Working Paper Series

Population, light, and the size distribution of cities

Christian Düben
Melanie Krause

ECINEQ WP 2019 - 488

ECINEQ 2019 - 488
January 2019

www.ecineq.org

Population, light, and the size distribution of cities*

Christian Düben†

Melanie Krause

Hamburg University, Germany

Abstract

We provide new insights on the city size distribution of countries around the world. Using geo-spatial data and a globally consistent city identification scheme, our data set contains 13,844 cities in 194 countries. City size is measured both in terms of population and night time lights proxying for local economic activity. We find that Zipf's law holds for many, but not all, countries in terms of population, while city size in terms of light is distributed more unequally. These deviations from Zipf's law are to a large extent driven by an undue concentration in the largest cities. They benefit from agglomeration effects which seem to work through scale rather than through density. Examining the cross-country heterogeneity in the city size distribution, our model selection approach suggests that historical factors play an important role, in line with the time of development hypothesis.

Keywords: Cities, Zipf's Law, Urban Concentration, Geo-spatial Data.

JEL Classification: R11, R12, O18, C18.

*This paper has been presented at the Urban Economics Conferences in New York City and Dusseldorf, the European Regional Science Congress in Cork, the Development Economics and Policy Conference in Zurich, the 8th ifo Dresden Workshop on Regional Economics in Dresden, and various seminars. We would like to thank Richard Bluhm, David Castells-Quintana, Lewis Dijkstra, Gilles Duranton, Martin Gassebner, Roland Hodler, Remi Jedwab, Christian Lessmann, Adam Storeygard, and David Weil for very helpful comments and suggestions. We gratefully acknowledge financial support from the Germany Science Foundation (DFG). All remaining errors are unfortunately ours.

†**Contact details**: Hamburg University, Department of Economics, Von-Melle-Park 5, 20146 Hamburg, Germany. E-mails: christian.dueben@uni-hamburg.de and melanie.krause@uni-hamburg.de.

1 Introduction

The rapid urbanization process around the world has brought renewed interest to the question of the size distribution of cities. Many countries are urbanizing at relatively early stages of development (Jedwab and Vollrath, 2018, Glaeser, 2014), and the total urban population of the world is set to grow by 2.7 bn people from 2015 to 2050 (United Nations, 2018). Whether this growth is distributed equally among cities of all size, or whether the largest metropolises are further increasing their share, is an important question for policymakers. Bigger cities can reap agglomeration benefits by pooling physical and human capital as well as by exploiting spillovers across and within industries, but they also suffer from plights such as congestion (Rosenthal and Strange, 2004, Desmet and Rossi-Hansberg, 2013).

It has long been argued that Zipf's Law should be an appropriate description of the size distribution of big cities within a country, implying that a city's rank is approximately inversely proportional to its size (Zipf, 1949). There are strong theoretical arguments for Zipf's Law as well as empirical evidence with population data, in particular for the U.S. (Gabaix, 1999, Rozenfeld et al., 2011). But cross-country studies have so far yielded mixed results (Rosen and Resnick, 1980, Soo, 2005). Unfortunately, they are often limited to countries with a high statistical quality and miss out on large parts of the rapidly urbanizing developing world. Population data on a country's cities ultimately rely on national statistical agencies. This implies not only large cross-country variations of data quality and availability across time, but also in terms of definitions of what constitutes a city. This is exemplified by a look at the database www.citypopulation.de (Brinkhoff, 2017), a regularly updated compilation of national census statistics which has been used by many cross-country studies (e.g. Soo, 2005, Henderson and Wang, 2007): The listed cities start at 50,000 inhabitants in the UK, 150,000 in Japan, and 15,000 in Egypt. By contrast, the global threshold of 300,000 inhabitants imposed by the database of the U.N. World Urbanization Prospects misses out on locally important cities in less populous countries. In fact, the World Bank (2009) argue that a threshold of 50,000 people is reasonable for a sizable settlement in developed and developing countries.

In this paper, we systematically examine the size distribution of cities with a truly global data set. We exploit geo-spatial data as well as recent theoretical advances on the origin of the city size distribution. In particular, we use a consistent city identification scheme across countries based on the European Commission's Global Human Settlement Layers (GHSL) and include all cities with more than 50,000 inhabitants in every country of the globe. In this way, we contour the issues of limited data availability and comparability plaguing earlier studies. Another advantage of working with such geo-spatial data is that we can define cities based on their de facto geographical extent rather than working with administrative city boundaries. This can capture the economic and social reality

of cities more appropriately. With our approach, we arrive at a data set of 13,844 cities in 194 countries.[1] Using this data set, our paper might be called the *Zipf paper for the geo-spatial age*. In particular, we carry out the following three investigations: (i) We analyze for each country whether Zipf's law holds in terms of population and examine deviations. (ii) In addition to population, our geo-spatial data allows us to measure city size by nighttime lights proxying for economic activity. To our knowledge, we are the first to investigate whether the size distribution of economic activity in countries also follows Zipf's Law. (iii) We investigate the geographical, institutional and historical determinants of the cross-country variation we observe in the city size distribution, and we look at changes over time.

