Comparing city size distributions: Gridded population vs. nighttime lights*

Miguel Puente-Ajovín[a], Marcos Sanso-Navarro[§a], and María Vera-Cabello[b]

[a]Departamento de Análisis Económico & IEDIS, Universidad de Zaragoza, Spain

[b]Centro Universitario de la Defensa de Zaragoza, Spain

Abstract

This paper compares the size distributions of cities when they are measured using gridded population and nighttime lights data. To do so, we exploit recent and accurate satellite imagery to proxy urban economic activity. Similarly to related studies, our results suggest that population is more equally distributed than lights at the country level. However, and calling assumptions established for urban nighttime lights into question, our findings do not support a Pareto function for their distribution. We also obtain evidence of a nonlinear and heterogeneous link between population and lights for a global sample of cities. Grounded on our empirical analysis, we develop a simple theoretical framework that relates the difference between the distributions of population and light emissions to the strength of agglomeration economies.

Keywords: City size distribution; Gridded population; Nighttime lights; Nonparametric methods; Urban scaling.

JEL classification: O10, O18, O57, R12.

*The authors have benefited from the valuable comments of the Editor (Siqi Zheng), two anonymous reviewers, and participants at the Brown Bag Seminars (Zaragoza), the XLVI International Conference on Regional Science (Madrid), the XXIV Applied Economics Meeting (Palma de Mallorca), the 62nd ERSA Congress (Alicante), and the 2023 Barcelona Workshop on Regional and Urban Economics (Barcelona). This work was supported by Fundación Ibercaja (Grant JIUZ-2020-SOC-11), Gobierno de Aragón (S39_23R ADETRE Research Group), and Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación (Grant PID2020-112773GB-I00).

§Corresponding author. E-mail: marcossn@unizar.es. Address: Facultad de Economía y Empresa. Departamento de Análisis Económico. Gran Vía 2. 50005 Zaragoza, Spain. Tel.: (+34) 876 554 629.

1 Introduction

The distribution of city sizes is one of the more extensively studied topics by urban economists due to its theoretical and policy-making implications. Following the seminal contributions of Gabaix (1999) and Eeckhout (2004), the related literature has mostly focused on testing whether the city size distribution fits the rank-size rule, also known as Zipf's law (Rosen and Resnick 1980). This empirical regularity quantifies the concept of urban hierarchy by stating that the size of the N-th city is 1/N times the size of the largest one. As pointed out by Arshad, Hu, and Ashraf (2018), Zipf's law is not universal, even if only the upper tail of the city size distribution is considered. The mixed evidence regarding the rank-size rule becomes especially apparent when the urban structures of different countries are analyzed (Puente-Ajovín, Ramos, and Sanz-Gracia 2020; Soo 2005). A shortcoming commonly found in these cross-country studies is that the definition of what is considered as a city differs across national data sources[1]. Fortunately, and taking advantage of satellite imagery and remote sensing techniques, several organizations and scholars have established harmonized definitions of cities and settlements that can represent all the urban areas worldwide in a homogeneous framework, making global analyses and cross-country comparisons more reliable; see Duranton (2021) and the references therein.

In the context of the study of the city size distribution, an early attempt was conducted by Decker, Kerkhoff, and Moses (2007) who, using clusters of nighttime lights (NTL, hereafter) as well as census data, tested the generality of Zipf's law across the entire range of city sizes in terms of both total area and population. These authors found more evidence against this empirical regularity for the size of urban agglomerations of light emissions than for the population of politically-defined cities. Small et al. (2011) also contend that NTL are a useful proxy – independent and complementary to population counts – for the quantification of the size, number, and spatial extent of human settlements worldwide. However, they conclude that the global size distribution of spatially contiguous patches of light emissions closely conforms to Zipf's law when they are measured in terms of area; see Jiang, Yin, and Liu (2015) for a similar analysis and findings covering all natural cities worldwide. Small et al. (2011) attribute these conflicting results to the different versions

[1] Fazio and Modica (2015), Ioannides and Skouras (2013), and Puente-Ajovín et al. (2020) show that this issue may lead to conflicting results even within a single country.

used of the 'stable night light images' collected by the Defense Meteorological Satellite Program (DMSP) Operational Linescan System.

Although there is a premise that areas with more light emissions generally have a higher population, other factors such as the level of economic activity can affect the intensity of NTL. Actually, since the pioneering contributions of Chen and Nordhaus (2011) and Henderson, Storeygard, and Weil (2012) night lights have been proven as a reliable proxy of economic activity. It is also well known that agglomeration plays a significant role in explaining why economic activity tends to concentrate where the population is dense (Glaeser 2008). Furthermore, strong agglomeration economies can, under some circumstances, lead to a situation where economic activity grows more than proportionally with population. Among the possible reasons, and in an urban context, it is worth noting the productivity gains experienced by firms and their workers in large cities due to lower transport costs, labor market pooling, and knowledge spillovers (Glaeser and Gottlieb 2009). Empirical studies about the relevance of these agglomeration economies, that cannot be observed directly, have mainly relied on regression analyses of income or productivity on city population or density (Combes and Gobillon 2015; Grover, Lall, and Timmis 2023) which, *inter alia*, have to deal with the issue of the simultaneous determination of population size and productivity (Brakman, Garretsen, and Marrewijk 2019; Ciccone and Hall 1996).

In the presence of strong increasing returns to scale at the local level, economic activity will grow disproportionately compared to population. This will widen the difference between their distributions in a given urban system, making economic activity to be more unevenly distributed than population. These arguments lead Düben and Krause (2021) to compare, at the country level, the distribution of urban population with the distribution of night lights – which may represent that of economic activity – as an alternative way of analyzing the magnitude of agglomeration economies. To do so, they calculate two alternative measures of city size by overlaying information of the geographical extent of urban centers, identified using a globally consistent scheme, with geospatial data of population and light emissions. Moreover, these authors take advantage of the data set created by Bluhm and Krause (2022) to correct the top-coding problem of DMSP night lights. Their main conclusion is that while urban population can be characterized by Zipf's law in most countries[2], this is not the case of light emissions, which are distributed less equally. Deviations from Zipf's law are mainly explained by primary cities being excessively inhabited and, especially, bright, what seems to be driven by the existence of agglomeration effects of scale and market access.

Adopting a supranational approach, and considering consistently defined functional urban areas (FUAs) and urban centres, Puente-Ajovín, Sanso-Navarro, and Vera-Cabello (2022) analyze the distributions of population and light emissions in the European Union and the United States. The main difference with the related aforementioned studies is that they exploit the more precise NTL data from the Visible Infrared Imaging Radiometer Suite (VIIRS) of instruments onboard the Suomi NPP satellite. These authors also find that urban population displays a more egalitarian distribution than light emissions, especially in the European Union. In addition, they obtain statistical evidence against both Zipf's law and a Pareto distribution for aggregate NTL within urban extents. Ribeiro et al. (2021) study the association between urban scaling – a power law relation between urban indicators and city population size (Bettencourt et al. 2007) – and Zipf's law in a global sample of FUAs using population and gross domestic product (GDP) data. At the country level, they show that there exists a direct and nonlinear relationship between the exponents that characterize these two fundamental paradigms for the science of cities (Batty 2013), suggesting that urban scaling and the distribution of population affect each other[3]. In particular, countries with a small number of large cities concentrate most complex economic activities in relatively fewer metropolises, hence intensifying the increasing returns to scale of urban GDP. Nonetheless, these authors acknowledge that other elements beyond the distribution of urban population – such as socio-economic development and historical factors – may also affect of the urban scaling of economic activity.

[2]These results disagree with those obtained by Ch, Martin, and Vargas (2021) who, after pre-processing DMSP luminosity data to correct some of its inherent problems, develop a procedure to identify the lit pixels that constitute urban footprint. These pixels are further employed to, on the basis of predetermined urban cores, construct a georeferenced data set with the location and extent of metropolitan areas of, at least, 50,000 inhabitants. This information is combined with gridded population data to estimate city sizes, finding no general support for Zipf's law at the national level in a sample of 55 countries.

[3]On the one hand, population of small cities may be attracted by the wealth and culture of large cities, as a result of agglomeration economies, affecting the distribution of city sizes in demographic terms. On the other hand, an urban system characterized by a few very large cities that agglomerate diverse economic sectors and businesses will display a higher value of the estimated urban scaling exponent.

The main aim of the present paper is to carry out a cross-country study of the difference between the distributions of urban population and light emissions making use of the current and accurate VIIRS images. Apart from providing further evidence about the global importance of urban agglomeration economies, and as a byproduct of our analysis, we are able to check the suitability of the top-coding correction of DMSP data proposed by Bluhm and Krause (2022), grounded on the assumption of a Pareto distribution for aggregate urban NTL. Given that urban agglomeration forces are not alike in developed and developing countries (Grover, Lall, and Timmis 2023), the focus is put on the differences across income groups rather than geographic regions. We assess the sensitivity of our results to the role played by primary cities, to the consideration of alternative gridded population and NTL data sets, and to the use of estimated gridded GDP to measure urban economic activity. In addition, the possible presence of a nonlinear and heterogeneous relationship between urban population and night lights has been explored using a global sample of cities. Taking these results as a starting point, we develop a theoretical explanation for the relationship between the strength of agglomeration economies and the different distributions displayed by urban population and light emissions.

The rest of the paper is structured as follows. Section 2 presents the urban units that conform our sample, and details the main sources of information from which the data exploited in our empirical analysis have been extracted. Section 3 studies the national distributions of urban population and aggregate night lights at the country level, applying parametric regressions and nonparametric tests. Section 4 evaluates the possible presence of a nonlinear and heterogeneous link between urban population and light emissions in a global sample of cities using kernel regression methods. Section 5 develops a simple theoretical framework to discuss of our main findings and, finally, Section 6 concludes. Appendices A and B contain further relevant information and results.

2 Georeferenced data: Urban centers, gridded population, and nighttime lights

The first key issue to conduct a cross-country study of the distribution of urban size is the adoption of a homogeneous definition for cities. Similarly to Düben and Krause

(2021), and for the sake of comparability, we have identified cities using the data contained in the Global Human Settlement Layer (GHSL), provided by the Joint Research Center of the European Commission; see Florczyk et al. (2019a) and Florczyk et al. (2019b). This database combines the information on built-up areas from Landsat images with the fourth version of the Gridded Population of the World[4] (GPW) to divide the globe into pixels (grid cells) of one square kilometer and classify them as belonging to a rural area or to an urban center and/or an urban cluster. Actually, GHSL urban centers correspond to the spatial extent of the cities considered in the present study, referred to the year 2015.

The GHSL consistently defines urban centers across geographical locations as areas with contiguous grid cells, where each of them has, at least, 1,500 inhabitants or 50 per cent built-up surface. In doing so, this database identifies contiguous settlements experiencing common agglomeration economies and congestion costs. Although GHSL urban centers only include areas with more than 50,000 inhabitants, this value corresponds to the threshold suggested by the World Bank (2008) to classify human settlements as urban in both developed and developing countries. The geospatial data with the shape and location of urban centers reveal that some of them belong to more than one country[5]. In these cases, we have assigned an urban area to a single country when it includes more than 75 per cent of the area. Applying this criterion, as well as only considering countries with more than 10 observations, our sample covers 12,852 urban centers of 100 countries.

Another relevant issue when dealing with urban size is its measurement. Just as the great majority of studies about the distribution of city sizes, we calculate them using population data. However, and also following Düben and Krause (2021), we exploit NTL satellite imagery to proxy urban economic activity. City size will be the sum of persons, on the one hand, and aggregate light emissions, on the other, in the pixels within the spatial extent of GHSL urban centers, according to the shapefile made available by this database. Regarding urban size measured in demographic terms, the GHSL also provides population estimates at the pixel level (GHS-POP). This information has been constructed

[4]Produced by Center for International Earth Science Information Network (CIESIN), within the Columbia University Earth Institute.

[5]The reason is that GHSL boundaries do not conform to the administrative definitions of cities, regions, or countries. Actually, some of the cities (urban centers) included in our sample contain several administrative cities.

by disaggregating GPW administrative area level population data from national censuses and registers[6] to grid cells according to their proportion of built-up area.

Although other gridded population data sets might be used (see Subsection 3.4.3), it is more reliable to build on GHS-POP for several reasons. First of all, these population data are produced by the same institution that establishes the definition of the urban units that has been adopted. In addition, the reliability of GPW estimates varies across countries, depending on the timeliness, accuracy, and spatial resolution of the census data used as an input, and on the suitability of the linear interpolation applied (Archila Bustos et al. 2020). The LandScan database refers to ambient population that, in contrast to resident population, not only represents where people live, but also where they work and travel. Leyk et al. (2019) suggest to use gridded population data constructed using information on human settlements or urban extents, such as GHS-POP, to study the distribution of urban population. Lastly, Chen et al. (2020) claim that this database is more opportune to analyze highly-urbanized areas.

To carry out their empirical analysis, Düben and Krause (2021) rely on the data set created by Bluhm and Krause (2022) to correct the top-coding problem of DMSP images. However, these NTL data are also affected by blurring, geolocation errors, lack of calibration, and coarse resolution; see Gibson (2021) and Gibson et al. (2021). Since April 2012, there are available more precise NTL images captured by the VIIRS onboard the Suomi NPP satellite. Its Day/Night Band was designed to measure the radiance of lights on earth in a wide variety of lighting conditions, and covers a dynamic range of about seven orders of magnitude (DMSP covers less than two), avoiding saturation problems and top-coding. VIIRS images are comparable over time and space, do not have blurring or geolocation errors, and display, at least, 45 times greater spatial resolution than DMSP data (Elvidge et al. 2017). For all these reasons, VIIRS images are superior at attributing lights to the place where they are emitted and, therefore, are a better proxy for urban economic activity than DMSP data; see Gibson, Olivia, and Boe-Gibson (2020) for a comparison of these two alternative NTL satellite imagery.

