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. In doing so, we exploit recent and accurate satellite imagery to proxy urban economic activity. Our results suggest that, at country level, urban population is more equally distributed than light emissions. Further, the degree of urbanization and the availability of natural resources are robustly related to the parameters that characterize national city size distributions. Calling assumptions established for urban nighttime lights into question, our findings do not support a Pareto function for their distribution. Moreover, we obtain evidence of a nonlinear and heterogeneous link between urban population and night lights. Grounded on our empirical analysis, we also provide a theoretical framework that relates the difference between the distributions of population and light emissions to the magnitude of agglomeration economies.

Keywords: City size distribution; Gridded population; Nighttime lights; Bayesian model averaging; Nonparametric methods.

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 XLVI International Conference on Regional Science (Madrid) and the XXIV Applied Economics Meeting (Palma de Mallorca). This work was supported by Fundación Ibercaja (Grant JIUZ-2020-SOC-11), Gobierno de Aragón (S39-20R 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

There is a well-established link between population, concentration, and economic activity at the urban level that, due to its theoretical and policy-making implications, motivates the study of the city size distribution. Following the seminal contributions of Gabaix (1999) and Eeckhout (2004), the related literature has mostly focused on testing whether the distribution of city sizes 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, see Soo (2005) and Puente-Ajovín, Ramos, and Sanz-Gracia (2020) for recent international comparisons. 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. Actually, this issue may lead to conflicting results even within a single country (Fazio and Modica 2015; Ioannides and Skouras 2013; Puente-Ajovín et al. 2020). Fortunately, there are several organizations that have established harmonized definitions of cities and settlements that can represent all the urban areas worldwide in a homogeneous framework.

Despite the relevance of the city size distribution from an urban economics point of view, most studies dealing with this topic measure the size of cities in demographic terms, taking for granted that the location of population determines the economic landscape. The main reasons are that it is difficult to find information about economic outcomes at the urban level and that, when available, it is not comparable across countries. Cities not only concentrate a large share of the population of a given country, but also of its economic activity. Moreover, the urban structure is the outcome of the dynamic interplay between economic activity and the growth process of cities (Arshad, Hu, and Ashraf 2018). Following Chen and Nordhaus (2011) and Henderson, Storeygard, and Weil (2012), this led Düben and Krause (2021) to make use of nighttime lights (NTL, hereafter) data compiled by satellites to proxy urban economic activity. The main conclusion drawn by these authors is that while the distribution of urban population can be characterized by Zipf's law in most countries, this is not the case of light emissions. To carry out their empirical analysis, Düben and Krause (2021) use the data set created by Bluhm and Krause (2022) to correct the top-coding problem of the 'stable night light images' collected by the Defense Meteorological Satellite Program (DMSP) Operational Linescan System. At this point, it is worth noting that these NTL data are also affected by blurring, geo-location 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 Visible Infrared Imaging Radiometer Suite (VIIRS) of instruments onboard the Suomi NPP satellite. The VIIRS 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 geo-location 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.

Taking into account previous arguments, the main aim of this paper is to contribute to the literature that compares the distributions of urban population and light emissions. Similarly to Puente-Ajovín, Sanso-Navarro, and Vera-Cabello (2022), we do so by proxying local economic activity with the NTL captured by VIIRS. Proceeding this way, 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), based on the assumption of a Pareto distribution for aggregate urban NTL. We also assess the sensitivity of our results to the role played by primary cities, and to the use of alternative gridded population and NTL data sets. Furthermore, we use the estimated power law coefficients from country rank-size regressions to search for robust determinants of city size distributions (Modica 2017; Sun et al. 2021; Wang, Wei, and Sun 2022). As another contribution, we explore the possible presence of a nonlinear and heterogeneous relationship between urban population and night lights.

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 distributions of urban population and aggregate nighttime lights at country level using parametric regressions and nonparametric tests. Adopting a Bayesian model averaging framework, Section 4 investigates the factors that display a robust relationship with the estimated coefficients characterizing national city size distributions. Section 5 evaluates the possible presence of a nonlinear and heterogeneous link between urban population and light emissions using kernel regression methods. Section 6 develops a simple theoretical framework to discuss of our main findings and, finally, Section 7 concludes. The Appendix contains further relevant information and results.

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

The first key issue when carrying out cross-country studies of the distribution of urban size is to adopt 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[1] (GPW) to divide the globe in 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. In fact, 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 the GHSL only includes areas with more than 50,000 inhabitants, this value corresponds to the threshold suggested

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

by the World Bank (2008) to classify human settlements as urban in both developed and developing countries. The geo-spatial data with the shape and location of urban centers reveals that some of them belong to more than one country[2]. 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.

The second relevant issue when dealing with urban size is its measurement. In line with the great majority of studies about the distribution of city size, we calculate it using population data. Nonetheless, and following Düben and Krause (2021), we also 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 by disagregating GPW administrative area level population data from national censuses and registers[3] to grid cells according to their proportion of built-up area.

In line with Puente-Ajovín, Sanso-Navarro, and Vera-Cabello (2022), and as suggested by Gibson (2021) and Gibson et al. (2021), we use the current and more precise VIIRS night lights to proxy urban economic activity. More specifically, we have been extracted the 'vcm-orm-ntl' annual composites[4] for 2015 from the website of the Earth Observation Group of the National Oceanic and Atmospheric Administration (US Department of Commerce)[5]. This 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). In the same manner as gridded population, NTL data have been aggregated for all pixels included within the extents of urban centers to calculate their

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

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

[4] VIIRS Cloud mask–Outlier removed–Nighttime lights.

[5] `https://www.ngdc.noaa.gov/eog/`.

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[6] 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.

3 The distribution of urban population and aggregate night-time lights at 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:

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

[6] See Table A1 in the Appendix for further details

where i is a city indicator, and n denotes the sample size. Zipf's law is equivalent to $\beta = 1$. In our 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 country level[7], calculating 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 country level.

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 city size distribution through the implementation of the Kolmogorov-Smirnov (KS) test statistic. 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 references: (i) a Pareto function imposing that the power law exponent is equal to one, 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:

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

[7] 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.

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 reference 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 observed city sizes.

[Insert Figures 2 and 3 about here]

We have first implemented the KS test against the null hypothesis that, at country level, city sizes are distributed as a Pareto function with power law exponent equal to one, i.e. the exact Zipf's law. 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 measured 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 more than 70 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. Nonetheless, this problem mainly affects larger cities which, according to the figures reported in Table 1, tend to be located in more developed countries. For this reason, we also carry out the analysis of how city sizes are distributed grouping countries by their level of income per capita.

3.3 Country income groups

Kernel density estimates of 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. Nonetheless, aggregate urban light emisssions are more unevenly distributed than population. The similarity between the distributions of population and night lights is directly related to the national income level. In particular, estimated Pareto coefficients for population (NTL) tend to increase (decrease) when GNI per capita decreases.

[Insert Figure 4 about here]

The upper panel of Table 2 reports, at different 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, high income countries 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 top-coding in 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 a measure of the degree of hierarchy in the urban system, such that a low coefficient is indicative of a high weight of large cities. Düben and Krause (2021) show that national primary shares are inversely related to the magnitude of estimated Pareto coefficients using both population and light emissions to measure city size. Moreover, these authors suggest that concentration in primary cities makes NTL to be more unevenly distributed than population. 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).

Primary cities in developing countries 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 our context, we are re-estimating expression (1) at country level once the largest city is removed from the sample. Kernel densities of resulting Pareto coefficients when city size is measured using population and NTL, grouped by national income per capita levels, are displayed in Figure 5. 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]

As expected, the distributions of estimated Pareto coefficients shown in Figure 5 tend to move to the right – reflecting higher values and, consequently, lower urban concentration – when primary cities are not included in national samples. Nothetheless, it can be observed that changes mainly affect rank-size regression results when city size is measured in demographic terms. In line with the related literature, the magnitude of the distributional shift is inversely related to the level of national income per capita. Therefore, this robustness check allows us to claim that the different distributions of urban light emissions and population are not driven by an excessive concentration in the largest cities. Actually, not considering primary cities lead to even greater differences between the estimated Pareto coefficients from the two alternative measures of urban size, especially in lower income countries. This is a surprising finding obtained from the use of more accurate satellite imagery than related studies.

3.4.2 Alternative nighttime lights data

For comparison purposes, 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[8] 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 light emissions in economic terms 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 country level using DMSP and VIIRS images are alike. However, the distribution of Pareto coefficients obtained using DMSP corrected data is more leptokurtic. This finding suggests 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 less affected by the top-coding problem of DMSP nighttime lights. 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 allows the estimated parameters that characterize the city size distribution to better reflect the higher degree of concentration of urban economic activity in LDCs.

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

[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 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 uncorrected 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[9] Leyk et al. (2019) and Archila Bustos et al. (2020) for two systematic reviews. In this section, we analyze the sensitivity of our results about the distribution of urban population at country level 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. LandScan and WorldPop adopt highly-modeled frameworks to disaggregate subnational census data that consist of implementing dasymetric mapping with more sophisticated statistical techniques – dynam-

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

ically 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 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 country level, similar to that of NTL. This is especially the case of LandScan and WoldPop, what can be related to their highly-modeled frameworks, and by the correlations between the variables included in their corresponding ancillary data sets. Furthermore, it is worth noting that WorldPop relies on DMSP images, among other information, to generate its population density predictions.

[Insert Figure 9 about here]

The distributions of Pareto coefficients at country level using the four gridded population data sets and grouped by income per capita levels are displayed in Figure 9. It can be observed that the differences between kernel densities are inversely related to national income. Urban sizes calculated using the GPW present the highest level of concentration and, with the exception of more developed countries, tend to display an average value around 0.5. As can be inferred from the descriptive statistics reported in Table A2 in the Appendix, GPW and, to a lesser extent, LandScan and Worldpop tend to underestimate the size of smaller urban centers as compared to GHS-POP, while this is not the case for the largest ones. This leads to an apparently more unequal distribution of population across urban centers and, as a result, lower estimated Pareto coefficients. Furthermore, 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. Corroborating previous 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.