More broadly, with our data analysis we extend the Zipf question into a general discussion of how to measure cities' economic and social importance. Satellite data of nighttime lights have been shown to be an appropriate proxy for local economic activity (Henderson et al., 2012, Donaldson and Storeygard, 2016) and are increasingly used for the study of cities in developing countries where up-to-date population data are lacking (Bluhm and Krause, 2018, Storeygard, 2016, Fetzer et al., 2016). But it is an open question whether light per capita is the same for all cities in a country and how light output in cities responds to changes in population. We address these points by comparing population and light across the whole city size distribution. We examine the particular role played by primary cities, which have been linked to autocratic structures (Ades and Glaeser, 1995), political centralization (Davis and Henderson, 2003) and low levels of development in general (Henderson, 2003). If luminosity in the largest cities is more responsive to population changes, it suggests that positive agglomeration effects are being reaped. But if primary cities are systematically brighter than the rest and relatively inelastic to population changes, other factors might play a role, such as wasteful political spending (Lipton, 1977).

Going further, our paper exploits the geo-spatial structure of the data and decomposes city size, in terms of both light and population, into the product of area and density. While larger cities are typically both more extended and denser than smaller ones, we examine which factor dominates. This carries vital insights for the types of agglomeration effects at play: Economies of scale and the wider market access of bigger cities (Krugman, 1991, Fujita et al., 1999) mainly operate through a larger area, while more frequent human capital interactions, as highlighted by Moretti (2004) and Bettencourt (2013), are arguably working through a higher density.

The main results of our paper are the following: (i) For many countries, the city size distribution in terms of population can be characterized by Zipf's law, given the

[1] We will use the term *city* for all observations in our data set, while keeping in mind that there are some agglomerations that do not reflect one but several cities in the administrative sense, such as the Pearl River Delta in China.

appropriate threshold is applied. (ii) The city size distribution of light is more unequal than that of population and Zipf's law in light can be rejected for most countries. The distributional difference between light and population is particularly pronounced in African countries. (iii) Deviations from Zipf's law can mostly be explained by the top end of the distribution, with primary cities being disproportionately populous and bright. In the biggest cities, economic activity is particularly concentrated and light is less sensitive to population changes. (iv) Decomposing city size into density and area, our results are mostly driven by the area effect of the largest cities. This suggests that the agglomeration effects of scale and market access dominate; reaping also human capital interaction effects requires a good inner-city infrastructure. (v) The observed cross-country variation in the city size distribution can mostly be accounted for by historical factors, which lends weight to the time of development hypothesis by Henderson et al. (2018). Despite relatively little variation over the last decades, our results on higher growth rates in primary cities suggest that the city size distribution of several conntries might become more unequal in the future.

The remainder of this paper is structured as follows: In Section 2 we link our paper to the literature on the size distribution of cities and urban primacy. In Section 3, we describe our city identification scheme, the resulting data set of city size in terms of light and population, as well as the econometric estimation approach. Section 4 contains the empirical estimation results on Zipf's law and their implications, while Section 5 investigates the determinants behind the cross-country variation in the city size distribution and gives an outlook. Section 6 concludes. An Online Appendix contains the accompanying material, such as supplementary information on the data set, a simulation exercise, detailed results for each country as well as numerous robustness checks.

2 Related literature

As focal points of economic and social activity, cities and their size have attracted researchers' interest for a long time. It was first suggested by Auerbach (1913) that the size distribution of cities in terms of population follows a Pareto distribution

$$N_y = A \cdot y^{-\alpha} \tag{1}$$

with N_y as the number of cities larger than population size y and shape parameter α. The special case of $\alpha = 1$, and constant A equal to the size of the largest city, is referred to as Zipf's Law (Zipf, 1949).

Various seminal papers have focused on the underlying theoretical processes. From the homogeneity of cities' growth processes with the same rate and variance independent of their size - so-called Gibrat's (1931) Law - one can derive that the entire distribution

of all cities and towns should be lognormal (Eeckhout, 2004). But augmenting this homogeneous growth process with a lower bound of city size and some shocks yields a distribution that is Pareto at the top (Gabaix, 1999). A Pareto distribution in city size also emerges from other theoretical consideration, for example the combination of cities' agglomeration and congestion effects (Rossi-Hansberg and Wright, 2007), the interplay between industry-specific shocks and firms' location decisions (Duranton, 2007), or a combination of sorting of individuals, productivity of firms as well as agglomeration effects (Behrens et al., 2014).