In the manner of Puente-Ajovín, Sanso-Navarro, and Vera-Cabello (2022), and as suggested by Gibson (2021) and Gibson et al. (2021), we use these VIIRS night lights to

[6] Adjusted to match estimates from the United Nations World Population Prospects.

proxy urban economic activity. More specifically, we have extracted the 'vcm-orm-ntl' annual composites[7] for 2015 from the website of the Earth Observation Group of the National Oceanic and Atmospheric Administration (US Department of Commerce)[8]. These data have been cleaned to exclude background noise, solar and lunar contamination, cloud cover degradation, and features unrelated to electric lighting (Elvidge et al. 2017). At the pixel level, reported radiance values are expressed in nano Watts per square centimeter per steradian, with a resolution of 15 arc seconds (approximately 450 meters at the equator). The same as gridded population, NTL data have been aggregated for all pixels included within the extents of urban centers to calculate their size. Although the pixels of VIIRS data are smaller than GHSL ones, this is not problematic because the aggregation of light emissions has been carried out considering the larger GHSL pixels.

[Insert Table 1 about here]

Table 1 reports descriptive statistics for the two measures of city size described above. This is done for the whole sample as well as by country income group, according to the World Bank classification[9] for 2015. It categorizes countries as 'Low income' if their Gross National Income (GNI) per capita was lower or equal than 1,025 U.S. Dollars (22 out of 100 countries in our sample); 'Lower-middle income' if it was between 1,026 and 4,035 USD (29); 'Upper-middle income' between 4,036 and 12,475 USD (27); and 'High income' if GNI per capita was higher than 12,475 USD (22). Average and median city size increase with the level of income, both in terms of population and aggregate light emissions. Nonetheless, this increase is more than proportional in the case of NTL as compared to population. Except in high income countries, there are cities for which no lights are attributed. It can also be observed that the largest cities in terms of aggregate NTL are located in countries that belong to the high income group.

[7]VIIRS Cloud mask–Outlier removed–Nighttime lights.
[8]https://www.ngdc.noaa.gov/eog/.
[9]See Table A1 in Appendix A for further details.

3 The distribution of urban population and aggregate night-time lights at the country level

3.1 Rank-size parametric regression

The rank-size rule implies that the city size distribution can be approximated by a Pareto function with power law exponent equal to one. For this reason, cross-sectional empirical analyses of the Zipf's law are generally based on a log-log linear regression between the rank of a city and its size. In order to reduce the bias of the OLS estimator in small samples, Gabaix and Ibragimov (2011) propose the following regression model:

$$log\left(Rank_i - 0.5\right) = \alpha - \beta \cdot log(Size_i) + \epsilon_i, \quad i = 1, \dots, n; \tag{1}$$

where i is a city indicator, and n denotes the sample size. Zipf's law is equivalent to $\beta = 1$. In the present context, a coefficient lower (greater) than one reflects that population and/or light emissions are more unequally (equally) distributed across the national urban system than predicted by the rank-size rule.

[Insert Figure 1 about here]

Figure 1 shows kernel densities for the estimated slope parameter in expression (1) at the country level[10], measuring city size in demographic terms (GHSPOP, orange) and when urban economic activity is proxied using NTL (VIIRS, blue). Estimated power law exponents are centered around values slightly higher than one when city size is calculated using gridded population. However, Pareto coefficients tend to be lower than one when urban size is expressed in terms of aggregate light emissions. Therefore, and corroborating the findings of Düben and Krause (2021) and Puente-Ajovín, Sanso-Navarro, and Vera-Cabello (2022), urban NTL are more unevenly distributed than population at the country level.

[10] Papua New Guinea has been omitted as an outlier. The estimated slope parameter in the rank-size regression for this country is 2.91 when city size is measured in population terms.

3.2 Nonparametric testing

The main purpose of the empirical model in expression (1) is to test the null hypothesis that the Pareto coefficient is equal to one; i.e., that Zipf's law holds. As a more flexible alternative, Gan, Li, and Song (2006) propose to investigate the distribution of city sizes through the implementation of the Kolmogorov-Smirnov (KS) test statistic. The idea is that this nonparametric method can be used to compare the city size distribution with a function of reference, determining the degree of (dis)similarity. With this aim, we have considered two benchmarks: (i) a Pareto function imposing that the power law exponent is equal to one (exact Zipf's law), and (ii) a Pareto function with the estimated β coefficient in expression (1) as the power law exponent.

The empirical distribution function of the n independent and identically distributed ordered size observations can be calculated as:

$$F_n(s) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,s]}(Size_i); \tag{2}$$

where $1_{(-\infty,s]}(Size_i)$ is an indicator function that takes a value equal to one if $Size_i \leq s$, zero otherwise.

The Pareto distribution function is given by:

$$F_P(s,\beta) = 1 - \left(\frac{Size_i}{s}\right)^{\beta}. \tag{3}$$

The calculation of the KS test statistic is based on the maximum difference between the empirical distribution of the data and the benchmark function:

$$KS = sup|F_n(s) - F_P(s,\beta)|. \tag{4}$$

The null hypothesis is that the observed data have been obtained from the probability distribution of reference. The resulting test statistic is compared to the critical values of the KS distribution to assess the validity of the reference function, such that the smaller the value of the test statistic the better the reference distribution function describes city sizes.

[Insert Figures 2 and 3 about here]

We have first implemented the KS test against the null hypothesis that, at the country level, city sizes are distributed as a Pareto function with power law exponent equal to one. The cumulative distribution function of the p-values that have been obtained for the two alternative measures of city size are plotted in Figure 2. In line with the kernel densities of estimated Pareto coefficients shown in Figure 1, the null hypothesis that city sizes adjust to Zipf's law can be more easily rejected when they are calculated using light emissions. As noted before, the KS test has also been calculated using the OLS estimate for the slope parameter in (1) as the power law exponent. The corresponding cumulative distribution functions displayed in Figure 3 show that, although there is a slightly higher evidence of a Pareto distribution for aggregate urban NTL, the null hypothesis can be rejected in 75 countries at the 1% significance level. Thus, we do not find supportive evidence using VIIRS images for the Pareto assumption established by Bluhm and Krause (2022) to correct for top-coding in DMSP data. This problem mainly affects larger cities which, according to the figures reported in Table 1, tend to be located in richer countries. For this reason, and given that agglomeration effects are not alike in developed and developing nations, we also investigate the distribution of city sizes grouping countries by their level of income per capita.

3.3 Country income groups

Kernel density estimates of national Pareto coefficients by country income group are plotted in Figure 4. The greatest resemblance between the distributions of urban population and NTL is found in high income countries. Aggregate urban light emissions are, however, more unevenly distributed than population. The similarity between the distributions of population and night lights is directly related to national income. In particular, estimated Pareto coefficients for population (lights) tend to increase (decrease) when GNI per capita decreases. These results corroborate the existing evidence that urban agglomeration effects are more important in developing countries; see Grover, Lall, and Timmis (2023). The explanations given by these authors for this fact are the existence of duality in labor and land markets, the poor quality of physical and transport infrastructures, and the favoritism of governments towards the largest cities.

[Insert Figure 4 about here]

The upper panel of Table 2 reports, for three significance levels, the percentage of rejections by the KS test of the null hypothesis that the city size distribution is a Pareto function with power law exponent equal to one. Corroborating the results in Figures 2 and 3, there is more evidence against the fullfilment of Zipf's law in the urban distribution of aggregate NTL than in the distribution of population when all countries in our sample are considered. Broadly speaking, nations with higher income per capita tend to display lower rejection rates than less developed countries (LDCs). The lower panel of Table 2 shows similar results when the KS test statistic is performed considering that the distribution of reference is a Pareto function with the estimated slope parameter in the rank-size regression as the power law exponent. In this case, and as expected, the evidence of a Pareto distribution for both urban population and light emissions is slightly higher than that for the exact Zipf's law. Nonetheless, the rejection rates for aggregate VIIRS night lights at the city level – higher than 50 per cent – do not support the Pareto assumption established by Bluhm and Krause (2022) to correct the top-coding problem of DMSP data.

[Insert Table 2 about here]

3.4 Robustness checks

3.4.1 The role of primary cities

The estimated Pareto coefficient from a rank-size regression at the country level can be interpreted as an indication of the degree of hierarchy in its urban system, such that a low coefficient reflects a high weight of large cities. In this line, Düben and Krause (2021) show that national primary shares[11] are inversely related to the magnitude of Pareto coefficients for city sizes calculated both in demographic terms and by aggregating light emissions. Nonetheless, primary cities may be outlying observations according to a power law, hence affecting the fit and estimated coefficients from rank-size regressions (Brakman, Garretsen, and Marrewijk 2019). To check whether this is the case in the present context, we re-estimate expression (1) after removing the largest city in terms of population[12] from each

[11]Urban primacy is a well-known feature of urbanization in LDCs (Duranton 2008), mainly driven by political and institutional factors (Ades and Glaeser 1995; Davis and Henderson 2003).

[12]This definition of primary city is the same to that used in related papers, see Düben and Krause (2021) and Bluhm and Krause (2022). Removing the largest city in terms of aggregate NTL in the corresponding

national sample. Resulting kernel densities of estimated Pareto coefficients by country income group are displayed in Figure 5. It can be observed that the main conclusions drawn in the previous subsections do not change when primary cities are excluded from national samples. That is, urban aggregate light emissions are less equally distributed than population, and the similarity between the distributions of NTL and population increases with national income.

[Insert Figure 5 about here]

The distributions of estimated parameters shown in Figure 4 tend to move to the right – reflecting higher values and, consequently, lower urban concentration – when primary cities are not included in national samples. The magnitude of the distributional shift is greater when city size is measured in terms of population. Hence, excluding the largest city makes aggregate urban light emissions and, especially, population to be more evenly distributed. Given that the estimated Pareto coefficient can be considered as an indirect indicator of city primacy (Ch, Martin, and Vargas 2021), this can be interpreted as evidence that fewer large cities dominate the urban landscape in demographic than in economic terms. Moreover, and for a given country, the size of the largest city seems to be more in line with the power law distribution of the other urban units when they are measured using aggregate light emissions than population. It can also be observed that changes affect all income groups in a similar manner, what can be related to the fact that developing countries have become as likely as developed countries to contain large agglomerations, mainly when they are measured in demographic terms (Jedwab, Loungani, and Yezer 2021).

In short, this robustness check excluding primary cities from national samples allows us to claim that the different distributions of light emissions and population are not driven by an excessive concentration in the largest urban centers. Actually, not considering primary cities lead to even greater differences between the estimated Pareto coefficients for the two alternative measures of urban size, especially in countries with lower income levels. This novel result obtained using more accurate satellite imagery, not affected by top-coding, than related studies can be an indication that agglomeration economies are not restricted to the largest cities, even in developing countries.

rank-size regression at the country level does not significantly change the results (available from the authors upon request).

3.4.2 Alternative nighttime lights data

For the sake of comparison, we have also proxied local economic activity with the 'stable night light images' collected by the DMSP, despite their limitations. Given that the production of DMSP images ended in 2013, we have used the information for that year. In addition, the top-coding correction of DMSP data proposed by Bluhm and Krause (2022) – referred to as DMSP_BK[13] in tables and figures – has been used to provide a broad perspective of all NTL data sources available, and to check the robustness of the results about the distribution of city sizes measured by aggregating urban light emissions to their choice.

[Insert Figures 6 and 7 about here]

Figure 6 shows that the density functions for the estimated slope parameters from expression (1) at the country level using DMSP and VIIRS images are alike. However, the distribution of Pareto coefficients resulting from DMSP corrected data is more leptokurtic. This finding points out that the top-coding correction proposed by Bluhm and Krause (2022) exerts a non-negligible influence on the estimated parameters from country rank-size rule regressions. Kernel densities plotted in Figure 7 show that the greatest similarity of estimated Pareto coefficients for urban aggregate NTL is found in lower-middle income countries. This result reflects that this group is not greatly affected by the top-coding problem of DMSP images. Even if this was also expected to be the case of low income countries, the distributions of estimated slope parameters for VIIRS and DMSP-based data are different in this group. This implies that the higher accuracy of VIIRS images, capturing the full brightness of the largest urban centers, allows the coefficients that characterize the city size distribution to better reflect the higher degree of concentration of urban economic activity in LDCs.

[Insert Table 3 about here]

Table 3 reports the percentage of rejections by the KS test of the null hypothesis that the city size distribution is a Pareto function with power law exponent equal to one (Panel

[13] Available at `https://lightinequality.com/`.

A), and that the distribution of reference is a Pareto function with the estimated slope parameter in the country rank-size regression as the power law exponent (Panel B). Obtained results for both the original and corrected DMSP images are similar to those in Table 2 for VIIRS data. Nonetheless, and with the exception of upper-middle income countries, there is a larger amount of evidence against Zipf's law and a Pareto distribution in urban economic activity when it is proxied using VIIRS images than with DMSP-based data.

3.4.3 Alternative gridded population data

Apart from GHS-POP, there are other global gridded population data sets intended to overcome the inconsistencies in the information provided by national censuses. In fact, it is by decoupling these data from their original administrative boundaries how population can be aggregated to other units such as urban centers. The differences across these gridded population databases are determined by the nature of the input data and the modeling approach adopted; see Leyk et al. (2019) and Archila Bustos et al. (2020) for two systematic reviews[14]. This subsection deals with the sensitivity of our results about the national distribution of urban population to the use of three alternative mainstream spatialized population data sets: GPW, LandScan, and WorldPop.