Independently of previous resuls, we consider that the analysis based on the information extracted from GHS-POP is the most trustworthy for several reasons. First of all, the GHS-POP data set is produced by the same institution that establishes the definition of the urban units that have been studied. 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. Actually, Chen et al. (2020) claim that this database is more opportune to analyze highly-urbanized areas.

4 Searching for robust determinants of the city size distribution

The estimated Pareto coefficients from the rank-size regression (1) can be further explored to study the factors that determine the inequality displayed by city size distributions at the country level (Modica 2017; Sun et al. 2021; Wang, Wei, and Sun 2022). With this aim, Düben and Krause (2021) applied a selection method grounded on a simplistic algorithm over all models up to seven regressors from a set of 36 potential covariates. As a more flexible alternative to identify the robust determinants of the city size distribution, and in order to control for model uncertainty in this context, we have implemented Bayesian model averaging (BMA); see Raftery, Madigan, and Hoeting (1997). This technique allows us to investigate the influence of a large number of regressors by estimating all candidate models and then computing a weighted average of their results, taking into account the implicit uncertainty conditional on a given model and across different models. In doing so, model selection, estimation, and inference are handled simultaneously.

Assuming that the estimated Pareto coefficient linearly depends on a vector of covariates x, its conditional mean is given by:

$$E(\hat{\beta}_c|x_c) = x_c'\theta, \quad c = 1, \dots, C; \tag{5}$$

where C is the number of countries and θ is a set of parameters, estimated using maximum likelihood.

Model uncertainty is related to the choice of the regressors to include in x. More specifically, and for a total number of q variables, there are 2^q models (sets of regressors) to be estimated M_j, $j = 1, ..., 2^q$; each of them depending on a set of parameters θ^j with conditional posterior probability:

$$g(\theta^j|\hat{\beta}, M_j) = \frac{f(\hat{\beta}|\theta^j, M_j)g(\theta^j|M_j)}{f(\hat{\beta}|M_j)}; \tag{6}$$

with $f(\hat{\beta}|\theta^j, M_j)$ and $g(\theta^j|M_j)$ denoting, respectively, the likelihood function and the prior.

For a given prior model probability $P(M_j)$, its posterior probability can be calculated applying Bayes' rule:

$$P(M_j|\hat{\beta}) = \frac{f(\hat{\beta}|M_j)P(M_j)}{f(\hat{\beta})} \tag{7}$$

Expressions (6) and (7) show that it is necessary to specify priors, updated according to the data, for both model parameters and probabilities. Leamer (1978) assumed that θ is a function of θ^j in order to obtain the posterior density function of the parameters for all candidate models using the law of total probability. It is also possible to calculate posterior inclusion probabilities (PIP) for the q regressors by adding the posterior probabilities of the models that include them. Actually, Steel (2020) considers these posterior inclusion and model probabilities as virtues of the BMA methodology. Using the BMS R package developed by Zeugner and Feldkircher (2015), the estimation of the whole set of 2^q models has been avoided through a Metropolis-coupled Markov-chain Monte Carlo (MC3) sampler. Given that it should converge to a suitable distribution, the first 500,000 draws ('burn-ins') have been disregarded. As a baseline, our empirical analysis considers two million subsequent iterations, a hyper-g prior for model-specific parameters, and a uniform prior over the model space.

[Insert Table 4 about here]

Taking the set of regressors considered by Düben and Krause (2021) as a starting point, we have excluded those variables with correlation coefficients higher than 0.70 to avoid multicolinearity problems. Proceeding this way, we have selected 26 covariates – four of them continental dummies – as potential determinants of city size distributions at country level. A description of these variables, as well as their sources[10], are included in Table 4. These regressors capture national demographic and economic structures, physical geography, and institutional quality. We have also included the square of the percentage of urban population and of GDP per capita in order to capture the possible presence of a nonlinear relationship between these variables related to economic development and the distribution of city sizes[11]. The first three columns of results in Table 5 show, for each covariate, and when urban size is measured in demographic terms using GHS-POP data, the PIP and the mean and standard deviation (SD) of estimated parameters[12]. While inclusion probabilities reflect the importance of the variables in explaining the data, the mean and standard deviation can be interpreted, respectively, as a BMA point estimation and standard error.

[Insert Table 5 about here]

In line with Modica (2017), the results reported in Table 5 suggest that national city size distributions are related to economic and geographical factors. In particular, the urbanization rate, its square, and the Asian continental dummy receive inclusion probabilities higher than 80 per cent. These findings imply that urbanization has a nonlinear relationship with the equality of the distribution of urban population, on the one hand,

[10] The missing values present in the original sources have been completed using alternative data sets, mainly the Economic Indicators provided by Moody's Analytics (https://www.economy.com/indicators).

[11] Table A3 in the Appendix reports descriptive statistics for the variables that have been considered as potential determinants of the city size distribution both for the whole set of countries as well as by income group.

[12] Table A4 in the Appendix shows the results from the BMA analysis when urban size is calculated using the alternative gridded population and NTL data sources discussed in the previous section.

and that Asian countries display more uneven city size distributions, on the other. Natural resource rents and government final consumption expenditures also display high PIPs (0.67 and 0.57, respectively), and are directly related to the magnitude of estimated power law coefficients. This last result is in contrast with Sun et al. (2021), who find that the quality of infrastructures has a negative association with the Pareto coefficient of the city size distribution in demographic terms. The figures reported in the lower panel of Table 5 show that more than one million models have been visited by the MC3 sampler, with an average size of, approximately, nine covariates. The correlation coefficient between iteration counts and analytical posterior model probabilities for the 500 best models (0.93) indicates an adequate degree of convergence. In addition, the average shrinkage factor over all models, which can be interpreted as a Bayesian goodness-of-fit measure, is 0.91.

The last three columns of Table 5 report the results for city sizes calculated using VIIRS data. The square of the percentage of urban population receives the highest inclusion probability, also displaying positive average estimated coefficients. The other two variables showing high PIPs are latitude (0.66), with positive average coefficients, and the year of independence (0.59), inversely related to the estimated slope parameters in country rank-size regressions. The results for continental dummies suggest that European countries tend to have a more equal distribution of aggregate light emissions across cities. The degrees of convergence and the average shrinkage factors using VIIRS are equal to those obtained when city size is measured using GHS-POP gridded population.

[Insert Figures 10 and 11 about here]

A visual summary of the results described above is shown in Figures 10 and 11 for urban population and aggregate light emissions, respectively. Each graph ranks, vertically, the potential determinants of the city size distribution according to their PIPs. Likewise, the best 500 models are ordered, horizontally, taking into account their posterior probability. A colored rectangle reflects that the covariate is included in the model, and indicates the sign of its estimated influence (blue when positive, red when negative). The variables that tend to display high PIPs are the percentage of urban population, its squared term, and natural resource rents. These two variables exert the opposite influence on the distributions of urban population and NTL at country level. While the urbanization rate is directly and nonlinearly related to the equality of the city size distribution, a higher percentage of rents from natural resources over GDP is associated with a more uneven distribution. Therefore, it can be stated that the degree of urbanization and the availability of natural resources contribute to the observed differences between the distributions of urban population and light emissions at country level.

The choice of model-specific parameters may be determining previous findings, see Steel (2020). In order to assess their sensitivity, Figures 12 and 13 display inclusion probabilities for the potential determinants of the parameters that characterize the distributions of urban population and night lights, respectively, under different prior specifications; see Zeugner and Feldkircher (2015), and Forte, Garcia-Donato, and Steel (2018) for a description. It can be observed that the PIPs of the urbanization rate and its square are not affected by the choice of the prior on model-specific parameters. With the exception of the local empirical Bayes prior ('EBL'), inclusion probabilities for the other regressors are lower when constant g priors are used. This is especially the case of the risk inflation criterion ('RIC') and benchmark ('BRIC') priors. Therefore, this robustness check allows us to state that the conclusions drawn about the variables that display a more robust relationship with the estimated power law coefficients at the country level are not significantly affected by changes in the specification of model-specific parameters.

[Insert Figures 12 and 13 about here]

5 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:

$$Lights_i = \mathrm{m}(Popul_i) + \varepsilon_i, \quad i = 1, \ldots, n; \tag{8}$$

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:

$$\hat{\text{m}}(Popul) = \sum_{i=1}^{n} w_i Lights_i. \tag{9}$$

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

$$w_i = \frac{\text{K}\left(\frac{Popul_i - Popul}{h}\right)}{\sum_{j=1}^{n} \text{K}\left(\frac{Popul_j - Popul}{h}\right)}, \tag{10}$$

with $\text{K}(\cdot)$ 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

$$\text{CV}(h) = \frac{1}{n}\sum_{i=1}^{n} \left[Lights_i - \hat{\text{m}}_{-i}(Popul_i)\right]^2 \text{M}(Popul_i), \quad 0 \leq \text{M}(\cdot) \leq 1; \tag{11}$$

where $\text{M}(\cdot)$ is a weighting function[13], and

$$\hat{\text{m}}_{-i}(Popul_i) = \frac{\sum_{l \neq i}^{n} Lights_l \text{K}\left(\frac{Popul_i - Popul_l}{h}\right)}{\sum_{l \neq i}^{n} \text{K}\left(\frac{Popul_i - Popul_l}{h}\right)}. \tag{12}$$

The criterion in expression (11) 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

$$\text{K}\left(\frac{Popul_i - Popul}{h}\right) \to \text{K}(0) \quad \text{when} \quad h \to \infty. \tag{13}$$

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

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:

$$Lights_i = a + b'(Popul_i - Popul) + e_i, \quad i = 1, \ldots, n; \tag{14}$$

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

$$\min_{a,b} \sum_{i=1}^{n} [Lights_i - a - b'(Popul_i - Popul)]^2 \mathrm{K}\left(\frac{Popul_i - Popul}{h}\right). \tag{15}$$

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 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 comparability with the results obtained[14] by Düben and Krause (2021), Table 6 reports the estimated elasticities from fitting standard parametric OLS regressions to the relationship between urban light emissions and population in (8). 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. In line with the existing evidence, the response of light emissions to population is lower in larger cities. However, and as a

[14] See Table 3, page 201. Estimated elasticities using DMSP data for our sample can be found in Table A5 in the Appendix.

novelty, we conclude that an increase in population of primary cities is associated with a less than proportional increase in aggregate NTL.