Most of the empirical papers on the topic focus on the U.S. and provide supportive evidence of a Zipf distribution for the biggest U.S. cities (see e.g. Rozenfeld et al., 2011, Fazio and Modica, 2015); for example Gabaix (2016) obtains an alpha coefficient estimate of 1.03. The implication of Zipf's law that a city's rank is approximately inversely proportional to its size is neatly illustrated in the U.S., where the biggest city (New York) has twice the population of the second-ranked city (Los Angeles) and three times the population of the third-ranked city (Chicago).[2]

This paper sets itself apart from these works by its global focus, investigating Zipf's law in countries around the world. It follows earlier cross-country studies which were limited in their data availability and comparability of city definitions. For example, Rosen and Resnick (1980) use census data from 44 countries from the 1970s; Soo (2005) works with the www.citypopulation.de data based on national statistical offices from 73 countries. But how cities are defined in terms of minimum size and administrative borders varies considerably across countries. It is therefore not clear to what extent the variability of alpha estimates across the sample - Rosen and Resnick's (1980) Pareto alphas range from 0.809 (Morocco) to 1.963 (Australia) - is due to systematic deviations from Zipf's law or how much may be due to measurement issues. With our globally consistent city identification scheme and our threshold discussion of the Pareto tail, we provide a thorough treatment of these topics which was lacking until now. Also, Brakman et al. (1999) have already remarked that there is more evidence in favor of Zipf's Law when agglomerations rather than core cities are considered as these tend to extend around the largest metropolises. With geo-spatial data, we are able to investigate this argument thoroughly. Although geo-spatial data are now widely used in regional and development economics, there only seems to be one other Zipf-related paper which measures the urban extent based on geo-spatial data: Small et al. (2011) look at the size distribution of the world's largest metropolises independent of countries.[3]

[2]Note that even if Zipf's Law holds exactly, this rank-size association remains an approximation. It is typically imprecise for the lower-ranked cities and provides a more adequate representation of the many higher-ranked cities (Gabaix and Ioannides, 2004).

[3]Their supportive evidence of Zipf's law in a sample of the world's largest cities irrespective of countries is intriguing. Yet, for policymakers, the city size distribution at the national level is arguably more relevant, and in the absence of a friction-free movement of people and capital across borders, the theoretical processes underlying city growth are more likely to hold at the national than at the global

To our knowledge, our paper is also the first to look at the size distribution of cities in terms of economic activity proxied for by light, in addition to population. Analyzing the drivers of the geographical distribution of economic activity within countries with nighttime lights is an active research area (see e.g. Henderson et al., 2018, Lessmann and Seidel, 2017), but so far a connection to Zipf's law is missing. We add up nighttime lights analogously to population within the identified city boundaries and study differences in the city size distributions of light and population. Nighttime lights capture consumption, investment, government spending, in particular in public infrastructure (Henderson et al., 2012, Elvidge et al., 2014), but the relation between light and population may be different in cities of different size. For example, primary cities might own their outsized role to both agglomeration effects (Rosenthal and Strange, 2004, Moretti, 2004) and disproportional public resources (Ades and Glaeser, 1995). Our comparison of population and light across the whole size distribution of cities directly contributes to this debate. One reason why nighttime lights have so far not been used widely for the study of largest cities is also the top-coding problem: The classical DMSP-OLS satellites fail to capture the brightness of big cities due to sensor saturation, thereby underestimating their economic size. In addition to the original data, we therefore work with the top-coding corrected lights data by Bluhm and Krause (2018) to contour this problem and measure the city size distribution of lights in an unbiased way.

By investigating Zipf's law in each country and its patterns of deviations and determinants in different world regions, our work is also related to continent-specific studies of cities. In particular, Africa with its rapidly increasing urbanization rates, high primacy ratios and often insufficient urban infrastructure is the topic of numerous papers (such as Jedwab and Vollrath, 2018, Castells-Quintana, 2017, Christiaensen and Todo, 2014, Barrios et al., 2006). It has been argued that the optimal city size distribution might depend on a country's level of development: Urban concentration is thought to be helpful for poor countries with weak physical and human capital resources as well as high transport costs, while a more balanced city size distribution should be beneficial for more advanced economies (Krugman, 1991, Henderson, 2003). Our worldwide data set allows to investigate to what extent deviations from Zipf's law follow this argument, as well as to give an outlook of the future city size distribution and its consequences.

level (Cristelli et al., 2012).