GPW implements the simplest method to redistribute the data from the administrative unit scale to the grid size (areal interpolation) by assuming that population is evenly distributed in space (areal weighting). Using remote sensing satellite imagery and geographic information, GSH-POP generates built-up areas and, according to their proportion in each grid and overlooking administrative boundaries, decomposes GPW data again using a dasymetric mapping method based on linear regression. In order to disaggregate subnational census data, LandScan and WorldPop adopt highly-modeled frameworks that consist of implementing dasymetric mapping with more sophisticated statistical techniques – dynamically adaptable and random forest algorithms, respectively – and broad ancillary data sets including land cover, roads, slope, and NTL, *inter alia*.

[Insert Figure 8 about here]

Figure 8 plots kernel densities for the estimated slope parameters from country rank-size regressions using the four gridded population data sets to calculate the size of urban

[14] See also the POPGRID Data Collaborative (https://www.popgrid.org/).

centers. This graph shows that the differences between the distributions of estimated Pareto coefficients are more evident than those found comparing NTL data sources. More specifically, the use of the three alternative gridded population data sets to measure city size in demographic terms results in a more uneven distribution of urban population at the country level, similar to that of light emissions. This is especially the case of LandScan and WoldPop, what can be related to their highly-modeled frameworks that exploit information capturing economic activities, and by the correlations between the variables included in their corresponding ancillary data sets. Actually, and among other information, it is worth noting that WorldPop relies on DMSP images to generate its population density predictions. In addition, it should be acknowledged that LandScan data refer to ambient population, which not only represents where people live, but also where they travel and work.

[Insert Figure 9 about here]

Figure 9 displays the distributions of areto coefficients using the four gridded population data sets and by country income group. It can be observed that the differences between kernel densities are inversely related to national income per capita. Urban sizes calculated using the GPW present higher levels of concentration and, with the exception of developed countries, tend to display an average value around 0.5. As can be inferred from the descriptive statistics reported in Table A2 in Appendix A, GPW, LandScan and, to a lesser extent, Worldpop tend to underestimate the size of smaller urban centers as compared to GHS-POP, while this is not the case larger ones. This leads to an apparently more unequal distribution of population across urban centers and, as a result, lower estimated Pareto coefficients. Corroborating these findings, Table 3 shows that the rejection rates of the KS test for the three alternative gridded population data sets considered in this robustness check are much higher than those for GHS-POP data for both the null hypothesis of exact Zipf's law and of a Pareto distribution function.

The similarity between the distributions of estimated slopes from rank-size regressions using LandScan and WorldPop data and the distribution with information from GPW (GHS-POP) decreases (increases) with national income per capita. This may be a reflection of the strong assumption established by GPW that population is equally distributed across administrative areas, on the one hand, and the lower data quality of national censuses and ancillary variables in LDCs, on the other; see Appendix B for a more elaborated explanation grounded on the uniform areal weighting approach implemented by GPW, and on the different administrative divisions considered across countries.

3.4.4 Economic activity measured using gridded GDP data

The patterns displayed by the estimated Pareto coefficients using alternative gridded population data sets point to a more uneven city size distribution, closer to that of urban aggregate light emissions, than the one obtained with GHS-POP. This leads us to ask whether the disparity found between national city size distributions using GHS-POP and VIIRS is really due to the different distributions of population and economic activity, or is simply as a consequence of the different nature of the two types of data. While gridded population is estimated using distinct frameworks, NTL are not grounded on any economic statistics. Therefore, and although there is an extant evidence that light emissions can be regarded as a proxy for urban economic activities[15], there might be some remaining concerns about whether this is the case in the present context.

Trying to make it more convincing that the difference between the distributions of city sizes calculated using GHS-POP and VIIRS data reflects the dissimilarity between the distributions of urban population and economic activity, we will check further how reliable and valid is to use VIIRS images to represent economic activities. With this aim, we exploit the gridded global data set for gross domestic product[16] (GDP) estimated by Kummu, Taka, and Guillaume (2018). The use of this data set is more convenient in our context than existing alternatives (Chen et al. 2022; Wang and Sun 2022) due to its spatial and temporal resolutions, underlying input data, and modelling approach adopted. In particular, and among other information, these authors provide GDP estimates in 30 arc-seconds resolution for the year 2015, expressed in 2011 (International) United States Dollars. Making use of both national and subnational information sources, they implement

[15] See Bluhm and Krause (2022), Phan (2023), and the references therein.

[16] As pointed out by Chen and Nordhaus (2019), economic statistics provided by governments and/or international organizations present inconsistencies in terms of definitions, measurement, and time frame. On the contrary, NTL avoid errors related to misreporting or methodological differences. Given that light emissions are measured objectively, updated regularly, and cover most of the globe, they can be considered as a more reliable source in predicting GDP values at different geographical levels. Actually, Hu and Yao (2022) exploit NTL to improve national accounts GDP growth measures.

areal weighting techniques to redistribute input data into grid cells. While the national GDP per capita[17] data come from the Central Intelligence Agency (CIA World Factbook) and the World Bank databases, the subnational information is based on Gennaioli et al. (2013). It is worth noting that, although Kummu, Taka, and Guillaume (2018) does not consider auxiliary variables, such as NTL, they use the GHSL population data to calculate GDP values in absolute terms.

In the manner of related work on this topic (Bluhm and McCord 2022; Gibson 2021), we study the predictive relationship between aggregate urban light emissions and gridded GDP using the following regression:

$$\log(GDP_i) = \phi + \theta \log(VIIRS_i) + \xi_i, \quad i = 1, \ldots, n; \tag{5}$$

Estimation results are reported in Table 4. Considering all the countries covered in our sample[18], we find that the variation in aggregate urban VIIRS light emissions predicts more than half the variation in GDP, as reflected by the coefficient of determination ($R^2 =$ 0.52; 0.64 if country fixed effects are included in the regression). When countries are grouped according to their level of development, this percentage increases in a 30% in high income countries, being the estimated elasticity closer to unity[19]. Both the estimated slope parameter and coefficient of determination decrease with national income. These results re similar to those obtained by Phan (2023), who shows that institutional quality and the level of development are two of the most important factors in explaining the difference between luminosity data and GDP across countries.

[Insert Table 4 about here]

Expression (5) has also been estimated with aggregate urban GHSL population on the left-hand side; i.e., as the dependent variable. Table 4 shows that 16% of the variance in overall urban population can be explained by light emissions. Thus, it can be claimed that their predictive power is more than three times higher for GDP than for population (2.5

[17] Purchasing power parity (PPP).

[18] Urban centers with a null estimated aggregate level of GDP have been removed from the analysis (Ribeiro et al. 2021).

[19] Given that GDP in developing countries is particularly error-prone and could be subject to manipulation (Keola, Andersson, and Hall 2015), the consideration of high-income countries separately can be understood as a benchmark for assessing the success of NTL as a proxy for economic activity (Gibson 2021).

times if country fixed effects are considered). When countries are grouped by their level of development, this higher predictive ability seems to be especially relevant for lower-middle income countries, advantage that also appears to be important in low income countries when country dummies are introduced in the regression.

[Insert Figures 10 and 11 about here]

Figure 10 plots kernel densities for the estimated slope parameters in national rank-size regressions – expression (1) – measuring cities using aggregate urban NTL, population and GDP. The same as light emissions, GDP is more unevenly distributed than population. In fact, the KS test statistic rejects more easily the null hypothesis that the distribution of Pareto coefficients for GDP is equal to that of population (p-value=0.00), than to that of lights (p-value=0.05). Figure 11 shows the same results, but grouping countries by their income level. These graphs make much more evident the differences between the national distributions of urban economic activity and population. In line with the results reported in Table 4, the similarity between kernel densities of Pareto coefficients for NTL and GDP increases with the level of national income. This is corroborated by the KS test statistic, that rejects an equal distribution of national urban light emissions and GDP only in the sub-sample of lower-income countries (p-value=0.03). Furthermore, and even if GHSL population data has been exploited by Kummu, Taka, and Guillaume (2018) to calculate their gridded GDP in 30 arc-seconds resolution, the KS test statistic rejects an equal distribution of Pareto coefficients for population and GDP in all cases at the 5% significance level[20].

To sum it up, the estimations carried out in this last robustness check suggest that, in general, aggregate urban light emissions are more useful in predicting GDP than population at the city level. Moreover, resulting Pareto coefficients from country rank-size regressions for urban gridded GDP are distributed more similarly to those obtained using VIIRS images than to those for GHS-POP data. These findings reinforce the claim that the differences observed in the national city size distributions using GHS-POP and VIIRS data reflect the disparities between the distributions of urban population and economic activity at the country level.

[20]Results available from the authors upon request.

4 The heterogeneous and nonlinear relationship between urban population and light emissions

This section takes a closer look at the relationship between urban population and light emissions by assessing the possible presence of heterogeneity and nonlinearities. With this aim, we implement nonparametric kernel regression methods that do not require a priori assumptions on the underlying functional form, and that provide observation-specific estimates.

A fully nonparametric specification to estimate the elasticity of urban light emissions to population is:

$$Lights_i = \mathrm{m}(Popul_i) + \varepsilon_i, \quad i = 1, \dots, n; \tag{6}$$

where $Lights_i$ denotes the logarithm of aggregate NTL in city i, $Popul_i$ is the logarithm of its number of inhabitants, ε_i is a zero-mean additive error, and $m(\cdot)$ is the smooth unknown function for the conditional mean. This function can be estimated by locally averaging the aggregate night lights of the urban centers with a similar size in demographic terms. This method is known as the local-constant – or Nadaraya-Watson – kernel estimator:

$$\hat{\mathrm{m}}(Popul) = \sum_{i=1}^{n} w_i Lights_i. \tag{7}$$

Weights are non-negative, their sum is equal to one, and they are given by:

$$w_i = \frac{\mathrm{K}\left(\frac{Popul_i - Popul}{h}\right)}{\sum\limits_{j=1}^{n} \mathrm{K}\left(\frac{Popul_j - Popul}{h}\right)}, \tag{8}$$

with $\mathrm{K}(\cdot)$ being a kernel function.

The amount of information used to calculate the local average is determined by the bandwidth h. A data-driven method to select this smoothing parameter is least-squares cross-validation (LSCV), which consists of choosing h so as to minimize

$$\mathrm{CV}(h) = \frac{1}{n} \sum_{i=1}^{n} \left[Lights_i - \hat{\mathrm{m}}_{-i}(Popul_i)\right]^2 \mathrm{M}(Popul_i), \;\; 0 \le \mathrm{M}(\cdot) \le 1; \tag{9}$$

where $\mathrm{M}(\cdot)$ is a weighting function[21], and

$$\hat{\mathrm{m}}_{-i}(Popul_i) = \frac{\sum_{l \neq i}^{n} Lights_l \mathrm{K}\left(\frac{Popul_i - Popul_l}{h}\right)}{\sum_{l \neq i}^{n} \mathrm{K}\left(\frac{Popul_i - Popul_l}{h}\right)}. \tag{10}$$

The criterion in expression (9) is a trimmed version of the sum of squared residuals from a leave-one-out estimator of the conditional mean function. LSCV bandwidth selection, in conjunction with the local-constant kernel estimator detects irrelevant regressors, which will be smoothed out as

$$\mathrm{K}\left(\frac{Popul_i - Popul}{h}\right) \to \mathrm{K}(0) \quad \text{when} \quad h \to \infty. \tag{11}$$

Instead of the local-constant approximation, a linear regression can be fitted for urban centers with a similar number of inhabitants. When a weighting function is included with this purpose, the estimation method is known as the local-linear kernel regression. The aim is to estimate the following expression:

$$Lights_i = a + b'(Popul_i - Popul) + e_i, \quad i = 1, \ldots, n; \tag{12}$$

In particular, the estimation is based on solving the following optimization problem:

$$\min_{a,b} \sum_{i=1}^{n} [Lights_i - a - b'(Popul_i - Popul)]^2 \mathrm{K}\left(\frac{Popul_i - Popul}{h}\right). \tag{13}$$

It has been demonstrated that the solutions $\hat{a} = a(Popul)$ and $\hat{b} = b(Popul)$ are consistent estimators of the conditional mean function, and of its partial derivative $m^{(1)}(Popul) = \partial \mathrm{m}(Popul)/\partial Popul$, respectively (Li and Racine 2007).

The local-linear kernel estimator nests OLS as a special case for sufficiently large values of the bandwidth parameters. Moreover, the LSCV bandwidth selection rule in the local-linear framework has the ability to assign a small value of h for regressors that have a nonlinear relationship with the dependent variable. Given that the kernel applied in the empirical analysis will be the Gaussian function, two times the sample standard deviation

[21] Following Racine and Li (2004), we have set $\mathrm{M}(\cdot) = 1$

of continuous covariates will be considered as the upper bound for their bandwidth; unity for the smoothing parameters of discrete regressors.

For the sake of comparison with the results obtained[22] by Düben and Krause (2021), Table 5 reports the estimated elasticities from fitting standard parametric OLS regressions to the relationship between urban lights and population in (6). In this case, the estimations are carried out using the whole sample of urban centers. Given the cross-sectional nature of our data set, we only include country fixed effects to control for unobserved heterogeneity as additional regressors. The estimated elasticities are of a higher magnitude than those previously found in the literature. Just as the existing evidence, the response of light emissions to population is lower in larger cities. However, and as a novelty, we conclude that an increase in the population of primary cities is associated with a less than proportional increase in their aggregate light emissions.