[Insert Tables 6 and 7 about here]

The upper panel of Table 7 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 7 shows 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, which rejects both a standard linear OLS model (HLR1) and a quadratic specification for population (HLR2) in favor of the estimated nonparametric regression.

Table 8 contains descriptive statistics for the distribution of the estimated partial effects for population using a local-linear kernel regression, and the bandwidth parameter reported in the middle panel of Table 7 for the specification that only includes country fixed effects as additional regressors. These gradients show that the elasticity of NTL to population is heterogeneous. Although the response of light emissions to population tends to be lower in larger cities, the difference in the magnitude of estimated elasticies with the whole sample is less important than when cities are classified according to country income groups. In particular, the figures displayed in the lower panel of Table 8 show that the elasticity of urban night lights to population sharply decreases with the level of development.

[Insert Table 8 about here]

6 Discussion

The results reported in Table 8, 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

$$Rank_i - 0.5 = e^{\alpha} e^{log\left(Size_i^{-\beta}\right)}. \tag{16}$$

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

$$Rank_i - 0.5 = ALights_i^{-\beta_L}, \tag{17}$$

and

$$Rank_i - 0.5 = BPopul_i^{-\beta_P}; \tag{18}$$

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

$$Lights_i = DPopul_i^{\gamma}, \tag{19}$$

where D is a normalization constant, and γ denotes the scaling exponent which, in our 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 measure used to calculate its size:

$$Rank_i - 0.5 = ALights_i^{-\beta_L} = BPopul_i^{-\beta_P}. \tag{20}$$

Dividing this expression for the primary city and for an arbitrary urban center of rank r, and taking into account the scaling relation in (19), it is obtained that

$$\left(\frac{Lights_1}{Lights_r}\right)^{\beta_L} = \left(\frac{Popul_1}{Popul_r}\right)^{\gamma\beta_L} = \left(\frac{Popul_1}{Popul_r}\right)^{\beta_P}. \tag{21}$$

This implies that there exists a linear relationship between the Pareto coefficients that characterize the distributions of urban population and light emissions that depends on the scaling exponent (elasticity of NTL to population):

$$\beta_P = \gamma\beta_L. \tag{22}$$

The results from country rank-size regressions 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 (22), this is equivalent to saying that the elasticity of NTL to population is greater than one, and is precisely what we find in Section 6 considering urban centers worldwide in the estimations.

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). 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). Therefore, 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 displayed in Table 8 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 (22), 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 that, so far, have been investigated independently. Using data for functional urban areas, these authors show that urban systems with a more balanced distribution of population tend to have less pronounced increasing returns and, therefore, 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 exponent determines the difference between the national distributions of urban population and light emissions, characterized by β_L.

7 Concluding remarks

This paper compares the distributions of urban population and nighttime lights at country level. The sample that has been analyzed covers 12,852 urban centers in 100 countries of 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. In fact, the null hypothesis that city sizes adjust to Zipf's law can be more easily rejected when they are measured using VIIRS night lights. Furthermore, there is a higher similarity between the distributions of urban population and light emissions the higher the level of national income per capita. As a byproduct of our analysis, we also provide evidence that casts doubt on the Pareto assumption adopted to correct the top-coding problem inherent to DMSP images.

Using Bayesian model averaging techniques, we show that the urbanization rate has a robust, direct, and nonlinear relationship with the size distribution of cities. To a lesser extent, the availability of natural resources at country level is also associated with the parameters that characterize city size distributions. We also find a nonlinear and heterogeneous relationship between urban population and aggregate nighttime lights. In this regard, it is worth noting that the nonparametric estimation framework adopted has led us to obtain higher estimated elasticities of urban light emissions to population than those previously established in the related literature. Moreover, the heterogeneity displayed by these elasticities seems to be driven by the level of national income per capita rather than by urban hierarchy. The empirical analysis carried out has allowed us to theoretically establish the magnitude 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.

Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg. 2003. “Fractionalization”. *Journal of Economic Growth* 8 (2): 155–194. doi:10.2307/40215942.

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.

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.

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

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.

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.

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.

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: 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.

Forte, Anabel, Gonzalo Garcia-Donato, and Mark Steel. 2018. "Methods and tools for Bayesian variable selection and model averaging in normal linear regression". *International Statistical Review* 86 (2): 237–258. doi:10.1111/insr.12249.

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.

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.

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.

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: Random House. ISBN: 9780394422961.

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.

Leamer, Edward E. 1978. *Specification searches: Ad hoc inference with nonexperimental data.* New York: John Wiley & Sons. ISBN: 9780471015208.

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.

Modica, Marco. 2017. "The impact of the European Union integration on the city size distribution of the Member States". *Habitat International* 70:103–113. doi:10.1016/j.habitatint.2017.10.011.

Nunn, Nathan, and Diego Puga. 2012. "Ruggedness: The blessing of bad geography in Africa". *The Review of Economics and Statistics* 94 (1): 20–36. doi:10.1162/REST_a_00161.

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.

Raftery, Adrian E., David Madigan, and Jennifer A. Hoeting. 1997. "Bayesian model averaging for linear regression models". *Journal of the American Statistical Association* 92 (437): 179–191. doi:10.2307/2291462.

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, no. 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.

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.

Steel, Mark F. J. 2020. "Model averaging and its use in economics". *Journal of Economic Literature* 58 (3): 644–719. doi:10.1257/jel.20191385.

Sun, Bindong, Tinglin Zhang, Yu Wang, Liangliang Zhang, and Wan Li. 2021. "Are megacities wrecking urban hierarchies? A cross-national study on the evolution of city-size distribution". *Cities* 108:102999. doi:10.1016/j.cities.2020.102999.

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, Yu, Yehua Dennis Wei, and Bindong Sun. 2022. "New economy and national city size distribution". *Habitat International* 127:102632. doi:10.1016/j.habitatint.2022.102632.

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.

Zeugner, Stefan, and Martin Feldkircher. 2015. "Bayesian model averaging employing fixed and flexible priors: The BMS package for R". *Journal of Statistical Software* 68 (4): 1–37. doi:10.18637/jss.v068.i04.

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 the Appendix 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 function

	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 function

	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: Potential determinants of the city size distribution at country level: Description of variables and data sources.

Variable	Description	Source
popgr	Annual population growth rate, per cent	World Development Indicators
urban	Urban population, as percentage of total population	World Development Indicators
pop1400	Population in the year 1400	Nunn and Puga (2012)
netmigr	Net migration, persons	World Development Indicators
ethnic	Ethnic fractionalization	Alesina et al. (2003)
rugged	Terrain ruggedness	Nunn and Puga (2012)
coastprox	Coastal proximity	Nunn and Puga (2012)
coastbord	Coastal border, kilometers	CIA World Factbook
area	Land area, square kilometers	World Development Indicators
extreme	Droughts, floods, and extreme temperatures; as percentage of total population	World Development Indicators
resourents	Total natural resource rents, as percentage of GDP	World Development Indicators
latitude	Latitude	Nunn and Puga (2012)
colonherit	Colonial heritage	CEPII GeoDist
govexp	General government final consumption expenditure, as percentage of GDP	World Development Indicators
democracy	Polity score	Center for Systemic Peace
intwar	Interstate war	Issue Correlates of War Project
indep	Time of independence	Issue Correlates of War Project
trade	Trade of goods and services, as percentage of GDP	World Development Indicators
gdp	Gross domestic product, in 2015 US Dollars	World Development Indicators
gdppc	Gross domestic product per capita, in 2015 US Dollars	World Development Indicators
manuf	Manufacturing, as percentage of GDP	World Development Indicators
services	Services, as percentage of GDP	World Development Indicators
africa	Country located in Africa, indicator variable	World Development Indicators
asia	Country located in Asia, indicator variable	World Development Indicators
europe	Country located in Europe, indicator variable	World Development Indicators
northam	Country located in North America, indicator variable	World Development Indicators

Table 5: Determinants of the city size distribution at country level. Bayesian model averaging.

	GHSPOP			VIIRS		
Variable	PIP	Mean	SD	PIP	Mean	SD
popgr	0.36	0.01	0.02	0.32	-4.66E-03	0.01
urban	0.99	-0.02	0.01	0.39	4.04E-04	2.34E-03
urbansq	0.81	8.92E-05	6.47E-05	0.80	3.16E-05	2.28E-05
pop1400	0.44	1.12E-09	2.03E-09	0.23	-3.32E-11	7.38E-10
netmigr	0.37	7.73E-09	2.05E-08	0.24	-5.95E-10	9.19E-09
ethnic	0.43	-0.05	0.09	0.27	-0.01	0.04
rugged	0.35	0.01	0.02	0.25	1.79E-03	0.01
coastprox	0.51	-0.04	0.06	0.33	-0.01	0.03
coastbord	0.31	4.53E-08	6.44E-07	0.24	1.00E-10	3.87E-07
area	0.32	1.20E-09	7.24E-09	0.40	-3.44E-09	6.13E-09
extreme	0.40	-4.93E-03	0.01	0.23	-5.46E-04	4.32E-03
resourents	0.67	4.82E-03	4.84E-03	0.46	-1.62E-03	2.41E-03
latitude	0.33	1.88E-04	8.99E-04	0.66	1.09E-03	1.04E-03
colonherit	0.33	-0.01	0.05	0.28	0.01	0.03
govexp	0.57	4.03E-03	0.01	0.25	4.04E-04	1.68E-03
democracy	0.52	-4.66E-03	0.01	0.28	8.47E-04	2.64E-03
intwar	0.35	-0.01	0.03	0.23	-1.55E-05	0.01
indep	0.42	-4.85E-05	9.20E-05	0.59	-6.90E-05	7.75E-05
trade	0.31	4.28E-05	3.91E-04	0.25	4.01E-05	2.44E-04
gdp	0.31	1.95E-11	7.56E-09	0.24	-2.94E-10	4.21E-09
gdppc	0.33	-6.71E-07	3.24E-06	0.24	-3.58E-08	1.36E-06
gdppcsq	0.32	6.48E-12	3.94E-11	0.24	2.59E-12	1.73-11
manuf	0.34	5.22E-04	2.76E-03	0.29	5.18E-04	1.58E-03
services	0.40	-1.11E-03	2.42E-03	0.23	3.33E-05	9.54E-04
africa	0.55	-0.06	0.09	0.32	0.01	0.03
asia	0.81	-0.13	0.10	0.42	-0.02	0.04
europe	0.34	-0.01	0.07	0.49	0.04	0.06
northam	0.33	1.72E-04	0.06	0.25	-2.72E-03	0.03
Models		1,363,101			1,364,142	
Size		9.43			9.43	
Correlation		0.93			0.93	
Shrinkage		0.91			0.91	

Note: The dependent variable is the estimated slope parameter from a rank-size OLS regression at country level. The number of observations is 100. The birth-death MC3 sampler has been implemented with 500,000 burn-ins and two million iteration draws. The hyper-g and uniform priors have been established, respectively, for parameters and models. PIP denotes the posterior inclusion probability of each variable. Mean and SD are the posterior mean and standard deviation from model averaging. The lower panel reports the number of models visited, their average size, the correlation between iteration counts and analytical posterior model probabilities, and the mean of the shrinkage factor.