3 Data and methodology

3.1 City identification and city size measurement

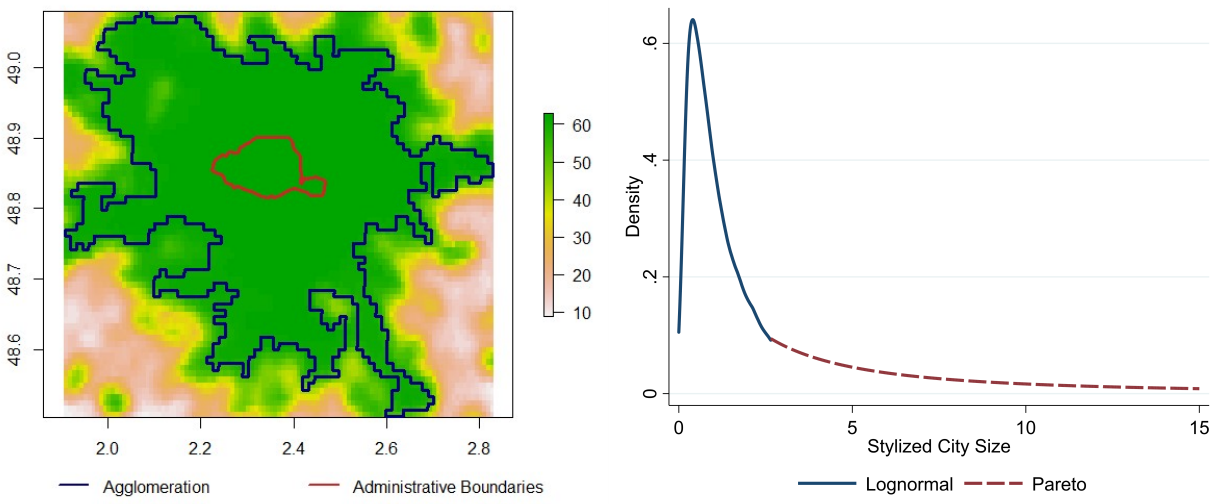
The key contribution of this paper is to conduct a global cross-country study of the size distribution of cities using a consistent city identification scheme.[4] We employ the European Commission's Global Human Settlement Layers (GHSL) to identify urban settlements hosting more than 50,000 residents in 2015. The lower bound of 50,000 inhabitants has been found to be reasonable for agglomeration effects to take place in countries around the world (World Bank, 2009). To determine the de facto spatial extent of a city, the GHSL data combine Landsat satellite imagery on built up area with census information (Pesaresi and Freire, 2016). Unlike other possible identification methods derived from the nighttime lights data itself (Small et al., 2011), this method does not suffer from overglow and identification problems at the bottom of the distribution. The resolution of the data grid is 1 km pixel of the globe. Figure 1a illustrates for the case of Paris that the identified urban boundaries (in blue) match the economic and social reality of a city much better than administrative borders (in red). Technical details on the city identification scheme, including the treatment of border cities and a robustness check with shape files from other years are given in Online Appendix A. In this way, we identify 13,844 cities in 194 countries.

For all the identified cities, we use two measures of size, namely geo-referenced population and the sum of lights within their urban extents.
We take the population data from the same data source as the city shapes, the GHSL. They combine Landsat information on built up area with NASA's Socioeconomic Data and Applications Center's fourth version of the Gridded Population of the World (GPW) to derive a population grid at the same resolution as the city shapes. Hence, we obtain population data for each of our identified cities as a panel for the years 1975, 1990, 2000 and 2015.
In contrast to this, our nighttime light data form a yearly panel from 1992 to 2013. We use the 'stable night light images' collected by the Defense Meteorological Satellite Program's Operational Linescan System (DMSP-OLS) operated by the National Oceanic Administration Agency (NOAA). The satellites monitor light emissions between 8.30 and 10.00 pm local time on a daily basis and the published pictures are yearly averages of light emissions on cloud-free days. Corrected for glare, auroral lights,

[4]For the reasons outlined above, we use a geo-spatial data set of cities based on a comparable definition for our cross-country analysis. But as a direct comparison with the existing literature, we repeat our analysis with the www.citypopulation.de data (Brinkhoff, 2017), which relies on definitions from national statistical agencies. This data set differs from ours in the number of cities included, the available years and the urban extents of the cities (administrative borders rather than geo-data based extent). The results are discussed in Online Appendix E.

Figure 1 – Measuring big cities

(a) Urban Extent of Paris **(b)** Lognormal Body of Cities with Pareto Tail



Notes: Figure 1a shows the administrative boundaries of Paris as well as the boundaries identified by GHSL. Lit areas (stable lights) are shown in the background. Figure 1b is a stylized city size distribution in line with the literature, showing a lognormal body of towns and smaller cities and a Pareto tail of big cities.

forest fire and gas flares, the resulting lights are assumed to be exclusively man-made. Light emissions at the pixel level are measured by a Digital Number (DN) ranging from 0 (dark) to 63 (fully illuminated). This 'stable lights' data set has been used extensively in applications in development and regional economics, see Donaldson and Storeygard (2016) for an overview. However, Bluhm and Krause (2018) point out that the 'stable lights' suffer from top-coding. Due to sensor saturation, the satellite cannot capture the full brightness of the biggest cities in which many pixels reach the end of the scale at 63 DN. As big cities form the focus of our analysis, we use as an additional data source the top-coding corrected nighttime lights provided by Bluhm and Krause (2018). For both the 'stable' and corrected lights separately, we add up the DN's for all the pixels within the city boundaries to obtain the total luminosity of the city.

Table 1 gives an overview of our data set in the year 2000 by presenting summary statistics for all the cities of the world (13,844) as well as those in some selected countries. While the median city has a population of 85,530 inhabitants, the world's largest city has more than 32m inhabitants (the Pearl River delta agglomeration in China covering inter alia Shenzhen and Guangzhou). Comparing, for example, the 323 cities in the US and 514 in Bangladesh, we see that the median cities in both countries are of similar size in terms of population but not in terms of sum of lights (34 DN vs 7295 DN). This underlines the importance of analyzing the city size distribution in each country separately and thus accounting for different development levels and country-specific heterogeneity. The table

also shows that the differences between the 'stable' and the top-coding corrected lights are larger in richer countries - and, for all countries, they are larger in the biggest cities than at the median.