[Insert Tables 5 and 6 about here]

The upper panel of Table 6 reports the bandwidth parameters selected using the LSCV method in a local-constant kernel regression framework. The magnitude of this smoothing parameter is below its upper bound for population in all specifications, implying that this variable is relevant to explain differences in urban light emissions worldwide. While this is also the case of the indicator variables for the primary and the 10 largest cities, as well as for country income groups, the bandwiths for their interactions with population are above their corresponding upper bounds. The only exception is the interaction term included to capture a differential response of urban NTL to population in low income countries. The middle panel of Table 6 shows the selected smoothing parameters for a local-linear kernel estimation. These figures suggest that, in general, there is a nonlinear relationship between night lights and population. This result is corroborated by the diagnostic test statistic developed by Hsiao, Li, and Racine (2007), reported in the lower panel, that rejects both a standard linear OLS model (HLR1) and a quadratic specification for population (HLR2) in favor of the estimated nonparametric regression.

Table 7 contains descriptive statistics of the distribution of the estimated partial effects for population using a local-linear kernel regression, and the bandwidth parameter reported

[22] See Table 3, page 201. Estimated elasticities using DMSP data for our sample can be found in Table A3 in Appendix A.

in the middle panel of Table 6 for the specification that only includes country fixed effects as additional regressors. These gradients show that the elasticity of urban light emissions to population is heterogeneous. Although the response of NTL to population tends to be lower in larger cities, the difference in the magnitude of estimated elasticies using the whole sample of urban centers is less important than when they are classified according to country income groups. In particular, the figures displayed in the lower panel of Table 7 show that the elasticity of urban lights to population sharply decreases with the level of development.

[Insert Table 7 about here]

5 Discussion

The results reported in Table 7, obtained considering all urban centers that conform our sample, can be theoretically related to the kernel densities of Pareto coefficients by income group displayed in Figure 4, estimated from rank-size regressions at the country level. To do so, let us begin by noting that, abstracting from the error term, expression (1) is equivalent to

$$Rank_i - 0.5 = e^{\alpha} e^{log\left(Size_i^{-\beta}\right)}. \tag{14}$$

Taking into account the two measures of urban size that have been studied throughout our empirical analysis, it can be stated that

$$Rank_i - 0.5 = ALights_i^{-\beta_L}, \tag{15}$$

and

$$Rank_i - 0.5 = BPopul_i^{-\beta_P}; \tag{16}$$

with β_L and β_P being, respectively, the Pareto coefficients that characterize the national distributions of urban light emissions and population. $A = e^{\alpha_L}$ and $B = e^{\alpha_P}$, with α_L and α_P two constant terms.

There is a recent strand of the literature showing that most urban properties vary continuously with population size; see, among others, Bettencourt et al. (2007), Bettencourt

(2013), and Lobo et al. (2013). This empirical observation has been described mathematically using power law scaling relations. On the basis of this formal framework, the relationship between urban light emissions and population can be written as

$$Lights_i = CPopul_i^{\gamma}, \tag{17}$$

where C is a normalization constant, and γ denotes the scaling exponent which, in the present context, corresponds to the elasticity of urban aggregate NTL to population at country level.

As long as $\gamma > 0$, it can be claimed that $Lights_i > Lights_j$ if $Popul_i > Popul_j$. Therefore, the rank of a given city i will not depend on the variable used to calculate its size:

$$Rank_i - 0.5 = ALights_i^{-\beta_L} = BPopul_i^{-\beta_P}. \tag{18}$$

Dividing this expression for the primary city and for an arbitrary urban center of rank r, and taking into account the scaling relation in (17), it is obtained that

$$\left(\frac{Lights_1}{Lights_r}\right)^{\beta_L} = \left(\frac{Popul_1}{Popul_r}\right)^{\gamma\beta_L} = \left(\frac{Popul_1}{Popul_r}\right)^{\beta_P}. \tag{19}$$

This implies that there exists a linear relationship between the Pareto coefficients characterizing the distributions of urban population and light emissions that depends on the scaling exponent:

$$\beta_P = \gamma\beta_L. \tag{20}$$

The results from rank-size regressions at the country level presented in Section 3 show that the estimated Pareto coefficients for the distributions of city sizes calculated using gridded population tend to be higher than those obtained aggregating light emissions within urban extents. According to expression (20), this is equivalent to saying that the elasticity of NTL to population is greater than one, and is precisely what we find in Section 5 considering a global sample of urban centers.

A scaling exponent greater than one is interpreted as evidence of a super-linear urban scaling regime, illustrated by the concept of agglomeration economies; see Duranton and

Puga (2004). Therefore, we are providing both further empirical evidence and theoretical support for the main hypothesis put forward by Düben and Krause (2021) that lights should be distributed more unevenly than population under supra-linear scaling. It implies that per capita economic output – as well as other socio-economic indicators such as wages or new inventions – increases with city population size (Bettencourt et al. 2007). That is, cities of different sizes display different features because, as complex systems, they are not only concentrations of people, but also of social interactions (Jacobs 1969). This reflects the role played by the 'second nature' factors that shape the distribution of economic activity across space through the interactions between agents and the increasing returns to scale created by dense interactions (Krugman 1991, 1993; Venables 2005). Thus, it is the importance of population size as a determinant of the socio-economic activity that takes place in urban centers what makes the distribution of aggregate NTL to be more uneven than that of population.

The statistics that describe the distribution of the estimated gradients at the urban center level reported in Table 7 show that the elasticities of light emissions to population significantly change across country income groups. These gradients tend to be slightly higher than one for cities in high income countries, explaining that this group displays the greatest similarity between the distributions of estimated Pareto coefficients for urban population and aggregate NTL. It can also be observed that the magnitude of the elasticities is inversely related to national income per capita what, in line with expression (20), explains that the greatest difference between the distributions of estimated Pareto coefficients for population and light emissions is found in LDCs. Similarly to Henderson et al. (2018), but with more recent and accurate satellite imagery, the use of NTL as a proxy for economic activity leads us to conclude that urban agglomeration benefits are more important than congestion costs in developing countries, as reflected by their higher elasticities estimated using nonparametric kernel regression methods.

As pointed out by Ribeiro et al. (2021), Zipf's law and urban scaling are two fundamental paradigms for the study of cities (Batty 2013) that, so far, have been investigated independently. Using data for FUAs, these authors show that urban systems with a more balanced distribution of population tend to have less pronounced increasing returns to scale at the local level and, hence, to display a smaller degree of agglomeration of economic activities. That is, Ribeiro et al. (2021) establish a direct relationship between the Pareto coefficient characterizing the distribution of city sizes in demographic terms β_P with the scaling exponent γ. As a further contribution, we have shown that this latter exponent determines the difference between the national distributions of urban population and light emissions, characterized by β_P and β_L, respectively.

6 Concluding remarks

This paper has conducted a comparison of the distributions of urban population and nighttime lights at the country level. The sample that has been analyzed covers 12,852 consistently-defined urban centers of 100 countries with different levels of development. In line with the results obtained by related studies, but using more recent and accurate satellite imagery to proxy economic activity, we show that aggregate urban light emissions are more unevenly distributed than population. Actually, the null hypothesis that city sizes adjust to Zipf's law can be more easily rejected when they are measured using lights than in demographic terms. Furthermore, there is a greater similarity between the distributions of urban population and lights the higher the level of national income per capita. As a byproduct of our analysis, we provide evidence that casts doubt on the Pareto assumption established to correct the top-coding problem inherent to DMSP images.

We also find a nonlinear and heterogeneous relationship between urban population and aggregate light emissions. In this regard, it is worth noting that the nonparametric estimation framework adopted has led us to obtain higher estimated elasticities of urban lights to population than those previously established in the related literature. Furthermore, the heterogeneity displayed by these elasticities seems to be driven by the level of national income per capita rather than by urban hierarchy. We have finally developed a theoretical framework that establishes the strength of agglomeration economies – reflecting super-linear scaling – as a determinant of the difference between the national distributions of urban population and night lights.

References

Ades, Alberto F., and Edward L. Glaeser. 1995. "Trade and circuses: Explaining urban giants". *Quarterly Journal of Economics* 110 (1): 195–227. doi:10.2307/2118515.

Archila Bustos, Maria F., Ola Hall, Thomas Niedomysl, and Ulf Ernstson. 2020. "A pixel level evaluation of five multitemporal global gridded population datasets: A case study in Sweden, 1990–2015". *Population and Environment* 42 (2): 255–277. doi:10.1007/s11111-020-00360-8.

Arshad, Sidra, Shougeng Hu, and Badar Nadeem Ashraf. 2018. "Zipf's law and city size distribution: A survey of the literature and future research agenda". *Physica A: Statistical Mechanics and its Applications* 492 (15): 75–92. doi:10.1016/j.physa.2017.10.005.

Batty, Michael. 2013. *The new science of cities.* Cambridge, MA: The MIT Press. ISBN: 9780262019521.

Bettencourt, Luís M. A. 2013. "The origins of scaling in cities". *Science* 340 (6139): 1438–1441. doi:10.1126/science.1235823.

Bettencourt, Luís M. A., José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B. West. 2007. "Growth, innovation, scaling, and the pace of life in cities". *Proceedings of the National Academy of Sciences* 104 (17): 7301–7306. doi:10.1073/pnas.0610172104.

Bluhm, Richard, and Melanie Krause. 2022. "Top lights: Bright cities and their contribution to economic development". *Journal of Development Economics* 157:102880. doi:10.1016/j.jdeveco.2022.102880.

Bluhm, Richard, and Gordon C. McCord. 2022. "What can we learn from nighttime lights for small geographies? Measurement errors and heterogeneous elasticities". *Remote Sensing* 14 (5). doi:10.3390/rs14051190.

Brakman, Steven, Harry Garretsen, and Charles van Marrewijk. 2019. *An introduction to geographical and urban economics: A spiky world.* Cambridge, United Kingdom: Cambridge University Press. doi:10.1017/9781108290234.

Ch, Rafael, Diego A. Martin, and Juan F. Vargas. 2021. "Measuring the size and growth of cities using nighttime light". *Journal of Urban Economics* 125:103254. doi:10.1016/j.jue.2020.103254.

Chen, Jiandong, Ming Gao, Shulei Cheng, Wenxuan Hou, Malin Song, Xin Liu, and Yu Liu. 2022. "Global 1 km × 1 km gridded revised real gross domestic product and electricity consumption during 1992–2019 based on calibrated nighttime light data". *Scientific Data* 9 (1): 202. doi:10.1038/s41597-022-01322-5.

Chen, Ruxia, Huimin Yan, Fang Liu, Wenpeng Du, and Yanzhao Yang. 2020. "Multiple global population datasets: Differences and spatial distribution characteristics". *ISPRS International Journal of Geo-Information* 9 (11). doi:10.3390/ijgi9110637.

Chen, Xi, and William D. Nordhaus. 2011. "Using luminosity data as a proxy for economic statistics". *Proceedings of the National Academy of Sciences* 108 (21): 8589–8594. doi:10.1073/pnas.1017031108.

— . 2019. "VIIRS nighttime lights in the estimation of cross-sectional and time-series GDP". *Remote Sensing* 11 (9). doi:10.3390/rs11091057.

Ciccone, Antonio, and Robert E. Hall. 1996. "Productivity and the density of economic activity". *American Economic Review* 86 (1): 54–70. doi:2023-10-19.

Combes, Pierre-Philippe, and Laurent Gobillon. 2015. "The empirics of agglomeration economies". In *Handbook of Regional and Urban Economics*, ed. by Gilles Duranton, J. Vernon Henderson, and William C. Strange, 5:247–348. Handbook of Regional and Urban Economics. Elsevier. doi:10.1016/B978-0-444-59517-1.00005-2.

Davis, James C., and J. Vernon Henderson. 2003. "Evidence on the political economy of the urbanization process". *Journal of Urban Economics* 53 (1): 98–125. doi:10.1016/S0094-1190(02)00504-1.

Decker, Ethan H., Andrew J. Kerkhoff, and Melanie E. Moses. 2007. "Global patterns of city size distributions and their fundamental drivers". *PLOS ONE* 2 (9): 1–6. doi:10.1371/journal.pone.0000934.

Düben, Christian, and Melanie Krause. 2021. "Population, light, and the size distribution of cities". *Journal of Regional Science* 61 (1): 189–211. doi:10.1111/jors.12507.

Duranton, Gilles. 2008. "Viewpoint: From cities to productivity and growth in developing countries". *Canadian Journal of Economics/Revue canadienne d'èconomique* 41 (3): 689–736. doi:10.1111/j.1540-5982.2008.00482.x.

— . 2021. "Classifying locations and delineating space: An introduction". *Journal of Urban Economics* 125:103353. doi:10.1016/j.jue.2021.103353.

Duranton, Gilles, and Diego Puga. 2004. "Micro-foundations of urban agglomeration economies", ed. by J. Vernon Henderson and Jacques-François Thisse, 4:2063–2117. Handbook of Regional and Urban Economics. Amsterdam, The Netherlands: Elsevier. ISBN: 9780444595171. doi:10.1016/S1574-0080(04)80005-1.

Eeckhout, Jan. 2004. "Gibrat's law for (all) cities". *American Economic Review* 94 (5): 1429–1451. doi:10.1257/0002828043052303.

Elvidge, Christopher D., Kimberly Baugh, Mikhail Zhizhin, Feng C. Hsu, and Tilottama Ghosh. 2017. "VIIRS night-time lights". *International Journal of Remote Sensing* 38 (21): 5860–5879. doi:10.1080/01431161.2017.1342050.

Fazio, Giorgio, and Marco Modica. 2015. "Pareto or log-normal? Best fit and truncation in the distribution of all cities". *Journal of Regional Science* 55 (5): 736–756. doi:10.1111/jors.12205.

Florczyk, Aneta, et al. 2019a. *Description of the GHS Urban Centre Database 2015.* JRC115586. Luxembourg: Publications Office of the European Union. doi:10.2760/037310.

Florczyk, Aneta, et al. 2019b. *GHSL Data Package 2019.* EUR 29788 EN. Luxembourg: Publications Office of the European Union. doi:10.2760/290498.

