areas to the MSAs defined by the United States Office of Management and Budget in 2010 (see Figure A-4). Although not legally constituted as administrative divisions, MSAs constitute a reference for the identification of metropolitan areas in the U.S., and are used for several statistical purposes. Out of the 383 MSAs defined for 2010, BEAM identifies 358 cases, and so the level of coincidence is 94%. The 6% discrepancy is explained, mostly, by the existence of MSAs with metropolitan areas with less than 100,000 inhabitants, the threshold established by BEAM.

Figure A-1: City sprawl: AUE and BEAM in 2010



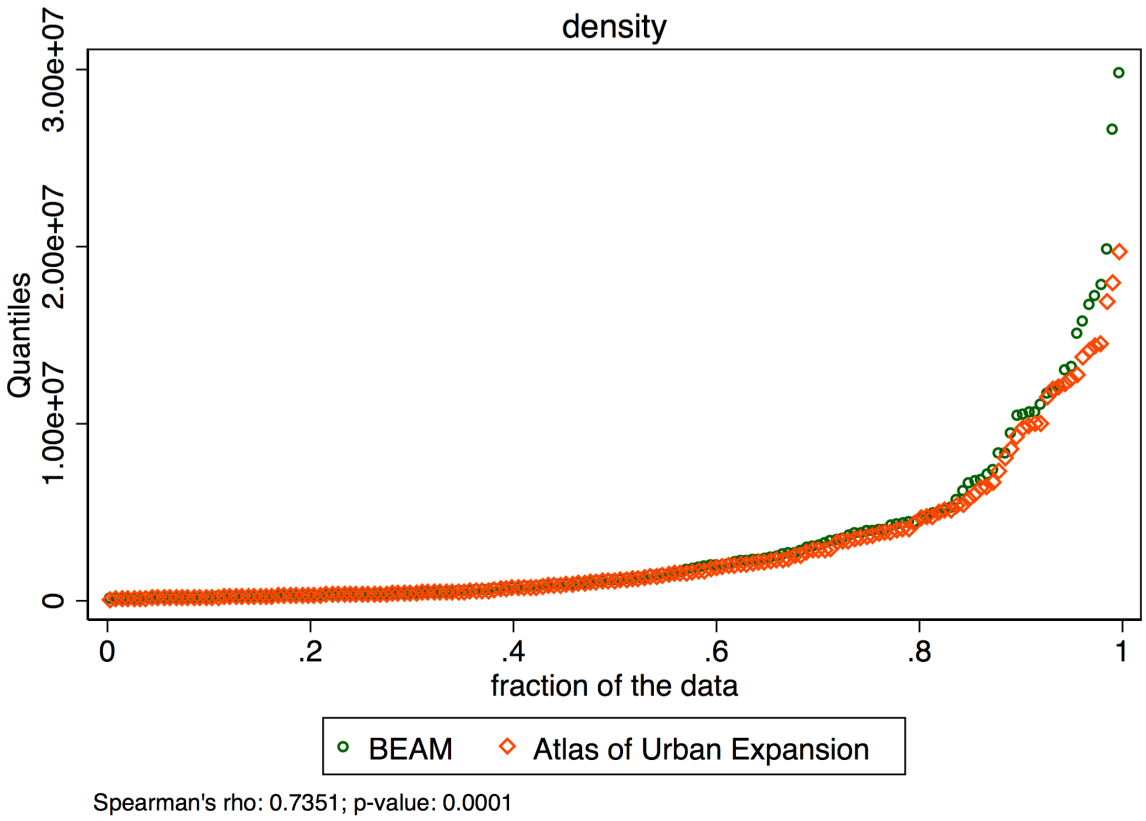
Note. The green area shows the urban sprawl as estimated by the Atlas of Urban Expansion (AUE). The red line follows *BEAM*'s city size. Gray lines represent each country's lowest administrative unit.

Figure A-2: City sprawl: AUE, BEAM and Google Earth, 2010



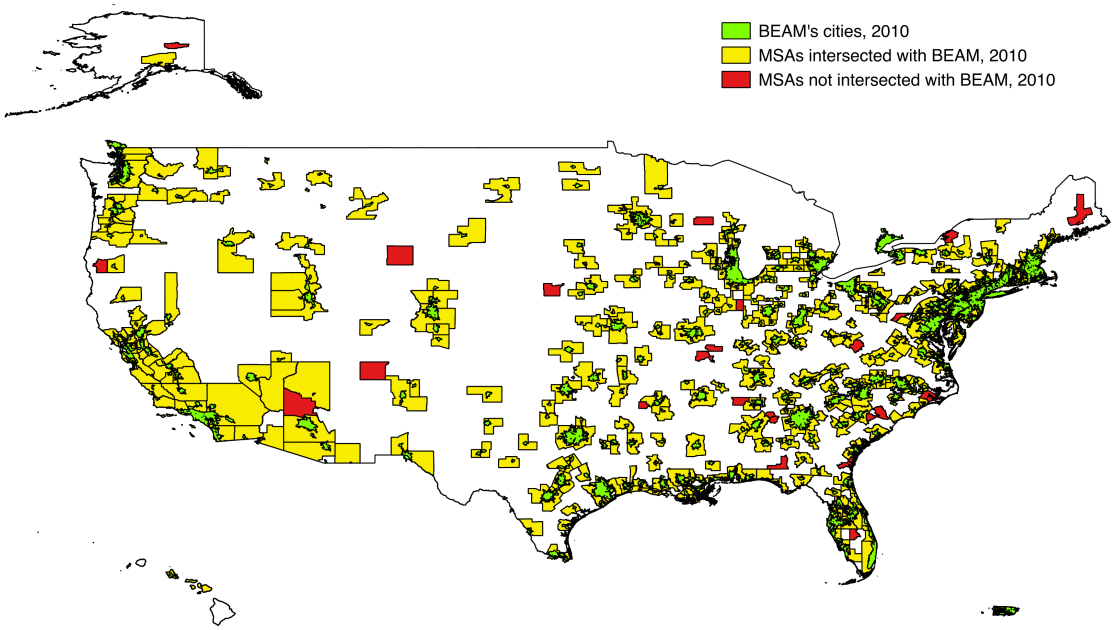
Note. The yellow line shows the urban sprawl as estimated by the Atlas of Urban Expansion (AUE) in 2010. The red line follows *BEAM*'s city size in 2010. The Google satellite image was taken in August 23th, 2018.

Figure A-3: Population rank comparison: BEAM and AUE in 2000



The graph shows the quantile to quantile comparison of city size (population) rank for the Atlas of Urban Expansion and *BEAM*. Spearman's rho rank is included, and shows that the null hypothesis of independent distributions is rejected with a significance level of 1%.

Figure A-4: BEAM's city coverage compared to Metropolitan Statistical Areas of the United States in 2010



Note. Metropolitan Statistical Areas (MSAs) from the United States Population Census, 2010. Green polygons represent *BEAM*'s; yellow polygons are MSAs that coincide with*BEAM*'s urban sprawl database, while red MSAs are those not considered by *BEAM*'s database. .