Table 6: 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 dependendent 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 7: VIIRS-GHSPOP elasticities. Least-squares cross-validation bandwidths and diagnostic test statistics for nonparametric kernel regressions.

	Upper	Local-constant estimation			
	bound	(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	Local-linear estimation			
	bound	(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 8: 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 7. Bootstrap standard errors (399 replications) in parentheses.

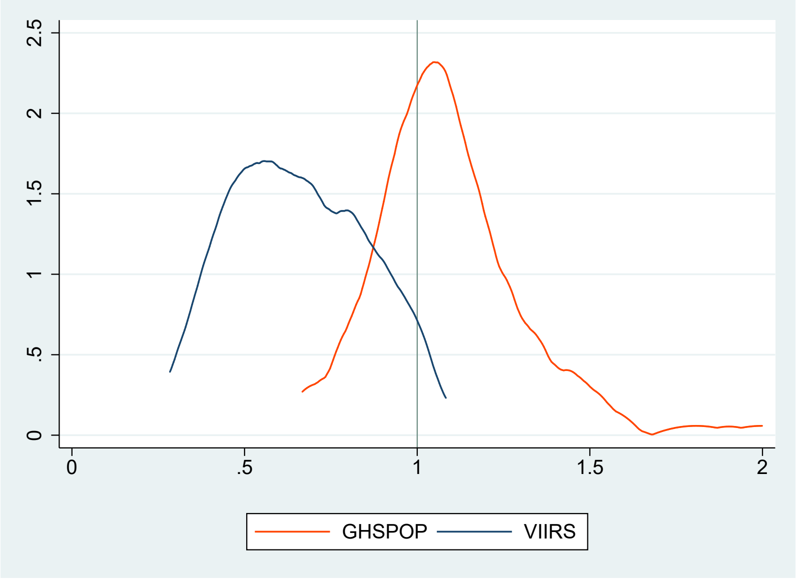

Figure 1: Kernel densities of estimated Pareto coefficients from a rank-size OLS regression at country level.

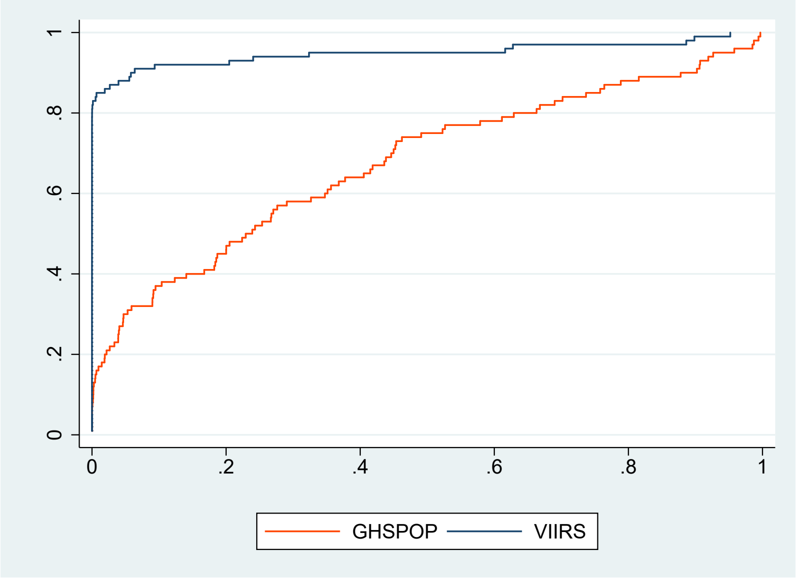

Figure 2: Cumulative distribution function of Kolmogorov-Smirnov test p-values using exact Zipf's law as a reference.

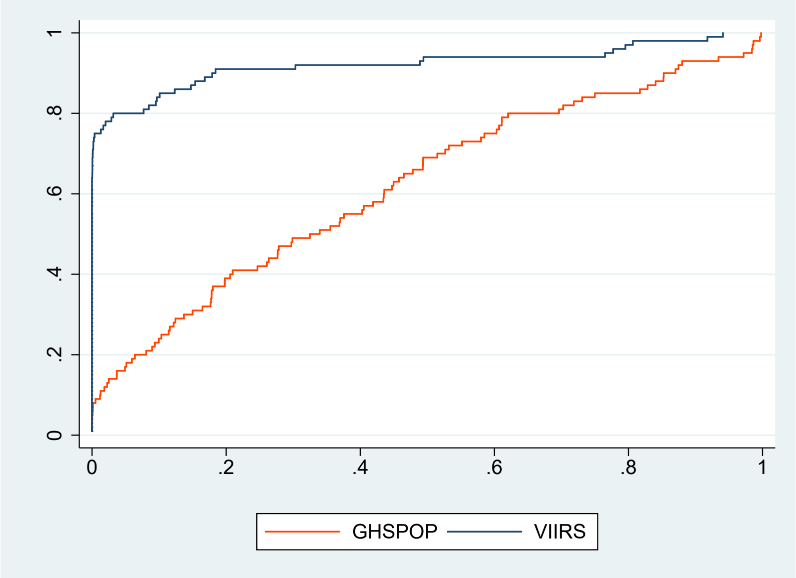

Figure 3: Cumulative distribution function of Kolmogorov-Smirnov test p-values using a Pareto distribution as a reference.

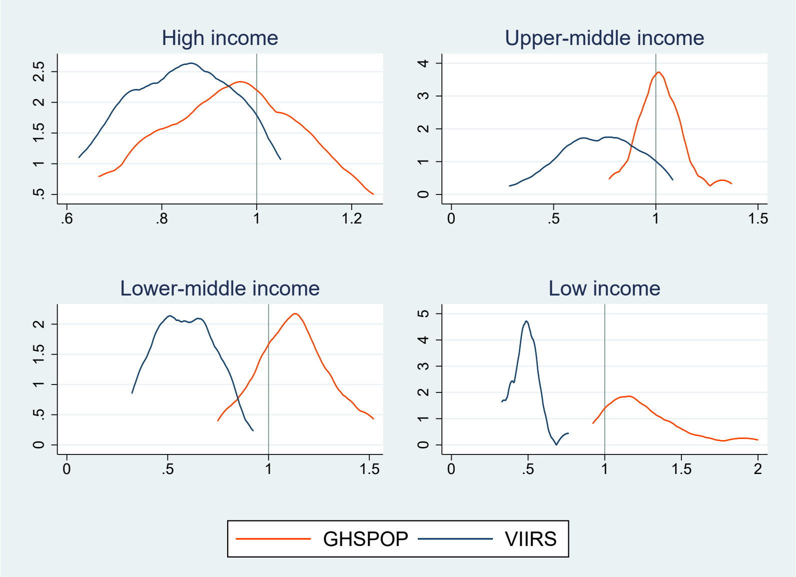

Figure 4: Kernel densities of estimated Pareto coefficients from a rank-size OLS regression at country level by income group, World Bank classification 2015.

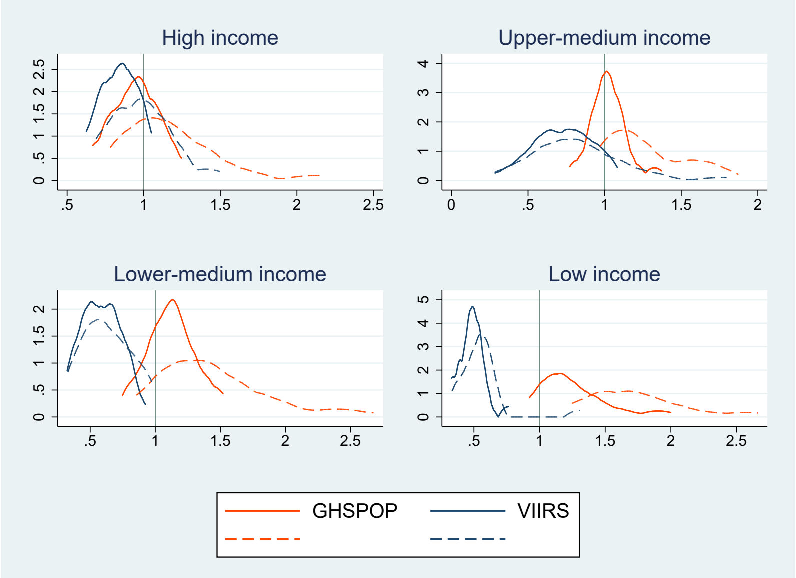

Figure 5: Robustness check: Kernel densities of estimated Pareto coefficients from rank-size OLS regressions at country level including (solid) and excluding (dashed) primary cities.