Gabaix, Xavier. 1999. "Zipf's law for cities: An explanation". *Quarterly Journal of Economics* 114 (3): 739–767. doi:10.2307/2586883.

Gabaix, Xavier, and Rustam Ibragimov. 2011. "Rank - 1/2: A simple way to improve the OLS estimation of tail exponents". *Journal of Business & Economic Statistics* 29 (1): 24–39. doi:10.1198/jbes.2009.06157.

Gan, Li, Dong Li, and Shunfeng Song. 2006. "Is the Zipf law spurious in explaining city-size distributions?" *Economics Letters* 92 (2): 256–262. doi:10.1016/j.econlet.2006.03.004.

Gennaioli, Nicola, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2013. "Human capital and regional development". *The Quarterly Journal of Economics* 128 (1): 105–164. doi:10.1093/qje/qjs050.

Gibson, John. 2021. “Better night lights data, for longer”. *Oxford Bulletin of Economics and Statistics* 83 (3): 770–791. doi:10.1111/obes.12417.

Gibson, John, Susan Olivia, and Geua Boe-Gibson. 2020. “Night lights in economics: Sources and uses”. *Journal of Economic Surveys* 34 (5): 955–980. doi:10.1111/joes.12387.

Gibson, John, Susan Olivia, Geua Boe-Gibson, and Chao Li. 2021. “Which night lights data should we use in economics, and where?” *Journal of Development Economics* 149:102602. doi:10.1016/j.jdeveco.2020.102602.

Glaeser, Edward L. 2008. *Cities, agglomeration, and spatial Equilibrium.* The Lindahl Lectures. Oxford, United Kingdom: Oxford University Press. ISBN: 9780191558597.

Glaeser, Edward L., and Joshua D. Gottlieb. 2009. “The wealth of cities: Agglomeration economies and spatial equilibrium in the United States”. *Journal of Economic Literature* 47 (4): 983–1028. doi:10.1257/jel.47.4.983.

Grover, Arti, Somik Lall, and Jonathan Timmis. 2023. “Agglomeration economies in developing countries: A meta-analysis”. *Regional Science and Urban Economics* 101:103901. doi:10.1016/j.regsciurbeco.2023.103901.

Henderson, J. Vernon, Tim Squires, Adam Storeygard, and David Weil. 2018. “The global distribution of economic activity: Nature, history, and the role of trade”. *The Quarterly Journal of Economics* 133 (1): 357–406. doi:10.1093/qje/qjx030.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. 2012. “Measuring economic growth from outer space”. *American Economic Review* 102 (2): 994–1028. doi:10.1257/aer.102.2.994.

Hsiao, Cheng, Qi Li, and Jeffrey S. Racine. 2007. “A consistent model specification test with mixed discrete and continuous data”. *Journal of Econometrics* 140 (2): 802–826. doi:10.1016/j.jeconom.2006.07.015.

Hu, Yingyao, and Jiaxiong Yao. 2022. “Illuminating economic growth”. *Journal of Econometrics* 228 (2): 359–378. doi:10.1016/j.jeconom.2021.05.007.

Ioannides, Yannis, and Spyros Skouras. 2013. "US city size distribution: Robustly Pareto, but only in the tail". *Journal of Urban Economics* 73 (1): 18–29. doi:10.1016/j.jue.2012.06.005.

Jacobs, Jane. 1969. *The economy of cities.* A Vintage Book, V-584. New York, NY: Random House. ISBN: 9780394422961.

Jedwab, Remi, Prakash Loungani, and Anthony Yezer. 2021. "Comparing cities in developed and developing countries: Population, land area, building height and crowding". *Regional Science and Urban Economics* 86:103609. doi:10.1016/j.regsciurbeco.2020.103609.

Jiang, Bin, Junjun Yin, and Qingling Liu. 2015. "Zipf's law for all the natural cities around the world". *International Journal of Geographical Information Science* 29 (3): 498–522. doi:10.1080/13658816.2014.988715.

Keola, Souknilanh, Magnus Andersson, and Ola Hall. 2015. "Monitoring economic development from space: Using nighttime light and land cover data to measure economic growth". *World Development* 66:322–334. doi:10.1016/j.worlddev.2014.08.017.

Krugman, Paul. 1991. "Increasing returns and economic geography". *Journal of Political Economy* 99 (3): 483–499. doi:10.1086/261763.

— . 1993. "First nature, second nature, and metropolitan location". *Journal of Regional Science* 33 (2): 129–144. doi:10.1111/j.1467-9787.1993.tb00217.x.

Kummu, Matti, Maija Taka, and Joseph H. A. Guillaume. 2018. "Gridded global datasets for gross domestic product and human development index over 1990–2015". *Scientific data* 5 (1): 1–15.

Lall, Somik, Mathilde Lebrand, Hogeun Park, Daniel Sturm, and Anthony Venables. 2021. *Pancakes to pyramids.* Tech. rep. Washington, DC: World Bank.

Leyk, Stefan, et al. 2019. "The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use". *Earth System Science Data* 11 (3): 1385–1409. doi:10.5194/essd-11-1385-2019.

Li, Qi, and Jeffrey S. Racine. 2007. *Nonparametric econometrics: Theory and practice.* Princeton, NJ: Princeton University Press. ISBN: 9780691121611.

Lobo, José, Luís M. A. Bettencourt, Deborah Strumsky, and Geoffrey B. West. 2013. "Urban scaling and the production function for cities". *PLOS ONE* 8 (3): 1–10. doi:10.1371/journal.pone.0058407.

Phan, Diep H. 2023. "Lights and GDP relationship: What does the computer tell us?" *Empirical Economics* 0:1–38. doi:10.1007/s00181-023-02377-y.

Puente-Ajovín, Miguel, Arturo Ramos, and Fernando Sanz-Gracia. 2020. "Is there a universal parametric city size distribution? Empirical evidence for 70 countries". *Annals of Regional Science* 65 (3): 727–741. doi:10.1007/s00168-020-01001-6.

Puente-Ajovín, Miguel, Arturo Ramos, Fernando Sanz-Gracia, and Daniel Arribas-Bel. 2020. "How sensitive is city size distribution to the definition of city? The case of Spain". *Economics Letters* 197:109643. doi:10.1016/j.econlet.2020.109643.

Puente-Ajovín, Miguel, Marcos Sanso-Navarro, and María Vera-Cabello. 2022. "The distribution of urban population and economic activity in the European Union and the United States". *Letters in Spatial and Resource Sciences* 15:517–522. doi:10.1007/s12076-022-00309-5.

Racine, Jeffrey S., and Qi Li. 2004. "Nonparametric estimation of regression functions with both categorical and continuous data". *Journal of Econometrics* 119 (1): 99–130. doi:10.1016/S0304-4076(03)00157-X.

Ribeiro, Haroldo V., Milena Oehlers, Ana I. Moreno-Monroy, Jürgen P. Kropp, and Diego Rybski. 2021. "Association between population distribution and urban GDP scaling". *PLOS ONE* 16 (1): 1–15. doi:10.1371/journal.pone.0245771.

Rosen, Kenneth T., and Mitchel Resnick. 1980. "The size distribution of cities: An examination of the Pareto law and primacy". *Journal of Urban Economics* 8 (2): 165–186. doi:10.1016/0094-1190(80)90043-1.

Small, Christopher, Christopher D. Elvidge, Deborah Balk, and Mark Montgomery. 2011. "Spatial scaling of stable night lights". *Remote Sensing of Environment* 115 (2): 269–280. doi:10.1016/j.rse.2010.08.021.

Soo, Kwok Tong. 2005. "Zipf's Law for cities: A cross-country investigation". *Regional Science and Urban Economics* 35 (3): 239–263. doi:10.1016/j.regsciurbeco.2004.04.004.

Venables, Anthony J. 2005. "Spatial disparities in developing countries: Cities, regions, and international trade". *Journal of Economic Geography* 5 (1): 3–21. doi:10.2307/26160603.

Wang, Tingting, and Fubao Sun. 2022. "Global gridded GDP data set consistent with the shared socioeconomic pathways". *Scientific Data* 9 (1): 221. doi:10.1038/s41597-022-01300-x.

World Bank. 2008. *World development report 2009: Reshaping economic geography*. Washington, DC: The World Bank. ISBN: 9780821376089. doi:10.1596/978-0-8213-7607-2.

Tables and figures

Table 1: Descriptive statistics of city sizes by country income group.

	All countries	High income	Upper-middle	Lower-middle	Low income
Countries	100	22	29	27	22
Urban centers	12,852	1,298	3,795	6,213	1,546
Mean					
GHSPOP	268,247	410,864	312,484.40	237,467.40	163,612.50
VIIRS	6,202.14	29,660.31	8,419.74	1,420.58	279.35
Median					
GHSPOP	99,755.16	108,721.70	106,719.20	97,808.61	90,814.05
VIIRS	460.99	8,257.89	2,060.96	162.58	7.93
Minimum					
GHSPOP	50,002.46	50,056.39	50,007.17	50,012.63	50,002.46
VIIRS	0	190.17	0	0	0
Maximum					
GHSPOP	4.06E+07	3.30E+07	4.06E+07	3.63E+07	5.62E+06
VIIRS	1.20E+06	1.20E+06	1.01E+06	4.01E+05	32,146.45

Note: GHSPOP is measured in number of persons, and VIIRS refers to aggregate nano Watts per square centimeter per steredian. Countries grouped according to the World Bank classification for the year 2015, see Table A1 in Appendix A for further details.

Table 2: Kolmogorov-Smirnov test. Percentage of rejections at different significance levels.

Panel A. H_0: Exact Zipf's law						
	GHSPOP			VIIRS		
	1%	5%	10%	1%	5%	10%
All countries	17.00	30.00	37.00	85.00	88.00	92.00
High income	0.00	9.09	18.18	63.64	77.27	81.82
Upper-middle	11.11	22.22	25.93	81.48	81.48	88.89
Lower-middle	20.69	44.83	55.17	96.55	96.55	100.00
Low income	36.36	40.91	45.45	95.45	95.45	95.45

Panel B. H_0: Pareto distribution						
	GHSPOP			VIIRS		
	1%	5%	10%	1%	5%	10%
All countries	9.00	17.00	24.00	75.00	80.00	84.00
High income	0.00	0.00	4.54	50.00	63.64	68.18
Upper-middle	7.41	18.52	25.93	77.78	77.78	85.19
Lower-middle	13.79	24.14	34.48	86.21	89.66	89.66
Low income	13.64	22.73	27.27	81.82	86.36	90.91

Table 3: Robustness check: Kolmogorov-Smirnov test. Percentage of rejections at different significance levels.

Panel A. H_0: Exact Zipf's law	DMSP			DMSP_BK			GPW			WorldPop			LandScan		
	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
All countries	78.00	85.00	88.00	81.00	87.00	89.00	86.00	90.00	91.00	75.00	79.00	83.00	86.00	88.00	90.00
High income	50.00	68.18	68.18	63.64	72.73	72.73	54.54	63.64	68.18	27.27	36.36	45.45	45.45	54.54	59.09
Upper-middle	85.18	85.18	88.89	85.18	88.89	88.89	88.89	96.30	96.30	70.37	77.78	85.19	92.59	92.59	92.59
Lower-middle	86.21	93.10	96.55	86.21	93.10	96.55	96.55	96.55	96.55	96.55	96.55	96.55	96.55	96.55	100.00
Low income	86.36	90.91	95.45	86.36	90.91	95.45	100.00	100.00	100.00	100.00	100.00	100.00	95.45	100.00	100.00

Panel B. H_0: Pareto distribution	DMSP			DMSP_BK			GPW			WorldPop			LandScan		
	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
All countries	72.00	76.00	81.00	72.00	75.00	79.00	71.00	74.00	78.00	66.00	74.00	77.00	80.00	84.00	88.00
High income	45.45	54.54	63.64	45.45	54.54	54.54	40.91	40.91	59.09	22.73	31.82	36.36	54.54	54.54	63.64
Upper-middle	81.48	85.19	88.89	81.48	81.48	88.89	70.37	77.78	77.78	62.96	66.67	70.37	77.78	88.89	92.59
Lower-middle	82.76	82.76	86.21	82.76	82.76	86.21	82.76	82.76	82.76	86.21	93.10	96.55	93.10	96.55	96.55
Low income	72.73	77.27	81.82	72.73	77.27	81.82	86.36	90.91	90.91	86.36	100.00	100.00	90.91	90.91	95.45

Table 4: Robustnest check: Relationship between aggregate urban light emissions and gridded GDP and population. OLS estimation.

Panel A. Excluding country fixed effects										
	GDP					GHSPOP				
	All countries	High income	Upper-middle	Lower-middle	Low income	All countries	High income	Upper-middle	Lower-middle	Low income
VIIRS	0.37***	0.80***	0.45***	0.28***	0.15***	0.09***	0.69***	0.34***	0.10***	0.05***
	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.00)	(0.02)	(0.02)	(0.01)	(0.00)
Intercept	18.05***	15.01***	17.73***	18.30***	17.44***	11.26***	5.56***	9.26***	11.22***	11.56***
	(0.04)	(0.16)	(0.18)	(0.06)	(0.05)	(0.01)	(0.15)	(0.16)	(0.02)	(0.01)
Observations	12,657	1,298	3,790	6,104	1,465	12,657	1,298	3,790	6,104	1,465
R^2	0.52	0.68	0.41	0.33	0.16	0.16	0.70	0.42	0.17	0.14
Panel B. Including country fixed effects										
	GDP					GHSPOP				
	All countries	High income	Upper-middle	Lower-middle	Low income	All countries	High income	Upper-middle	Lower-middle	Low income
VIIRS	0.28***	0.90***	0.53***	0.29***	0.16***	0.15***	0.85***	0.43***	0.14***	0.06***
	(0.01)	(0.02)	(0.03)	(0.02)	(0.01)	(0.00)	(0.01)	(0.03)	(0.01)	(0.00)
Intercept	16.653	13.81***	16.42	18.35***	16.73***	11.47***	4.41***	8.10***	11.16***	11.52***
	(0.24)	(0.21)	(0.42)	(0.26)	(0.25)	(0.11)	(0.16)	(0.32)	(0.26)	(0.08)
Observations	12,657	1,298	3,790	6,104	1,465	12,657	1,298	3,790	6,104	1,465
R^2	0.64	0.82	0.52	0.39	0.31	0.25	0.86	0.53	0.22	0.23

Note: The dependent variables are aggregate urban gridded GDP (Kummu, Taka, and Guillaume 2018) and population. All variables are expressed in (natural) logarithms. Robust standard errors in parentheses. ***$p<0.01$.