Table 1 – Summary Statistics of the Data Set in the Year 2000

	USA	DEU	CHN	NGA	BGD	World
Number of Cities	323	86	2,266	428	514	13,844
Median City Size						
Population	100,909	105,124	99,455	72,434	98,226	85,530
Stable Light	7,295	4,153	474	53	34	284
Corrected Light	16,485	4,833	474	53	34	284
Maximum City Size						
Population	14,853,624	7,477,014	32,343,639	7,789,496	15,452,476	32,343,639
Stable Light	474,339	299,255	502,338	43,631	58,281	502,338
Corrected Light	1,983,732	379,765	845,757	45,687	62,815	1,983,732
Top City's Name						
Population	New York[a]	*Ruhrgebiet*[b]	Pearl River Delta[c]	Lagos	Dhaka[d]	Pearl River Delta[c]
Stable Light	Los Angeles[e]	*Ruhrgebiet*[b]	Pearl River Delta[c]	Lagos	Dhaka[d]	Pearl River Delta[c]
Corrected Light	Chicago[f]	*Ruhrgebiet*[b]	Pearl River Delta[c]	Lagos	Dhaka[d]	Chicago[f]

Notes: The sums of light within the city boundary are measured in DN. a: New York includes Newark, Paterson. b: Ruhrgebiet (Essen, Duisburg, etc.) includes Düsseldorf, Cologne, Bonn. c: Pearl River Delta includes Shenzhen, Guangzhou, Huizhou, Dongguan, Foshan, Jiangmen. d: Dhaka includes Narayanganj. e: Los Angeles includes Pesadena, Irvine, San Bernadino etc. f: Chicago includes Waukegan, Evanston, Gary, Joliet, Aurora, Elgin

3.2 Estimation approach

To test for Zipf's law, we have to limit the analysis to those 103 countries from our data set with a sufficient number of cities, set to 10 here.[5] Within each country, the question remains where the threshold between the lognormal body of towns and smaller cities on the one hand and the Pareto tail of big cities on the other hand should be, see Figure 1b. While our data set only contains cities with at least 50,000 inhabitants, a Chinese city of that size might still belong to the lognormal part of its country's distribution. Ioannides and Skouras (2013) argue that in the U.S. the switch between the two portions of the distribution occurs between 30,000 and 60,000 inhabitants, but a cross-country discussion of this issue is missing so far. In Online Appendix B, we conduct a Monte Carlo simulation which shows the potential distortion of the coefficient estimate when the Pareto estimation is carried out using too low a threshold. Based on this result, a thorough discussion of possible thresholds and a numeric identification algorithm, we follow a twofold strategy in our empirical estimation: In each country, we use (i) all cities in our data set, (ii) only

[5] In the Online Appendix B, we motivate why we set a minimum requirement in the quantity of given cities, showing that Pareto tails cannot be established empirically for tiny countries' fewer cities.

those cities above the median. The later follows from an optimization approach between ensuring a pure Pareto tail in as many countries as possible (and therefore avoiding bias by not setting the threshold too low), while using a large number of city observations per country (and hence ensuring precision through not too high a threshold).[6]

For our actual estimation, we follow the literature in using log-rank regressions. If city size y is Pareto distributed with shape parameter α above the threshold y_c as determined above, we have $\text{rank}(y) \approx N y_c^{\alpha} y^{-\alpha}$, or, in logarithms, $\log \text{rank}(y) - \log N \approx \alpha \log y_c - \alpha \log y$. OLS estimation of this equation underestimates the true coefficients and standard errors in small samples due to the ranking procedure (Gabaix and Ioannides, 2004). However, subtraction of one half from the rank has been shown to improve the estimation of the Pareto alpha (Gabaix and Ibragimov, 2011) so that we will estimate the following log-rank regression by OLS

$$log\Big(rank(y) - \frac{1}{2}\Big) - log(N) = cons. - \alpha \cdot log(y) + \epsilon \tag{2}$$

with the corrected OLS standard errors given as $\sqrt{2/N} \cdot \hat{\alpha}$ by Gabaix and Ibragimov (2011).[7]

4 Results on Zipf's Law around the world

4.1 Results in terms of population and light

Running (2) individually for all countries with at least 10 cities, and measuring city size respectively by population and light, we determine whether or not Zipf's law holds.

Figure 2 shows the densities of the obtained Pareto alpha coefficients in the year 2000. The main insight is that Zipf's law is an appropriate characterization of the city size distribution for many countries in terms of population but not in terms of economic activity as proxied for by light.