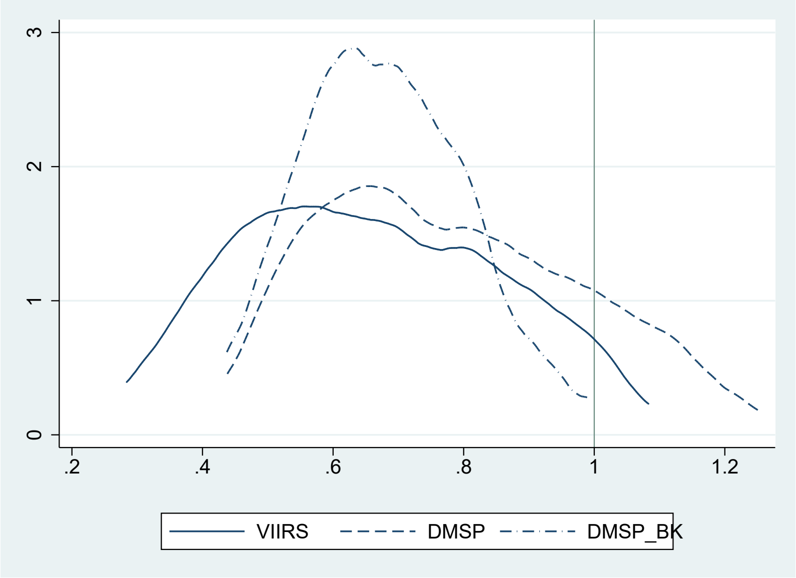

Figure 6: Robustness check: Kernel densities of estimated Pareto coefficients from rank-size OLS regressions at country level using alternative nighttime lights data.

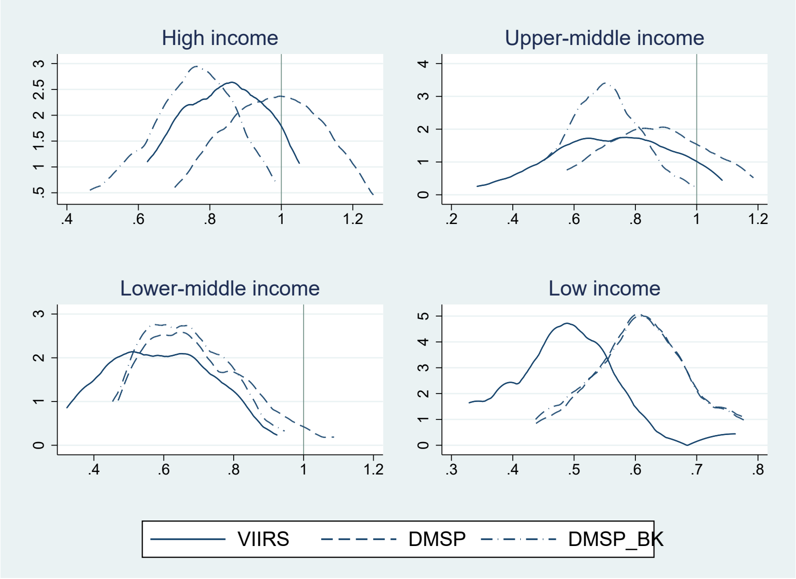

Figure 7: Robustness check: Kernel densities of estimated Pareto coefficients from country rank-size OLS regressions by income group using alternative nighttime lights data.

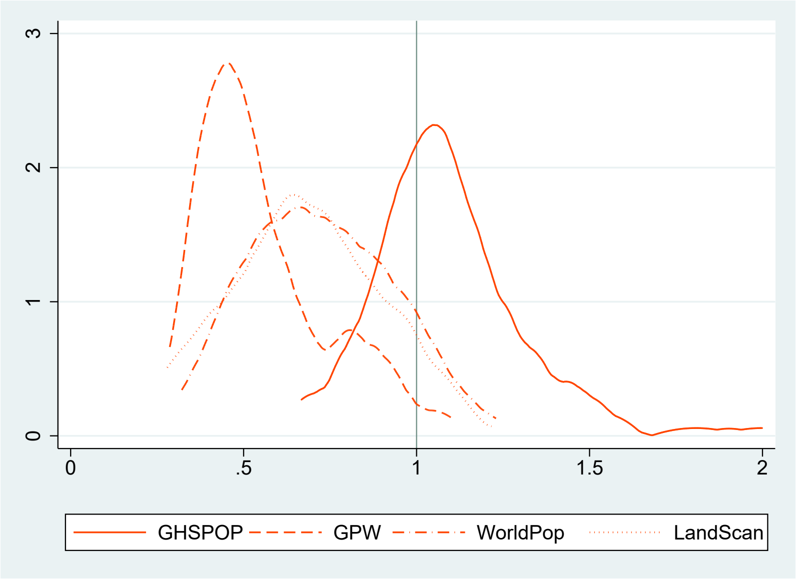

Figure 8: Robustness check: Kernel densities of estimated Pareto coefficients from rank-size OLS regressions at country level using alternative gridded population data.

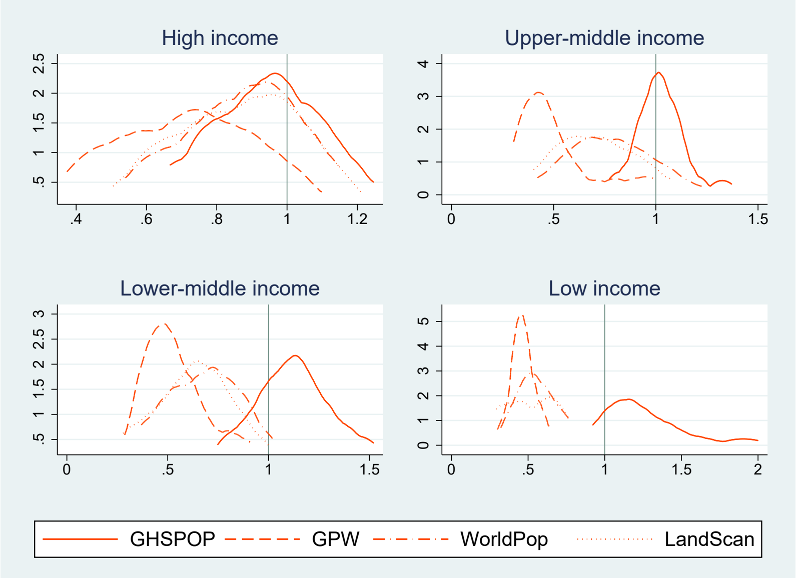

Figure 9: Robustness check: Kernel densities of estimated Pareto coefficients from country rank-size OLS regressions by income group using alternative gridded population data.

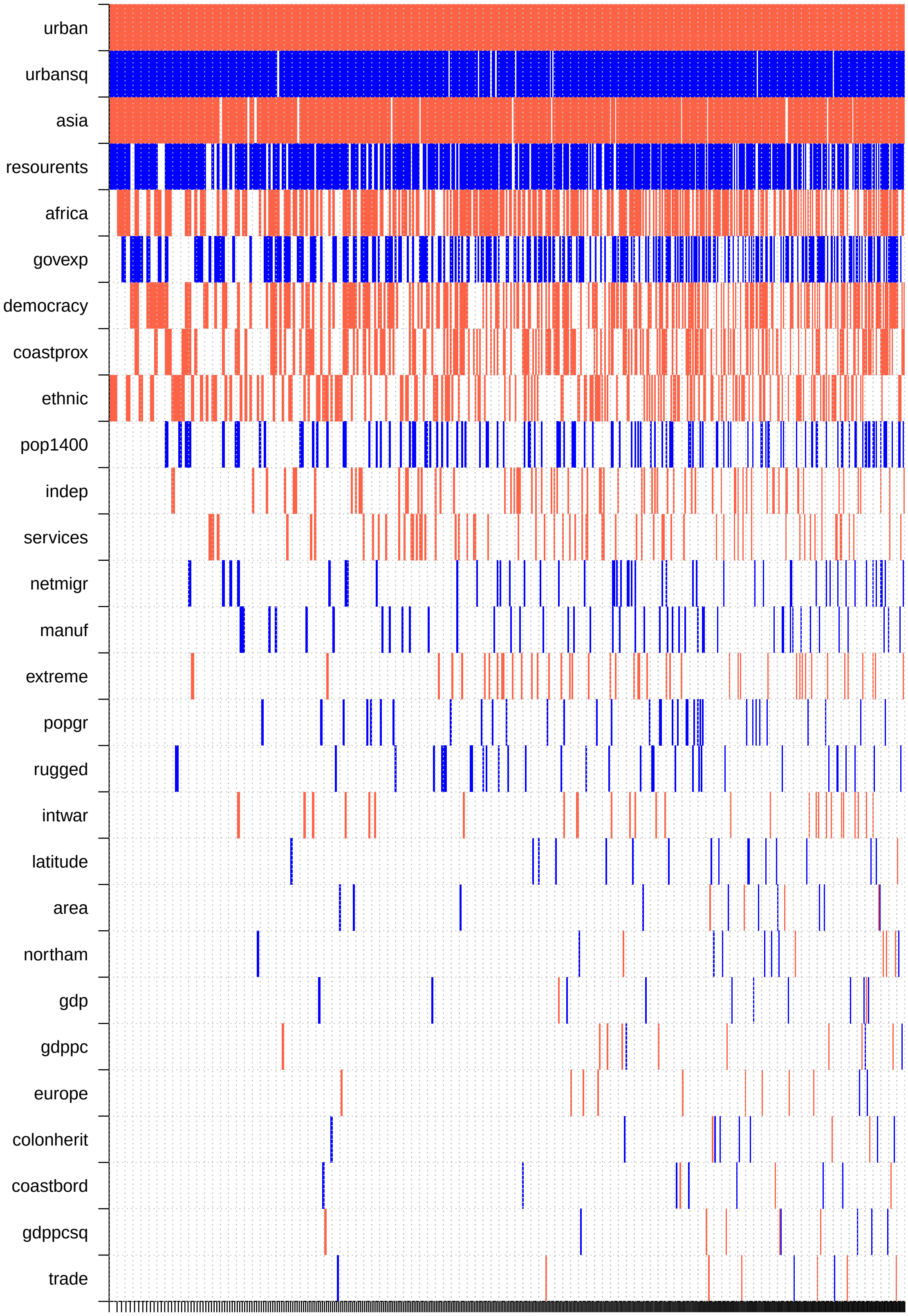

Figure 10: GHSPOP urban population: MC3 sampler results of the 500 best models. Colored areas reflect the inclusion of variables in the model, and whether their estimated parameters are positive (blue) or negative (red).