Table 5: VIIRS-GHSPOP elasticities. OLS estimation.

	(1)	(2)	(3)
GHSPOP (in logs)	1.50***	1.52***	1.54***
	(0.08)	(0.09)	(0.10)
Primacy		12.64***	
		(4.19)	
GHSPOP*Primacy		-0.85***	
		(0.27)	
Top10			5.71***
			(1.25)
GHSPOP*Top10			-0.41***
			(0.09)
Intercept	-17.09***	-17.29***	-17.53***
	(0.93)	(1.01)	(1.11)
R^2	0.63	0.63	0.63

Note: The dependent variable is aggregate VIIRS nighttime lights (in logs). The sample is made up of 12,852 observations. All estimations include country fixed effects. Clustered standard errors are reported in parentheses.*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

Table 6: VIIRS-GHSPOP elasticities. Least-squares cross-validation bandwidths and diagnostic test statistics for nonparametric kernel regressions.

	Upper bound	Local-constant estimation (1)	(2)	(3)	(4)
GHSPOP (in logs)	1.74	0.24	0.23	0.24	0.24
Primacy	1.00		0.04		
GHSPOP*Primacy	2.66		6.21E+06*		
Top10	1.00			0.50	
GHSPOP*Top10	7.06			2.16E+06*	
Upper-middle	1.00				0.26
GHSPOP*Upper-middle	10.48				1.79E+06*
Lower-middle	1.00				0.03
GHSPOP*Lower-middle	11.76				1.57E+05*
Low income	1.00				0.43
GHSPOP*Low income	7.54				0.16
	Upper bound	**Local-linear estimation (1)**	**(2)**	**(3)**	**(4)**
GHSPOP (in logs)	1.74	1.16	1.29	1.48E+06**	1.21
Primacy	1.00		0.50		
GHSPOP*Primacy	2.66		1.71E+06**		
Top10	1.00			0.50	
GHSPOP*Top10	7.06			0.80	
Upper-middle	1.00				0.50
GHSPOP*Upper-middle	10.48				1.03E+06**
Lower-middle	1.00				0.50
GHSPOP*Lower-middle	11.76				1.31E+06**
Low income	1.00				0.40
GHSPOP*Low income	7.54				5.29E+05**
R^2		0.65	0.65	0.43	0.65
HLR1		8.08	8.25	8.22	4.76
		(0.00)	(0.00)	(0.00)	(0.00)
HLR2		11.57	11.57	10.00	6.85
		(0.00)	(0.00)	(0.00)	(0.00)

Note: The dependent variable is aggregate VIIRS nighttime lights (in logs). The sample is made up of 12,852 observations. All estimations include country fixed effects. * denotes that the variable is smoothed out of the regression, and ** indicates that the regressor enters linearly. The Hsiao, Li, and Racine (2007) test statistic has been calculated for a standard OLS model (HLR1) and a quadratic specification (HLR2). P-values are reported in parentheses.

Table 7: VIIRS-GHSPOP elasticities. Local-linear kernel regression.

	Mean	Q1	Q2	Q3
All countries	1.76	1.25	1.40	1.52
	(0.37)	(0.06)	(0.05)	(0.03)
Primary cities	1.50	0.98	1.18	1.56
	(0.03)	(0.06)	(0.07)	(0.22)
10 largest cities	1.77	1.07	1.27	1.85
	(0.40)	(0.06)	(0.08)	(0.42)
High income	1.07	1.00	1.07	1.11
	(0.44)	(0.06)	(0.06)	(0.04)
Upper-middle	1.31	1.11	1.36	1.42
	(0.15)	(0.11)	(0.10)	(0.04)
Lower-middle	1.69	1.38	1.41	1.53
	(0.24)	(0.06)	(0.04)	(0.28)
Low income	3.73	3.33	3.43	4.59
	(0.84)	(0.33)	(0.39)	(0.46)

Note: Reported partial effects are the estimated derivatives from a local-linear kernel regression using GHSPOP urban population (in logs) and country fixed effects as covariates, and the bandwidths displayed in Table 6. Bootstrap standard errors (399 replications) in parentheses.

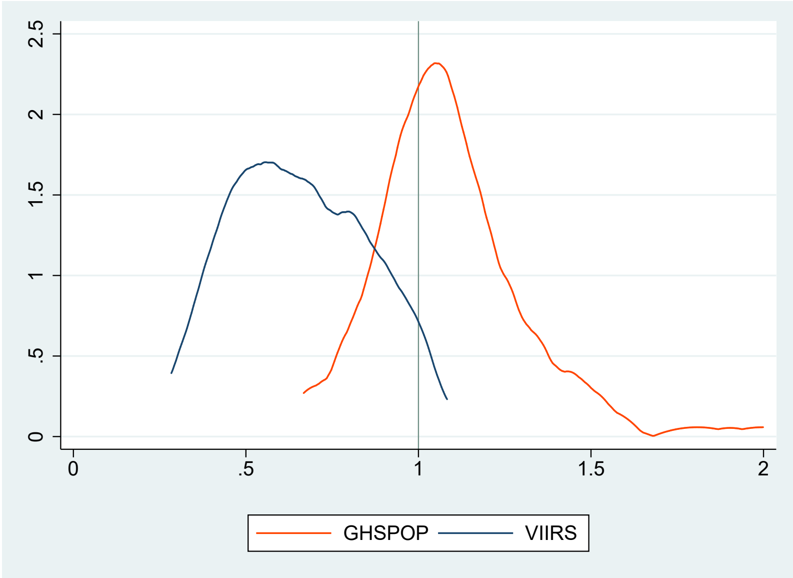

Figure 1: Kernel densities of estimated Pareto coefficients from a rank-size OLS regression at the country level.

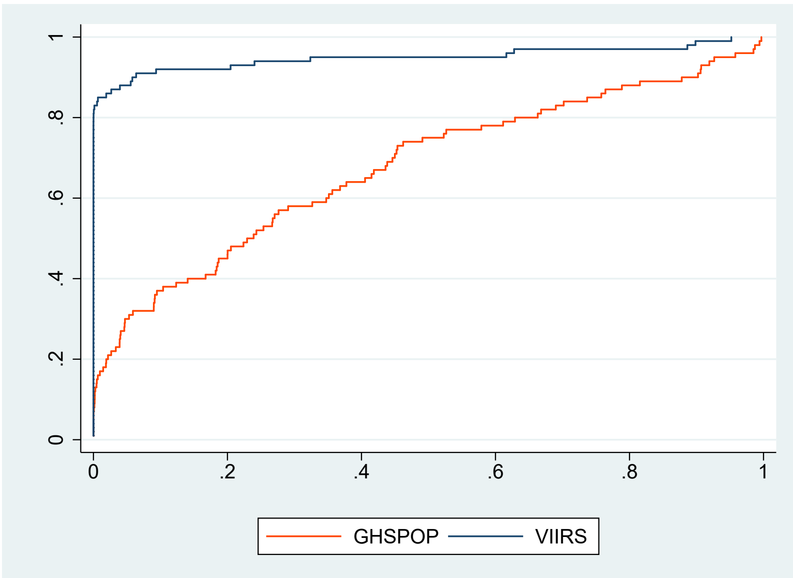

Figure 2: Cumulative distribution function of Kolmogorov-Smirnov test p-values using exact Zipf's law as a reference.

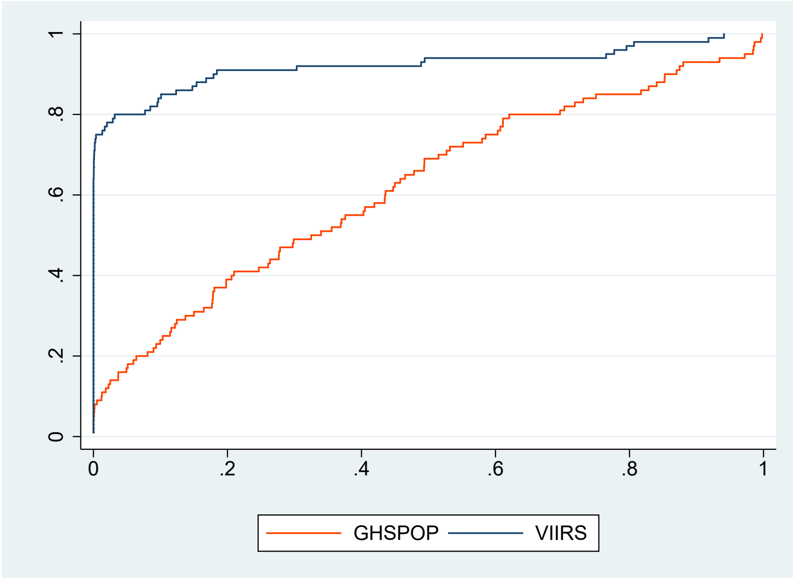

Figure 3: Cumulative distribution function of Kolmogorov-Smirnov test p-values using a Pareto distribution as a reference.

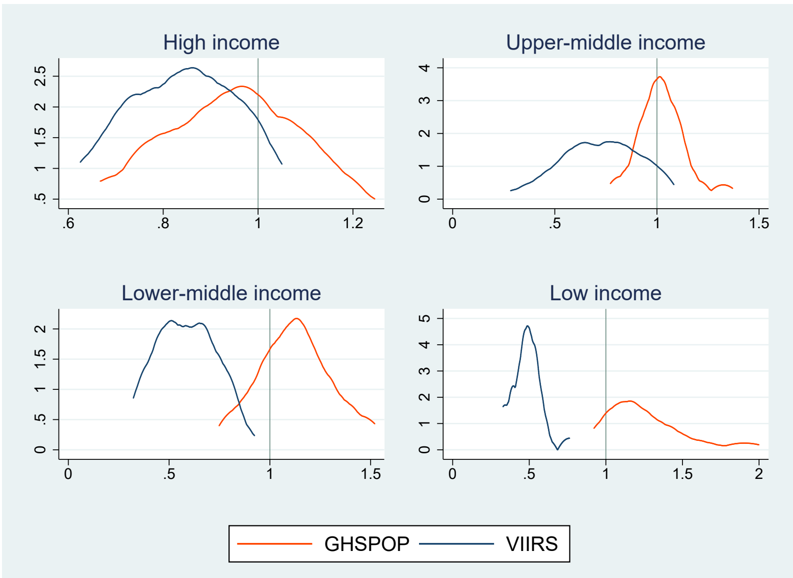

Figure 4: Kernel densities of estimated Pareto coefficients from a rank-size OLS regression at country level by income group, World Bank classification 2015.

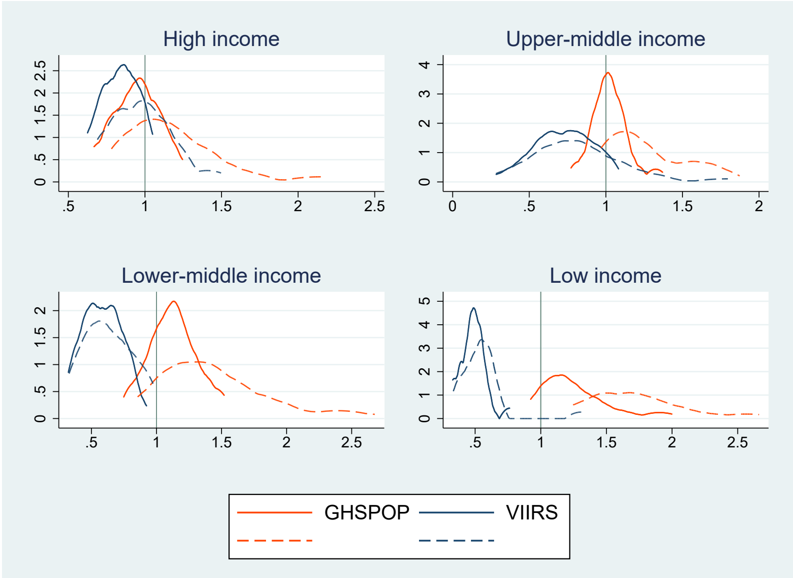

Figure 5: Robustness check: Kernel densities of estimated Pareto coefficients from rank-size regressions at the country level including (solid) and excluding (dashed) primary cities.