The green curve of the population alpha coefficients is centered around 1, and for 77% of countries the coefficients lie within 95% confidence bands around $\alpha = 1$.[8] When using the full distribution of cities above 50,000 inhabitants in each country (Figure 2a), the alpha density is slightly more left-skewed than when only including cities above the

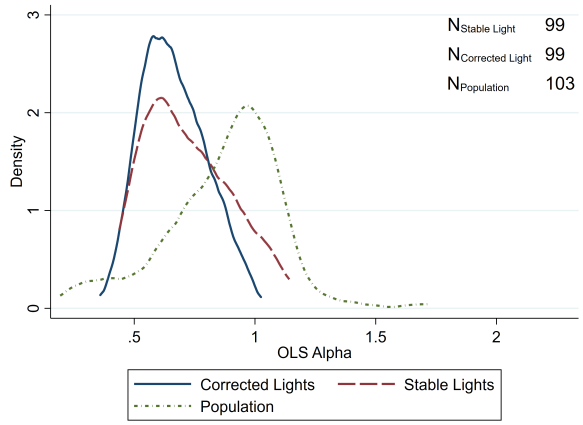
[6] We also test other thresholds. In Online Appendix B we discuss a linear combination of a cutoff that is first horizontal and then increasing in the number of cities per country, analogous to the literature on international poverty lines (see for example Ravallion and Chen, 2011).

[7] An alternative to OLS is the Hill (1975) estimator, which is the maximum likelihood estimator if the data is Pareto distributed but which is less robust to deviations. In Online Appendix D we repeat our analysis with the Hill estimator as a robustness check for our results. They are very similar for most countries; for example, the correlation between the OLS and Hill alpha estimates is 0.813 for 'stable lights' in the above median sample.

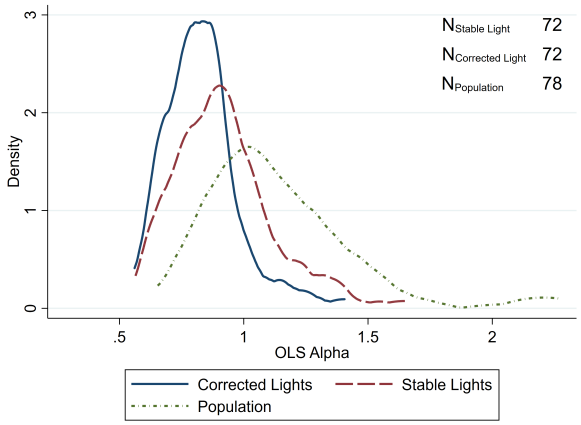
[8] In Online Appendix C we list Pareto alpha coefficients for all countries with their standard errors.

Figure 2 – Density Plots of Countries' Estimated Pareto Alphas in the Year 2000

(a) Full Distribution of Cities



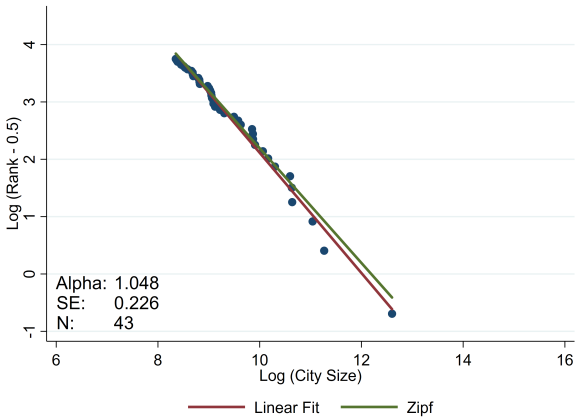
(b) Cities Above Median



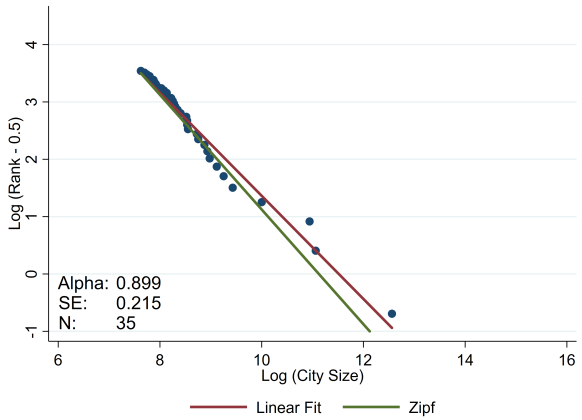
Notes: The estimated Pareto alphas of all countries with more than 10 cities. Figure 2a estimates (2) for the whole distribution of countries, Figure 2b uses only cities above the median. Light is based on satellite F15.

median (Figure 2b). As the former potentially features observations from the lognormal body, inequality in the distribution is slightly overstated. Nevertheless, the mode at $\alpha = 1$ is a robust characteristic. Our result that an approximately equal share of countries have coefficient estimates slightly smaller and larger than 1 differs from the previous literature, which typically finds a dominating share of countries with larger Pareto alphas (Rosen and Resnick, 1980, Soo, 2005). Apart from the larger number of countries in our data set and the consistent city identification scheme, this may be due to the different urban extent we measure. Administrative boundaries fail to capture the economic and social extent of larger cities and suggest a more egalitarian city size distribution than in our results.