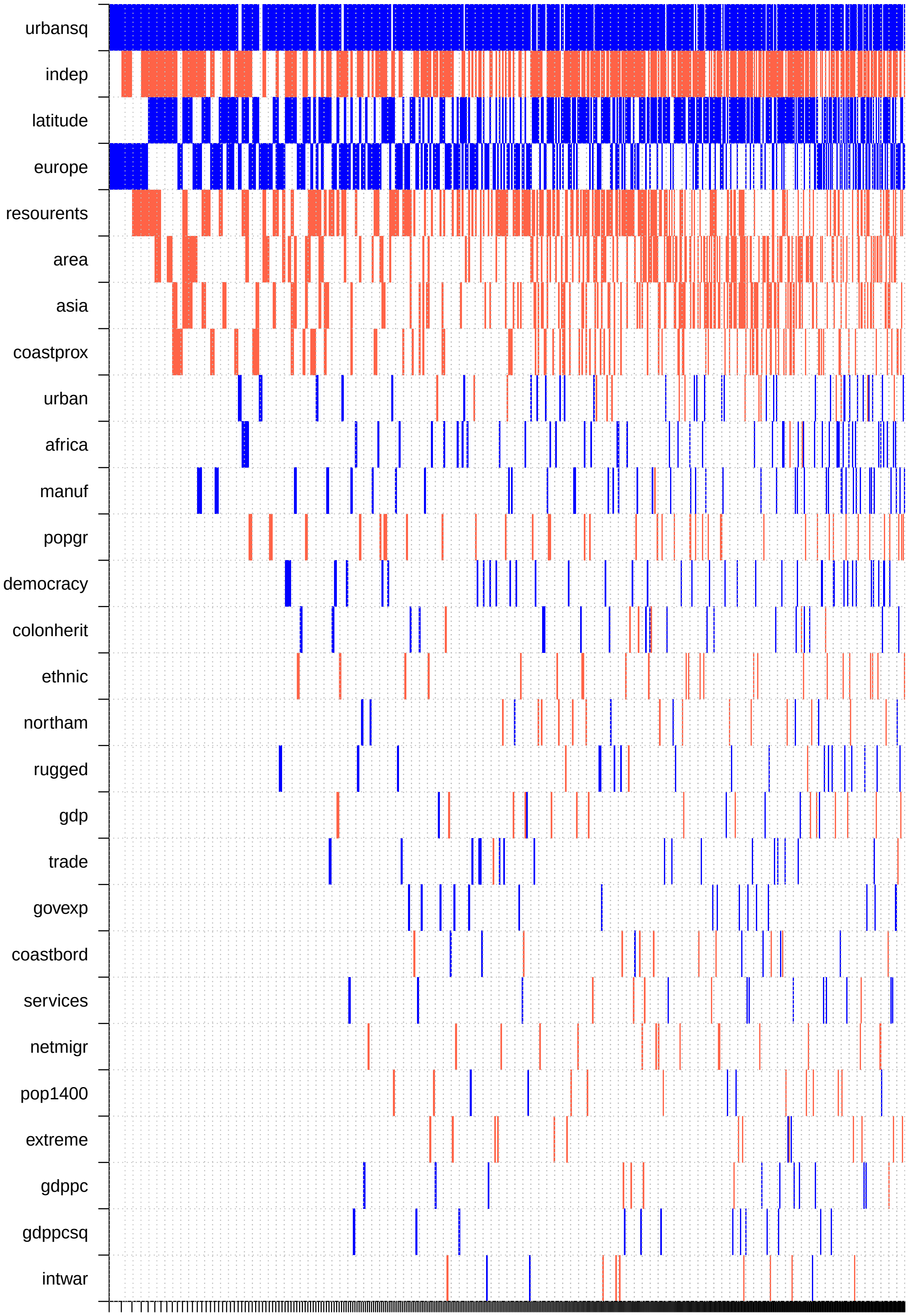

Figure 11: VIIRS nighttime lights: MC3 sampler results of the 500 best models. Colored areas reflect the inclusion of variables in the model, and whether their estimated parameters are positive (blue) or negative (red).

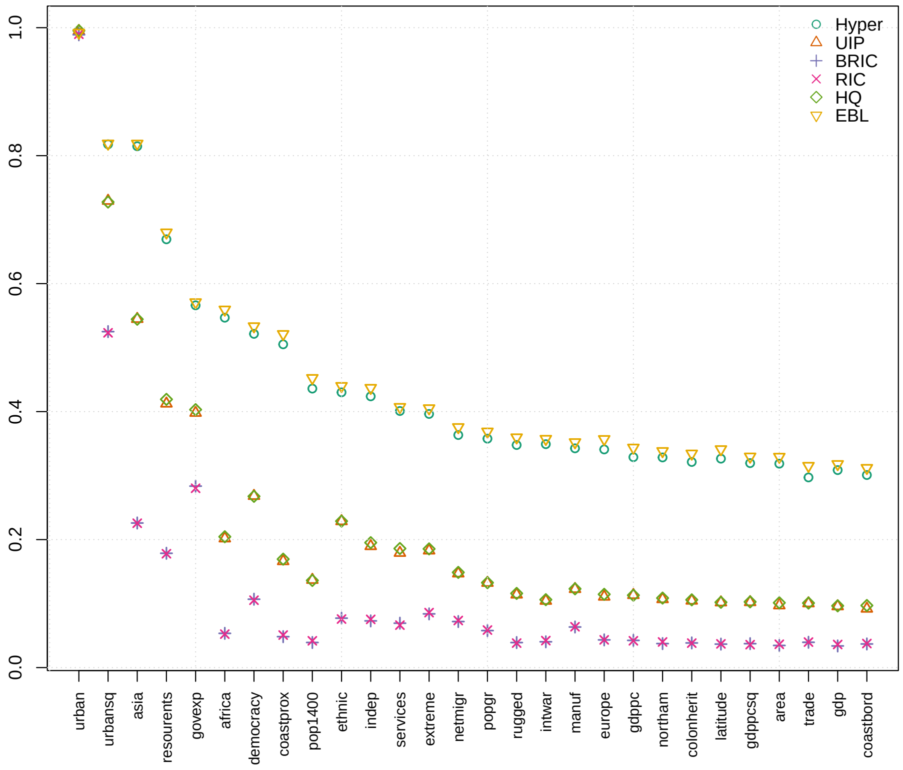

Figure 12: GHSPOP urban population: Posterior inclusion probabilities. Sensitivity analysis to alternative specifications of the prior for model-specific parameters.

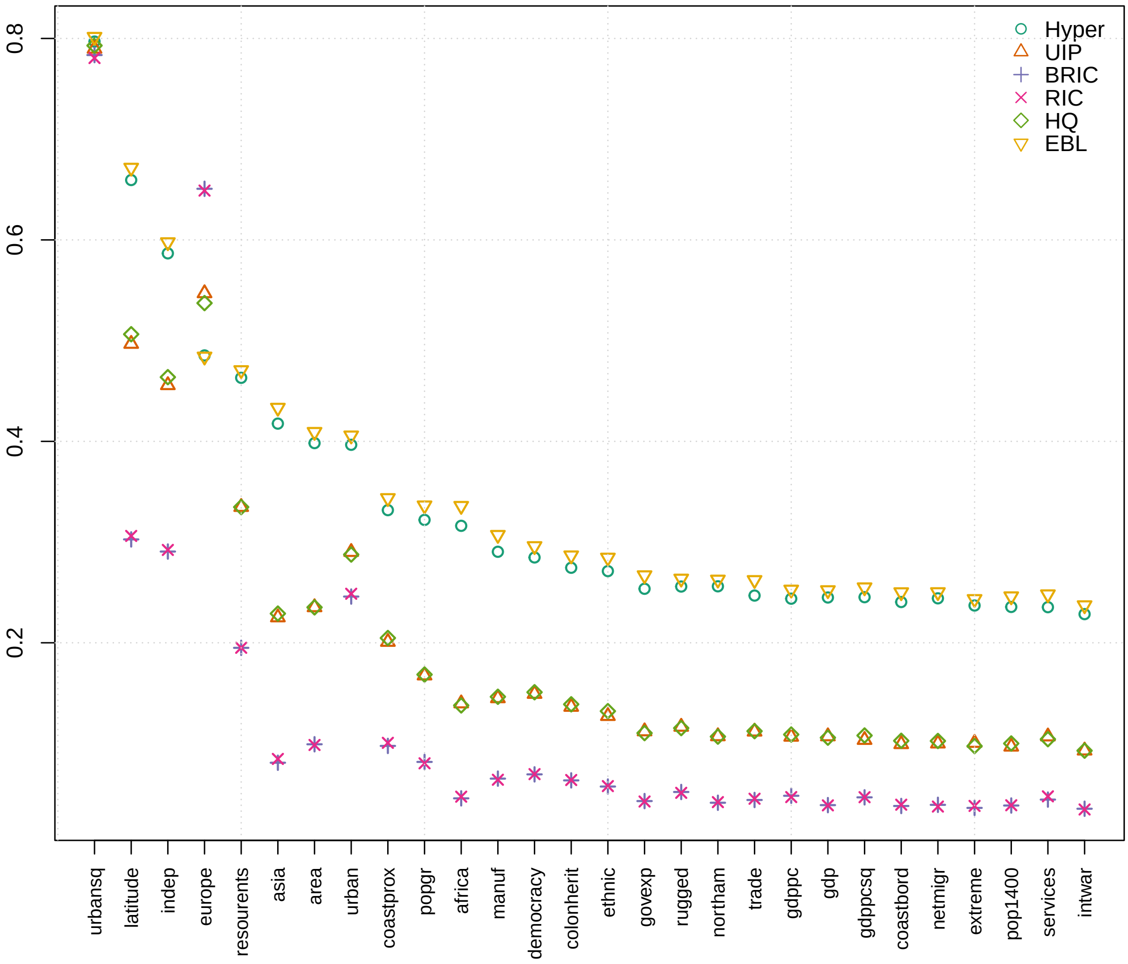

Figure 13: VIIRS nighttime lights: Posterior inclusion probabilities. Sensitivity analysis to alternative specifications of the prior for model-specific parameters.