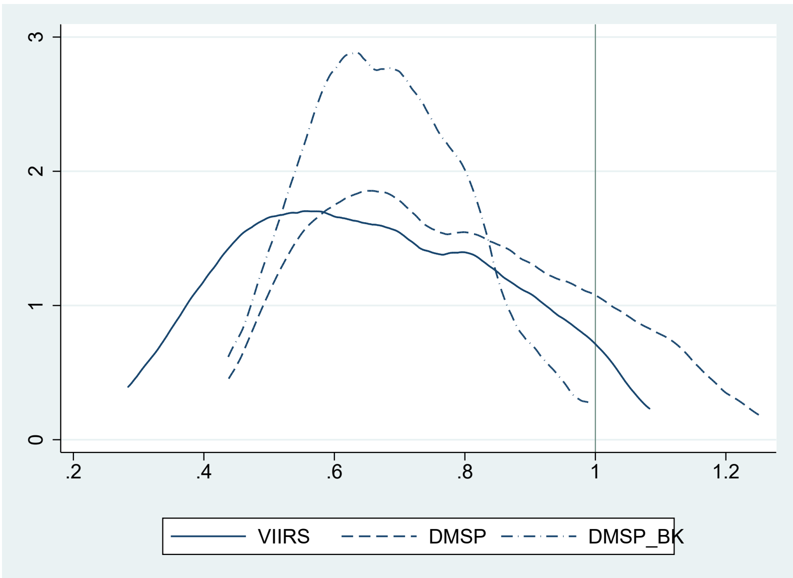

Figure 6: Robustness check: Kernel densities of estimated Pareto coefficients from rank-size regressions at the country level using alternative nighttime lights data.

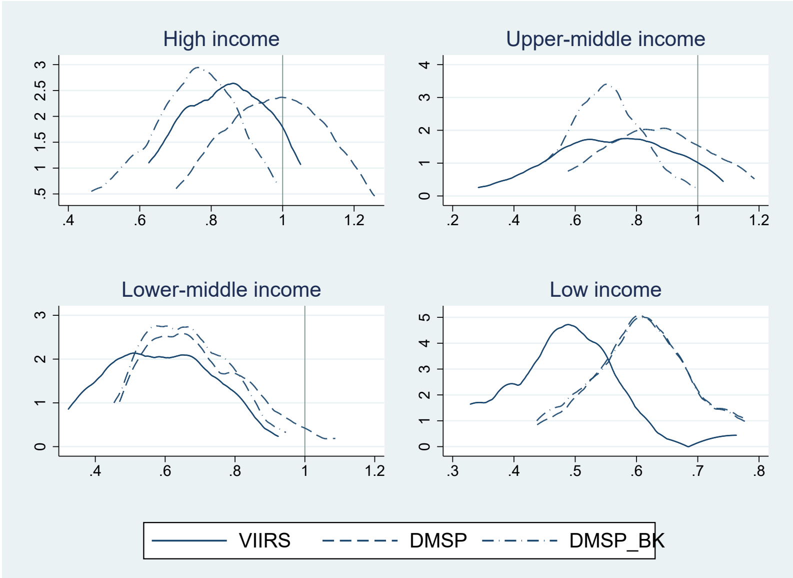

Figure 7: Robustness check: Kernel densities of estimated Pareto coefficients from country rank-size regressions by income group using alternative nighttime lights data.

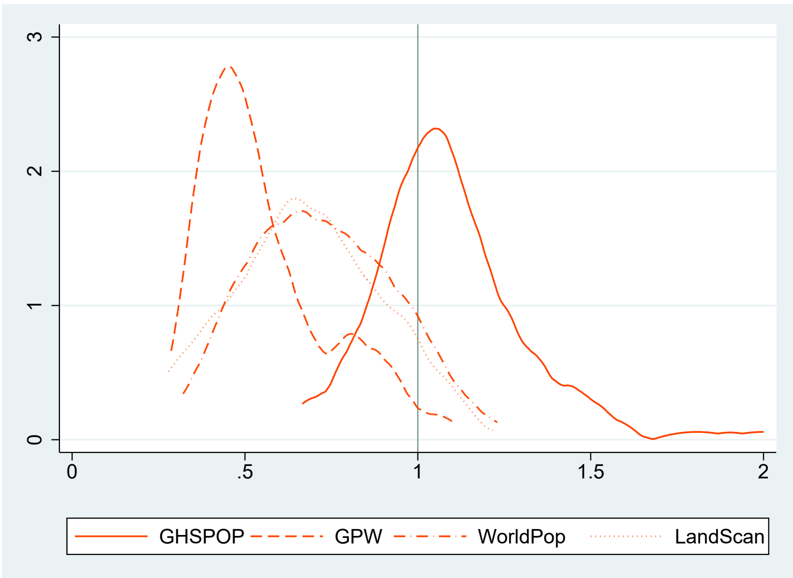

Figure 8: Robustness check: Kernel densities of estimated Pareto coefficients from rank-size regressions at the country level using alternative gridded population data.

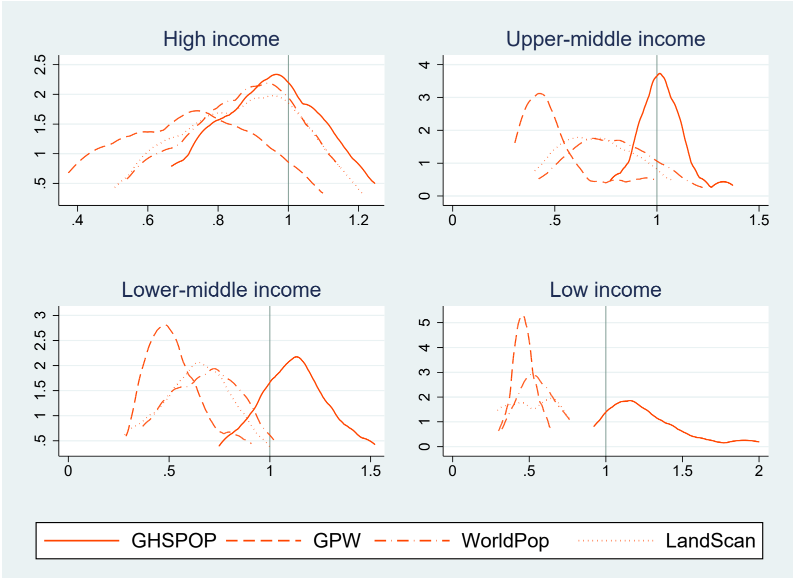

Figure 9: Robustness check: Kernel densities of estimated Pareto coefficients from country rank-size regressions by income group using alternative gridded population data.

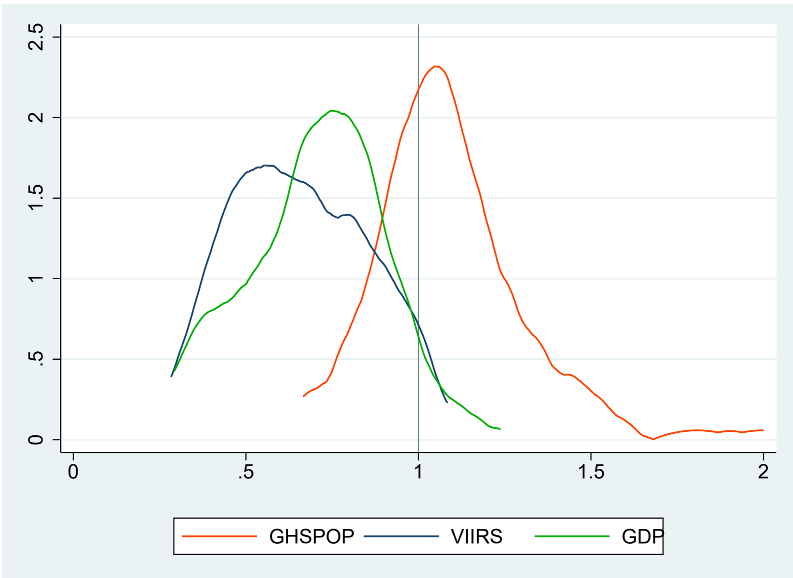

Figure 10: Robustness check: Kernel densities of estimated Pareto coefficients from country rank-size regressions. Comparison with gridded GDP data.

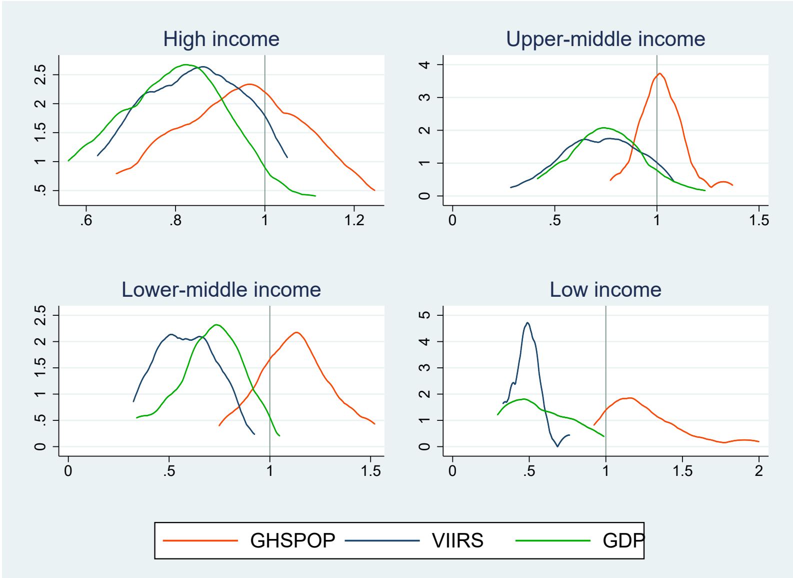

Figure 11: Robustness check: Kernel densities of estimated Pareto coefficients from country rank-size regressions by income group. Comparison with gridded GDP data.

Appendix A

Table A1: Countries included in the sample, grouped according to the World Bank classification for the year 2015.

High income	Upper-middle income	Lower-middle income	Low income
Australia [27] (Oceania)	Algeria [96] (Africa)	Bangladesh [307] (Asia)	Afghanistan [74] (Asia)
Belgium [12] (Europe)	Angola [58] (Africa)	Bolivia [13] (South America)	Benin [25] (Africa)
Canada [49] (North America)	Argentina [72] (South America)	Cambodia [11] (Asia)	Burkina Faso [44] (Africa)
Chile [33] (South America)	Azerbaijan [17] (Asia)	Cameroon [54] (Africa)	Burundi [43] (Africa)
Czechia [12] (Europe)	Belarus [15] (Europe)	Côte d'Ivoire [35] (Africa)	Chad [51] (Africa)
France [77] (Europe)	Brazil [352] (South America)	Egypt [190] (Africa)	Congo [159] (Africa)
Germany [89] (Europe)	China [1,851] (Asia)	Ghana [59] (Africa)	Ethiopia [557] (Africa)
Greece [10] (Europe)	Colombia [92] (South America)	Guatemala [48] (North America)	Guinea [18] (Africa)
Hungary [11] (Europe)	Cuba [19] (North America)	Honduras [13] (North America)	Haiti [23] (North America)
Italy [91] (Europe)	Dominican Republic [16] (North America)	India [3,252] (Asia)	Korea (Democratic People's Republic of) [91] (Asia)
Japan [109] (Asia)	Ecuador [31] (South America)	Indonesia [393] (Asia)	Madagascar [24] (Africa)
Korea (Republic of) [39] (Asia)	Iran [182] (Asia)	Kenya [45] (Africa)	Mali [16] (Africa)
Netherlands [37] (Europe)	Iraq [81] (Asia)	Morocco [63] (Africa)	Mozambique [90] (Africa)
Oman [11] (Asia)	Kazakhstan [27] (Asia)	Myanmar [126] (Asia)	Nepal [28] (Asia)
Poland [48] (Europe)	Libya [15] (Africa)	Nicaragua [18] (North America)	Niger [44] (Africa)
Saudi Arabia [53] (Asia)	Malaysia [38] (Asia)	Nigeria [484] (Africa)	Senegal [34] (Africa)
Spain [73] (Europe)	Mexico [168] (North America)	Pakistan [302] (Asia)	Somalia [36] (Africa)
Sweden [12] (Europe)	Paraguay [10] (South America)	Papua New Guinea [47] (Oceania)	South Sudan [55] (Africa)
Switzerland [17] (Europe)	Peru [51] (South America)	Philippines [93] (Asia)	Tanzania [46] (Africa)
Taiwan [21] (Asia)	Romania [30] (Europe)	Sri Lanka [22] (Asia)	Togo [21] (Africa)
United Kingdom [138] (Europe)	Russian Federation [209] (Europe)	Sudan [124] (Africa)	Uganda [34] (Africa)
United States of America [329] (North America)	Serbia [14] (Europe)	Syrian Arab Republic [26] (Asia)	Zimbabwe [33] (Africa)
	South Africa [77] (Africa)	Tajikistan [16] (Asia)	
	Thailand [48] (Asia)	Tunisia [26] (Africa)	
	Turkey [136] (Asia)	Ukraine [78] (Europe)	
	Turkmenistan [11] (Asia)	Uzbekistan [56] (Asia)	
	Venezuela [79] (South America)	Viet Nam [163] (Asia)	
		Yemen [100] (Asia)	
		Zambia [49] (Africa)	

Note: The number of urban centers included in national samples are reported in brackets.

Table A2: Robustness check: Descriptive statistics of city sizes by country income group.