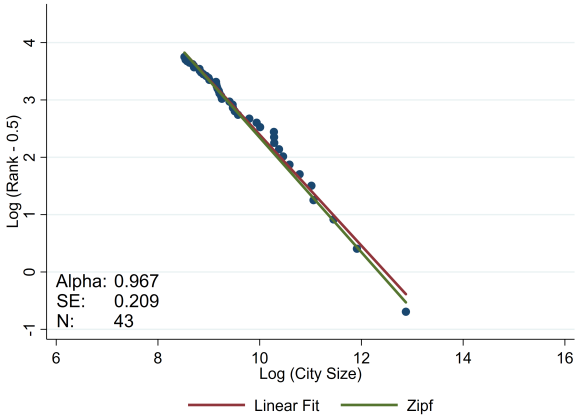
Turning towards light, Figure 2 clearly shows that the Pareto alphas are more unequally distributed than for population. Their distributions are centered around coefficient estimates smaller than 1. Only 52% of 'stable light' alpha estimates, and 39% of 'corrected light' estimates, lie within a 95% confidence interval around $\alpha = 1$. In line with our expectations, top-coding corrected lights are even more inegalitarian than the 'stable lights' as they capture the full brightness of the largest cities. This is particularly evident when only cities above the median are considered, of which large metropolises command a larger share (Figure 2b).



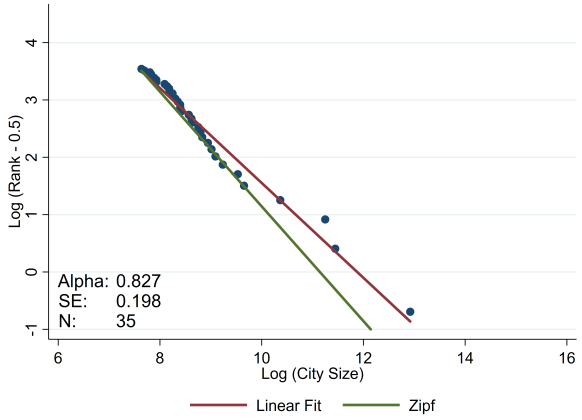
(a) Stable Light: Germany



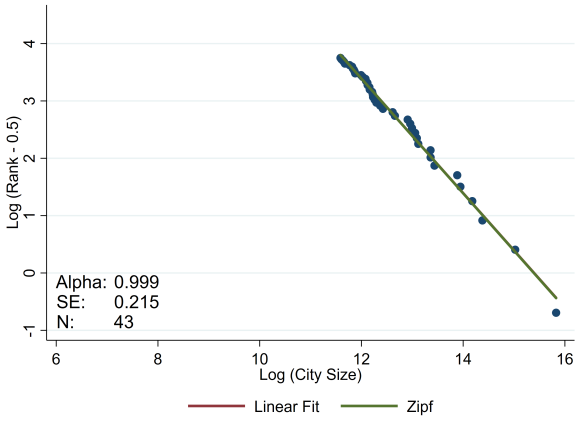
(b) Stable Light: South Africa



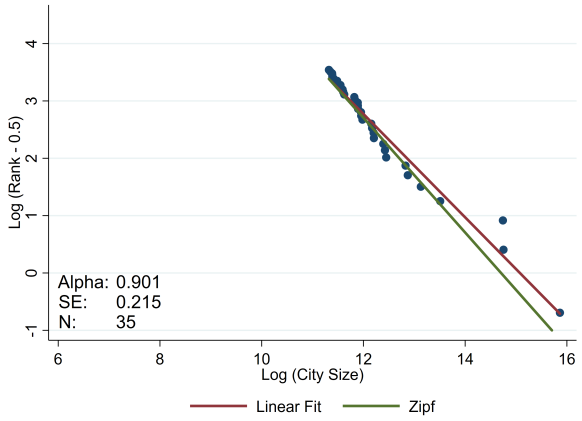
(c) Corrected Light: Germany



(d) Corrected Light: South Africa



(e) Population: Germany



(f) Population: South Africa

Figure 3 – Zipf Plots for Germany and South Africa (Above Median Setting, Year 2000, Satellite F15)

There is, however, evidence that the size distribution of cities in terms of light follows a Pareto distribution; it is just not the special Zipf case of $\alpha = 1$ for many countries. Figure 3 uses the example of the two otherwise very different countries of Germany and South Africa, which both display remarkably linear Zipf plots in terms of

population, 'stable' and corrected light. It is the slopes of the OLS best-fit lines that differ and deviate slightly more strongly from the green Zipf line ($\alpha = 1$) for light than for population.

Exploiting the global nature of our data set, we assemble the population and 'stable light' alpha coefficients for available countries in the scatter plot in Figure 4. The clustering of countries in the lower left and upper right quadrants shows that those whose city size distribution is slightly more (less) egalitarian in terms of population is also more (less) egalitarian in terms of light. Yet, the sizable number of countries in the upper-left quadrant of more equality in the population size distribution and more inequality in light suggests a more nuanced relation. Clearly, population increases do not always translate one to one into increases in economic activity as proxied for by light, but the effects may differ across the city size distribution, as we will examine in the following.