Appendix

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
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
Minimum					
DMSP	0	194	0	0	0
DMSP_BK	0	194	0	0	0
GPW	0.58	18.54	5.03	1.08	0.58
WorldPop	3.16	817.94	53.32	3.16	3.24
LandScan	0	1,144	9	0	2.90
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

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.

Table A3: Potential determinants of the city size distribution at country level: Descriptive statistics by income group.

	All countries				High income				Upper-middle income				Lower-middle income				Low income			
Variable	Mean	Median	Minimum	Maximum	Mean	Median	Minimum	Maximum	Mean	Median	Minimum	Maximum	Mean	Median	Minimum	Maximum	Mean	Median	Minimum	Maximum
popgr	1.49	1.36	-3.89	5.79	0.79	0.55	-0.66	5.79	1.12	1.19	-0.49	3.44	1.55	1.69	-3.89	3.07	2.55	2.79	0.40	3.88
urban	56.33	55.6	12.08	97.88	80.43	81.30	60.28	97.88	69.52	73.36	47.69	91.50	43.58	46.28	13.01	69.06	32.82	33.39	12.08	61.28
urbansq	3,666.18	3,091.15	145.88	9,579.71	6,532.81	6,610.42	3,633.44	9,579.71	4,972.35	5,381.40	2,274.72	8,372.80	2,120.15	2,142.21	169.31	4,769.42	1,234.45	1,116.07	145.88	3,754.87
pop1400	3.43E+06	9.91E+05	39,145	8.09E+07	2.80E+06	1.06E+06	174,110	1.25E+07	4.39E+06	9.31E+05	39,145	8.09E+07	4.79E+06	1.32E+06	1.00+05	7.72E+07	1.09E+06	740786	51954	3776350
netmigr	-21,450.43	-50,001	-5.39E+06	4.96E+06	7.32E+05	3.78E+08	-520,442	4.96E+06	66,992.96	-38,001	-1.55E+06	1.80E+06	-5.62E+05	-1.50E+05	-5.39E+06	1.32E+05	-170832.23	-85,000.50	-2.04E+06	5.22E+05
ethnic	0.47	0.49	0	0.93	0.24	0.16	0.01	0.71	0.48	0.54	0.15	0.79	0.48	0.48	0	0.86	0.66	0.72	0	0.93
rugged	1.19	0.87	0.04	5.3	1.41	0.99	0.04	4.76	1.07	0.94	0.16	2.62	1.20	0.76	0.19	5.30	1.11	0.68	0.14	5.04
coastprox	0.44	0.31	0.02	2.21	0.25	0.15	0.02	1.43	0.54	0.38	0.02	2.21	0.38	0.28	0.02	1.65	0.57	0.48	0.02	1.25
coastbord	6,023.51	1,314	0	2.02E+05	15,571.02	2,929	0	2.02E+05	4,245.48	2,414	0	37,653	4,416.62	853	0	54,716	776.27	18.50	0	4,828
area	1.20E+06	4.43E+05	25,680	1.64E+07	1.48E+06	307,845	30,280	9.15E+06	2.11E+06	882,050	48,310	1.64E+07	6.07E+05	4.46E+05	61,893	2.97E+06	5.95E+05	4.84E+05	25,680	2.27E+06
extreme	1.22	0.41	0	7.95	0.58	0.03	0	7.95	0.92	0.17	0	7.95	1.65	0.83	0	6.64	1.66	0.86	0	7.53
resourents	5.89	3.04	0.01	34.25	2.72	0.29	0.01	24.11	7.60	3.68	0.21	34.25	4.48	2.96	0.12	22.18	8.81	10.32	0.10	20.12
latitude	19.72	18.92	-37.94	62.78	37.33	46.14	-37.94	62.78	18.61	23.94	-35.40	61.99	16.17	14.82	-16.71	49.01	8.15	9.14	-19.38	40.14
colonherit	0.16	0	0	1	0.45	0	0	1	0.19	0	0	1	0	0	0	0	0.05	0	0	1
govexp	15.06	14.38	5.15	32.21	19.90	19.57	11.39	30	15.40	14.60	7.41	32.21	12.22	11.76	5.40	19.82	13.52	13.35	5.15	25.93
democracy	5.26	6	-7	10	8.86	10	0	10	4.22	5	-7	9	4.69	7	-7	9	3.68	5	-7	7
intwar	0.60	1	0	1	0.91	1	0	1	0.70	1	0	1	0.45	0	0	1	0.36	0	0	1
indep	1,837.76	1,947.50	0	1,993	1,614.18	1,800.50	0	1,993	1,825	1,844	1,368	1,991	1,943.31	1,956	1,825	1,991	1,937.86	1,960	1,768	1,975
trade	67.48	59.24	7.36	178.77	83.45	68.64	27.76	167.24	65.01	59.70	22.49	131.37	65.55	53.92	7.36	178.77	57.09	57.43	16.54	124.11
gdp	7.11E+06	89,081.98	3,104.39	1.82E+07	2.01E+06	9.51E+05	68,420.26	1.82E+07	7.62E+05	1.84E+05	27,842.13	1.11E+07	1.98E+05	62,186.19	8,271.45	2.10E+06	19,789	13,981.08	3,104.39	64,589.33
gdppc	10,718.32	3,928.16	293.46	84,776.14	35,537.45	35,799.41	12,578.50	84,776.14	7,602.34	6,229.10	4,166.98	17,300	2,303.46	2,085.10	978.40	3,994.64	815.68	760.36	293.46	1,700
gdppcsq	3.6E+08	1.54E+07	86,115.94	7.19E+09	1.58E+09	1.28+09	1.58E+08	7.19E+09	6.74E+07	3.88E+07	1.74E+07	2.99E+08	6.13E+06	4.35E+06	9.57E+05	15957124	7.99E+05	5.78E+05	86,115.94	2.89E+06
manuf	13.73	12.69	1.78	47.60	15.25	12.69	6.29	36	14.87	14.42	2.18	28.95	13.13	13.69	1.78	21.31	11.58	9.36	2.78	47.60
services	52.17	52.12	23.36	76.78	63.94	66.99	50.20	76.78	52.94	53.48	34.90	72.76	48.71	48.57	23.36	61.45	44.01	44.79	27.42	58.65
africa	0.32	0	0	1	0	0	0	0	0.15	0	0	1	0.34	0	0	1	0.82	1	0	1
asia	0.27	0	0	1	0.23	0	0	1	0.22	0	0	1	0.45	0	0	1	0.14	0	0	1
europe	0.21	0	0	1	0.59	1	0	1	0.26	0	0	1	0.03	0	0	1	0	0	0	0
northam	0.09	0	0	1	0.09	0	0	1	0.11	0	0	1	0.10	0	0	1	0.05	0	0	1

Table A4: Robustness check: Determinants of the city size distribution at country level. Bayesian model averaging.