	All countries	High income	Upper-middle	Lower-middle	Low income
Countries	100	22	29	27	22
Urban centers	12,852	1,298	3,795	6,213	1,546
Mean					
DMSP	3,458.43	14,123.17	4,482.78	1,371.28	375.15
DMSP_BK	8,522.06	44,610.24	10,346.06	13,909.08	400.51
GPW	142,001.90	335,741.70	209,846.50	85,462.46	40,019.59
WorldPop	191,171.60	390,319	271,547.70	133,642.30	57,865.73
LandScan	207,950.10	413,222.70	274,650.80	156.315.80	79,380.55
GDP	3.98E+09	1.73E+10	5.31E+09	1.34E+09	1.83E+08
Median					
DMSP	689	4,665	1,713	258	15
DMSP_BK	694	8,099.96	1,901.52	258	15
GPW	12,329.90	77,282.58	39,938.60	6,477.77	806.80
WorldPop	43,980.82	98,391.75	80,787.69	25,587.56	6,411.31
LandScan	53,406	110,747	79,606	37,847	14,335.50
GDP	6.30E+08	4.08E+09	1.34E+09	3.73E+08	5.06E+07
Minimum					
DMSP	0	194	0	0	0
DMSP_BK	0	194	0	0	0
GPW	0.58	18.543	5.029	1.084	0.58
WorldPop	3.155	817.94	53.325	3.155	3,24
LandScan	0	1,144	9	0	2.90
GDP	0	9.31E+06	0	0	0
Maximum					
DMSP	509,507	509,507	505,237	269,129	29,844
DMSP_BK	2.37E+06	2.37E+06	1.55E+06	5.69E+05	35,082.50
GPW	3.71E+07	3.15E+07	3.71E+07	2.69E+07	5.27E+06
WorldPop	3.98E+07	3.34E+07	3.98E+07	3.26E+07	6.06E+06
LandScan	3.49E+07	3.24E+07	3.49E+07	2.83E+07	8.18E+06
GDP	1.43E+12	1.43E+12	7.70E+11	5.03E+11	2.78E+10

Note: DMSP light intensities are recorded at the pixel level as integerized digital numbers (DN) ranging from 0 to 63 in the original (truncated) version, and from 0 to 2,000 in the corrected data set created by Bluhm and Krause (2022). City sizes have been calculated by aggregating the DN of the pixels within the spatial extent of urban centers. WorldPop, GPW, and LandScan refer to the number of persons. GDP values are expressed in 2011 (International) United States Dollars (Kummu, Taka, and Guillaume 2018).

Table A3: DMSP-GHSPOP elasticities. OLS estimation.

	DMSP			DMSP_BK		
	(1)	(2)	(3)	(5)	(5)	(6)
GHSPOP (in logs)	1.62***	1.66***	1.71***	1.75***	1.79***	1.82***
	(0.15)	(0.16)	(0.20)	(0.13)	(0.15)	(0.18)
Primacy		17.46***			17.34***	
		(5.53)			(5.36)	
GHSPOP*Primacy		-1.21***			-1.19***	
		(0.36)			(0.35)	
Top10			10.85***			9.82***
			(2.24)			(2.08)
GHSPOP*Top10			-0.80***			-0.72***
			(0.18)			(0.16)
Intercept	-19.22***	-19.68***	-20.36***	-20.78***	-21.19***	1.82***
	(1.70)	(1.88)	(2.32)	(1.50)	(1.67)	(0.18)
R^2	0.54	0.54	0.54	0.56	0.56	0.56

Note: The dependent variable is aggregate DMSP nighttime lights (in logs). The sample is made up of 12,852 observations. All estimations include country fixed effects. Clustered standard errors are reported in parentheses.*p $<$ 0.10, **p $<$ 0.05, ***p $<$ 0.01.

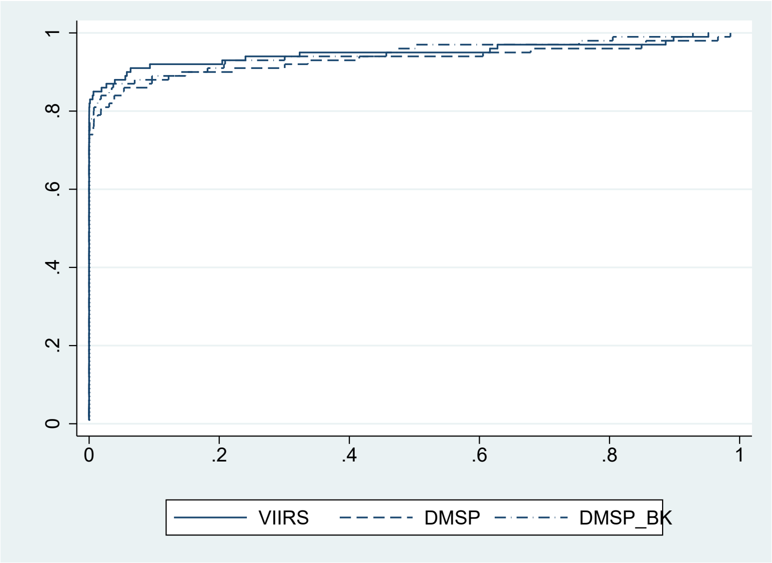

Figure A1: Robustness check: Cumulative distribution function of Kolmogorov-Smirnov test p-values using exact Zipf's law as a reference and alternative nighttime lights data.

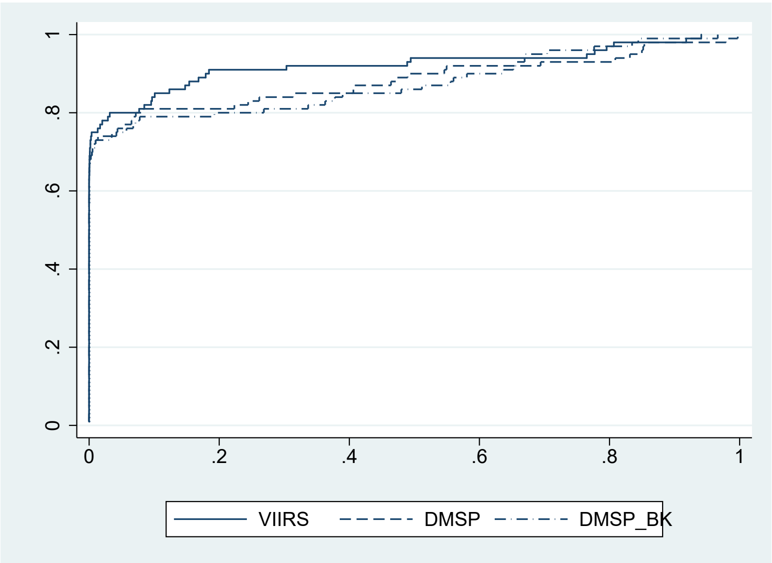

Figure A2: Robustness check: Cumulative distribution function of Kolmogorov-Smirnov test p-values using a Pareto distribution as a reference and alternative nighttime lights data.

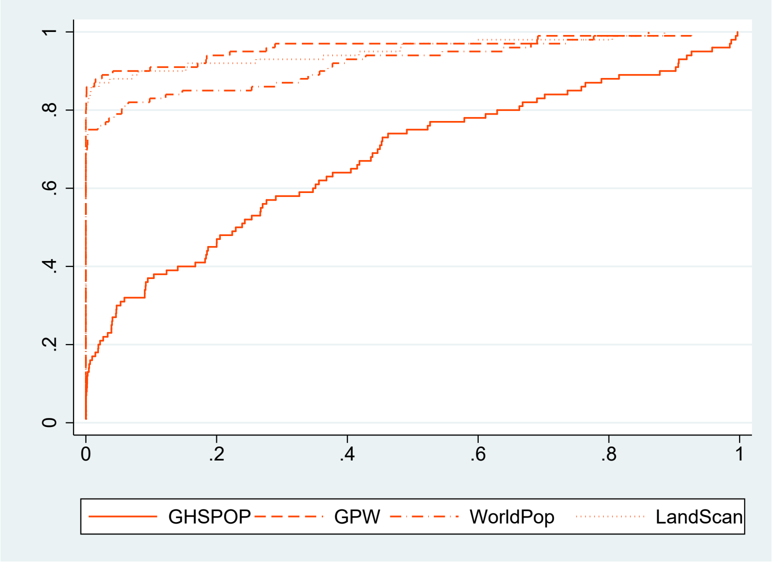

Figure A3: Robustness check: Cumulative distribution function of Kolmogorov-Smirnov test p-values using exact Zipf's law as a reference and alternative gridded population data.

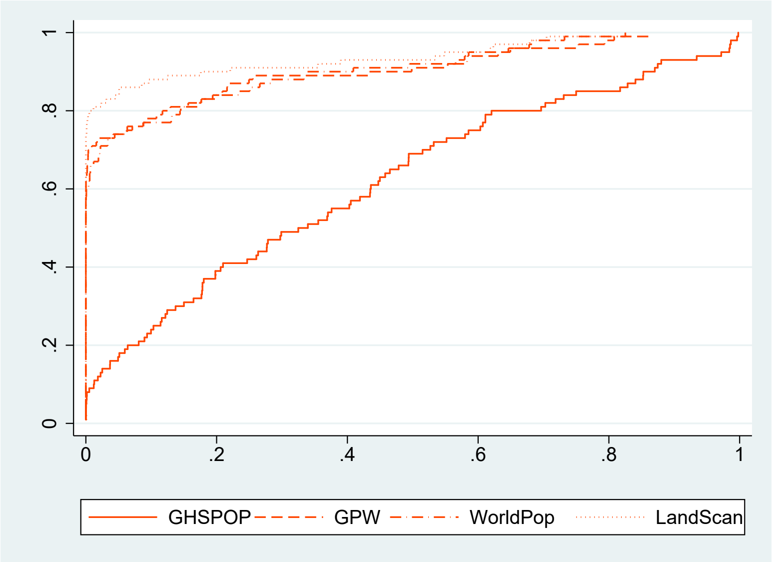

Figure A4: Robustness check: Cumulative distribution function of Kolmogorov-Smirnov test p-values using a Pareto distribution as a reference and alternative gridded population data.

Appendix B

The distributions of estimated Pareto coefficients using the GPW data set represented in Figures 8 and 9 can be related to the uniform areal weighting approach that this database implements to allocate population into grid cells. This method implies that if an administrative unit has a total population of P and contains M pixels, each of them is assigned a value of P/M. This leads to an unrealistic population distribution that does not consider neither the existence of heterogeneous levels of urbanization over space, nor the presence of natural barriers such as mountains or rivers. By assuming an even population distribution, GPW does not accurately represent the true spatial diversity of population density, bringing about a distorted understanding of where people reside.

Another issue of GPW data is that it relies on different administrative levels across countries, making national estimated Pareto coefficients to be not completely comparable. That is to say, when analyzing the cross-country concentration of population, discrepancies may arise not only from its actual distribution, but also from the different national administrative divisions considered. Furthermore, even when two countries adopt the same administrative level, its definition can be distinct in each of them[23]. In those countries where the population is more evenly distributed over space and a more disaggregated geographical level is considered, a larger percentage of the population will reside outside the boundaries of GHSL urban areas. Consequently, their population will be significantly under-represented in certain regions, rendering the calculation of the Pareto coefficient unreliable.

In order to illustrate this under-representation, let us think about an administrative unit with a total population of 90 and a dimension of a 3x3 grid, where the population lives in the center. While the GPW considers the whole area of the administrative unit to distribute the population at the pixel level, GHSL only takes into account the area where people reside. For this reason, and as exemplified in Figure B1, population is under-represented when calculated using GPW pixels and GHSL areas.

[23] As an example, Brazil, Canada, and France adopt Level 2. It corresponds to municipalities in Brazil, census divisions in Canada, and departments in France.

GPW		
10	10	10
10	10	10
10	10	10

GHSPOP		
	90	

Figure B1: Example: Distribution of population in GPW and GHS-POP data sets.

Therefore, it can be stated that the city sizes calculated combining the information provided by GPW and the urban centers defined by GHSL ($Popul_{GPW}$) will be a share of actual population ($Popul_{GHSPOP}$), determined by the ratio between the area of the urban center and that of the administrative unit, expected to range between zero and one:

$$Popul_{GPW} = \frac{Area_{GHSL}}{Area_{GPW}} Popul_{GHSPOP}. \tag{B.1}$$

In other words, the population included in the areas of the GPW raster can be defined as a function of that in the GHS-POP data set. It can also be assumed that there is a direct link between the extent of an administrative unit and the number of persons that reside in its urban center. Therefore, a higher population will tend to increase its density ($d \geq 0$):

$$\frac{Area_{GHSL}}{Area_{GPW}} = C \left[Popul_{GHSPOP}\right]^{d}. \tag{B.2}$$

Combining previous expressions, defining $c = \log(C)$ and $D = d + 1$, and taking natural logarithms, it can be written that:

$$\log(Popul_{GPW}) = c + D \log(Popul_{GHSPOP}). \tag{B.3}$$

Taking into account expression (1), and disregarding the error term, the relationship between the Pareto coefficient estimated using the gridded population provided by the GHSL (β_{GHSPOP}) and that obtained from GPW data (β_{GPW}) can be derived as follows:

$$
\begin{aligned}
\log(Rank - 0.5) &= \alpha_{GHSPOP} - \beta_{GHSPOP} \log(Popul_{GHSPOP}) = \\
&= \alpha_{GHSPOP} - \beta_{GHSPOP} \left[\frac{\log(Popul_{GPW}) - c}{D} \right] = \\
&= (\alpha_{GHSPOP} + \beta_{GHSPOP} \frac{c}{D}) - \frac{\beta_{GHSPOP}}{D} \log(Popul_{GPW}) = \\
&= \alpha_{GPW} - \beta_{GPW} \log(Popul_{GPW});
\end{aligned} \tag{B.4}
$$

The disparity between β_{GPW} and β_{GHSPOP} is determined by the value of D. This parameter will be equal to one if the percentage of the area of the GPW that is represented in the GHSL is independent of population size, meaning $d = 0$. This will make the coefficients statistically equivalent, even with different urban shapes and sizes. However, the positive correlation between the population size and the ratio of the areas considered by both GPW and GHSL data sets introduces a downward bias in the resulting Pareto coefficient from GPW. As shown in Figure 9, the importance of the difference between the estimated coefficients varies among income groups. Thinking about the case of richer countries, they generally exhibit higher urbanization rates and possess taller buildings (Jedwab, Loungani, and Yezer 2021; Lall et al. 2021). Thus, an increase in the population of an urban unit does not necessarily imply a larger area, but rather a higher density. As a consequence, the direct relationship between urban population and area becomes is weak, leading to a reduced value of D, hence making β_{GPW} and β_{GHSPOP} more alike.