	DMSP			DMSP_BK			GPW			WorldPop			LandScan		
Variable	PIP	Mean	SD	PIP	Mean	SD	PIP	Mean	SD	PIP	Mean	SD	PIP	Mean	SD
popgr	0.22	1.37E-03	0.01	0.36	-9.03E-04	0.01	0.38	-0.01	0.01	0.57	-0.01	0.02	0.28	-1.16E-03	0.01
urban	0.33	5.25E-05	2.27E-03	0.45	3.15E-05	1.24E-03	0.55	-1.17E-03	1.89E-03	0.52	1.05E-03	1.86E-03	0.58	1.69E-03	2.37E-03
urbansq	0.88	3.91E-05	2.23E-05	0.50	5.70E-06	1.26E-05	0.47	-3.99E-06	1.71E-05	0.47	5.99E-06	1.69E-05	0.58	1.54E-05	2.19E-05
pop1400	0.20	3.67E-11	6.21E-10	0.37	1.48E-10	6.55E-10	0.29	7.38E-11	8.49E-10	0.32	3.60E-10	1.06E-09	0.28	3.08E-12	9.27E-10
netmigr	0.20	-3.55E-10	7.42E-09	0.35	-2.37E-10	7.47E-09	0.38	6.54E-09	1.47E-08	0.29	1.79E-11	1.25E-08	0.31	-4.09E-09	1.30E-08
ethnic	0.21	2.56E-03	0.03	0.35	-1.29E-04	0.03	0.29	-4.79E-03	0.04	0.45	-0.04	0.07	0.34	-0.02	0.05
rugged	0.22	1.64E-03	0.01	0.36	2.37E-04	0.01	0.33	-3.05E-03	0.01	0.27	-5.16E-04	0.01	0.50	-0.01	0.02
coastprox	0.20	1.71E-04	0.02	0.37	-2.28E-03	0.02	0.51	-0.03	0.04	0.29	-4.41E-03	0.03	0.29	-1.34E-03	0.03
coastbord	0.22	-5.68E-08	3.42E-07	0.39	-1.11E-07	3.48E-07	0.37	2.56E-07	5.69E-07	0.27	-3.05E-08	4.38E-07	0.47	-5.22E-07	8.08E-07
area	0.20	-2.88E-10	2.97E-09	0.37	-6.64E-10	3.19E-09	0.38	2.60E-09	6.14E-09	0.33	2.15E-09	5.50E-09	0.42	-3.96E-09	7.26E-09
extreme	0.28	-2.02E-03	5.28E-03	0.42	-1.89E-03	4.39E-03	0.29	-6.64E-04	4.80E-03	0.28	-1.08E-03	5.28E-03	0.31	-1.83E-03	0.01
resourents	0.24	3-47E-04	1.29E-03	0.38	-3.27E-04	1.14E-03	0.61	-2.59E-03	2.87E-03	0.40	-1.22E-03	2.32E-03	0.53	-2.23E-03	2.95E-03
latitude	0.89	2.02E-03	1.06E-03	0.73	7.88E-04	6.99E-04	0.74	1.26E-03	1.04E-03	0.79	1.68E-03	1.22E-03	0.88	2.06E-03	1.16E-03
colonherit	0.23	-3.75E-03	0.03	0.40	-0.01	0.03	0.63	-0.06	0.06	0.33	-0.01	0.04	0.51	-0.05	0.07
govexp	0.35	1.26E-03	2.39E-03	0.47	9.65E-04	1.73E-03	0.35	8.35E-04	2.14E-03	0.35	1.02E-03	2.40E-03	0.29	4.55E-04	1.98E-03
democracy	0.21	4.87E-05	1.81E-03	0.41	8.00E-04	2.07E-03	0.44	2.15E-03	3.80E-03	0.41	2.09E-03	3.86E-03	0.39	1.93E-03	3.89E-03
intwar	0.22	3.40E-03	0.02	0.35	9.78E-04	0.01	0.68	0.04	0.04	0.31	0.01	0.02	0.28	-2.89E-03	0.02
indep	0.47	-4.79E-05	6.71E-05	0.61	-4.23E-05	5.15E-05	0.84	-1.48E-04	9.92E-05	0.64	-8.66E-05	9.01E-05	0.57	-8.16E-05	9.97E-05
trade	0.26	8.48E-05	2.58E-04	0.37	3.54E-05	1.91E-04	0.30	2.79E-05	2.58E-04	0.27	-4.27E-06	2.56E-04	0.27	2.10E-05	2.73E-04
gdp	0.21	-1.95E-10	3.51E-09	0.36	3.21E-10	3.45E-09	0.32	1.37E-09	5.37E-09	0.28	8.76E-10	5.16E-09	0.30	1.41E-09	5.86E-09
gdppc	0.21	-5.44E-08	1.10E-06	0.42	-5.07E-07	1.93E-06	0.49	1.09E-06	2.27E-06	0.31	-5.39E-07	2.26E-06	0.29	2.77E-07	1.99E-06
gdppcsq	0.21	1.24E-12	1.39E-11	0.46	1.13E-11	2.53E-11	0.45	1.07E-11	2.76E-11	0.29	6.27E-12	2.79E-11	0.29	-4.84E-12	2.53E-11
manuf	0.21	7.46E-05	1.05E-03	0.41	4.24E-04	1.16E-03	0.30	-5.09E-06	1.40E-03	0.40	1.13E-03	2.19E-03	0.40	1.15E-03	2.28E-03
services	0.24	2.50E-04	9.94E-04	0.37	1.88E-04	8.51E-04	0.76	3.16E-03	2.49E-03	0.34	6.15E-04	1.53E-03	0.29	2.10E-04	1.25E-03
africa	0.29	0.01	0.03	0.45	0.01	0.02	0.38	0.02	0.04	0.35	0.01	0.04	0.30	0.01	0.03
asia	0.30	-0.01	0.03	0.51	-0.02	0.03	0.33	-2.78E-03	0.03	0.50	-0.03	0.05	0.37	-0.02	0.04
europe	0.49	0.04	0.06	0.51	0.02	0.04	0.43	0.03	0.05	0.50	0.05	0.07	0.36	0.02	0.05
northam	0.42	-0.03	0.05	0.42	-0.01	0.03	0.32	0.01	0.04	0.31	0.01	0.04	0.29	4.82E-03	0.04
Models		1,180,227			1,898,094			1,571,543			1,568,664			1,518,180	
Size		8.59			11.93			12.61			10.87			10.96	
Correlation		0.98			0.98			0.57			0.69			0.75	
Shrinkage		0.93			0.93			0.85			0.87			0.86	

Note: See Table 5 .

Table A5: 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 dependendent 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.

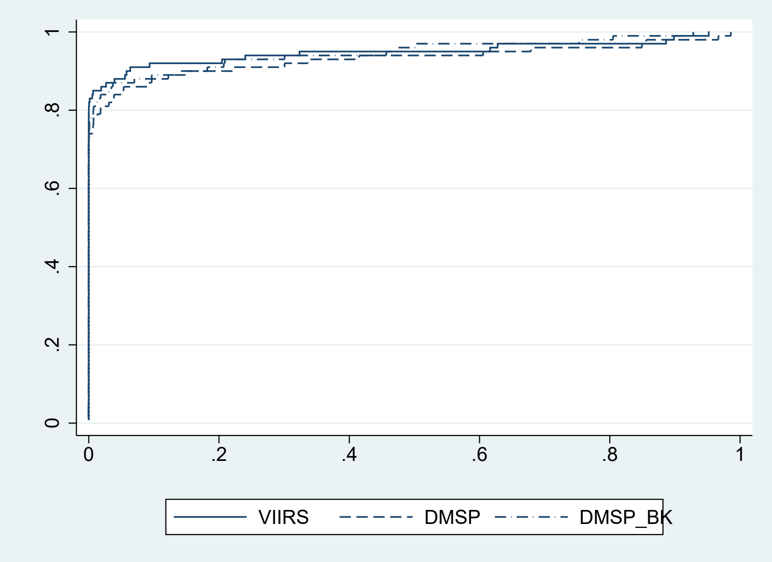

Figure A1: Robustness check: Cumulative distribution function of Kolmogorov-Smirnov test p-values using exact Zipf's law as a reference and alternative nighttime lights data.

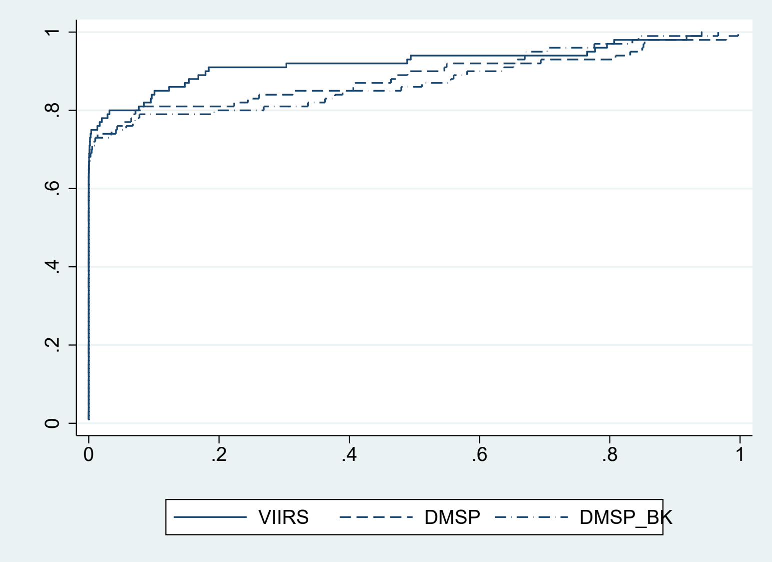

Figure A2: Robustness check: Cumulative distribution function of Kolmogorov-Smirnov test p-values using a Pareto distribution as a reference and alternative nighttime lights data.

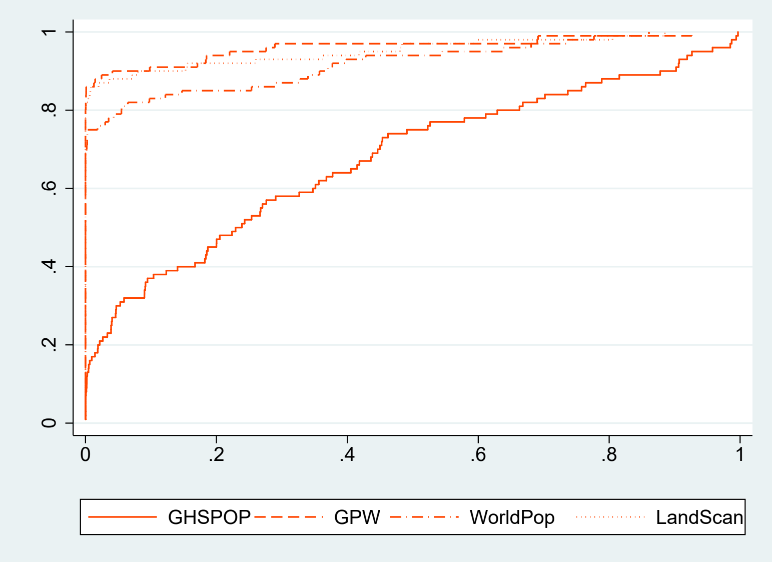

Figure A3: Robustness check: Cumulative distribution function of Kolmogorov-Smirnov test p-values using exact Zipf's law as a reference and alternative gridded population data.

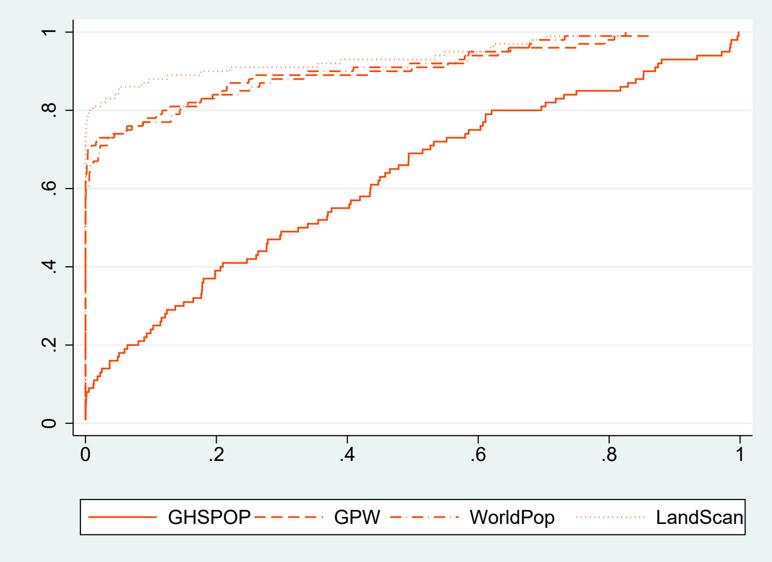

Figure A4: Robustness check: Cumulative distribution function of Kolmogorov-Smirnov test p-values using a Pareto distribution as a reference and alternative gridded population data.