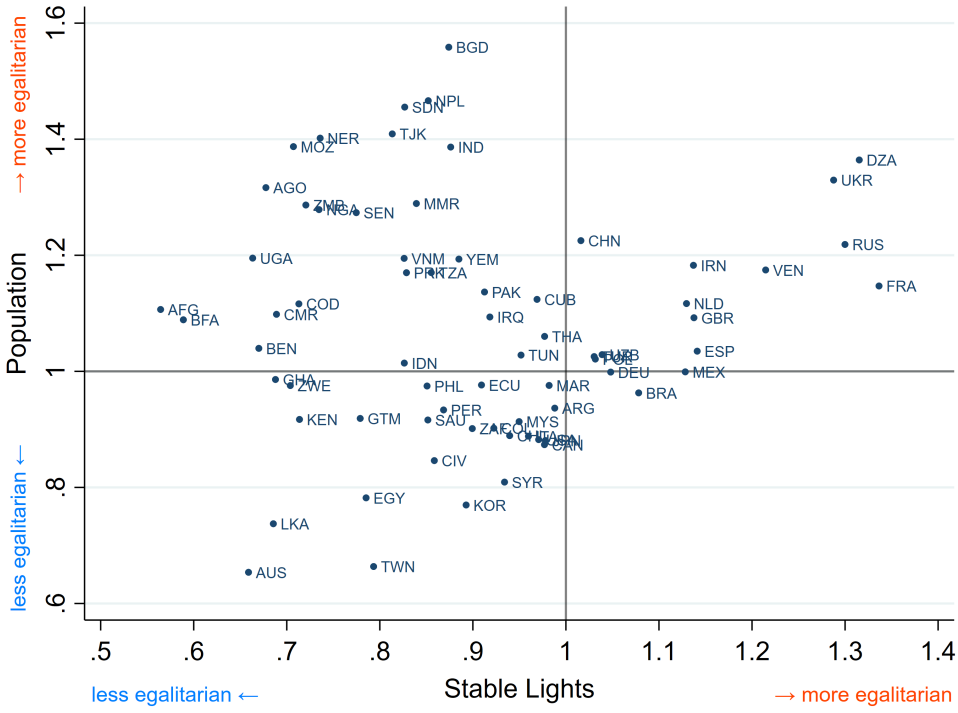


Figure 4 – Scatter Plot of Population and Stable Light Pareto Alpha Coefficients (Above Median Setting, Omits Outliers: Kazakhstan, Romania, Ethiopia)

Furthermore, we note patterns of cross-country heterogeneity in the estimated coefficients: Table 2 shows that on average, population and light are distributed more equally in Europe than in Africa, Asia, and the Americas. In terms of population, this confirms the cross-country pattern observed by Soo (2005). Going beyond that, our results show that (i) this holds for light as well, even as light is more unequally distributed than population, (ii) working with the full distribution of cities rather than those above the median yields coefficients of a larger magnitude but with similar cross-

Table 2 – Summary Statistics for Pareto Alphas (Year 2000)

			Africa	Americas	Asia	Europe	World
Stable	Full	Mean	0.626	0.706	0.643	0.955	0.709
Light	Dist.	(SD)	(0.126)	(0.153)	(0.125)	(0.130)	(0.178)
		N	29	17	32	17	99
	Above	Mean	0.776	0.964	0.886	1.145	0.904
	Median	(SD)	(0.156)	(0.118)	(0.161)	(0.129)	(0.204)
		N	22	12	26	9	72
Corrected	Full	Mean	0.609	0.624	0.607	0.829	0.654
Light	Dist.	(SD)	(0.102)	(0.096)	(0.104)	(0.106)	(0.135)
		N	29	17	32	17	99
	Above	Mean	0.746	0.810	0.816	0.979	0.817
	Median	(SD)	(0.113)	(0.082)	(0.128)	(0.133)	(0.152)
		N	22	12	26	9	72
Population	Full	Mean	0.811	0.891	0.875	0.989	0.868
	Dist.	(SD)	(0.263)	(0.117)	(0.271)	(0.085)	(0.244)
		N	31	17	32	17	103
	Above	Mean	1.195	0.955	1.152	1.094	1.137
	Median	(SD)	(0.294)	(0.098)	(0.331)	(0.130)	(0.308)
		N	25	13	27	9	78

Note: The table presents the summary statistics of OLS alpha estimates by continent. Luminosity in the year 2000 is here defined as the average of the two DN values obtained from the two satellites that were active in that year (F14 and F15). Asia includes Oceania.

continent patterns, (iii) the distributional differences between light and population are not as pronounced in Europe as on other continents. Across African countries, the mean alpha estimate is 0.746 for the above-median corrected light distribution and 0.609 for the full distribution, indicating strong deviations from Zipf's law.

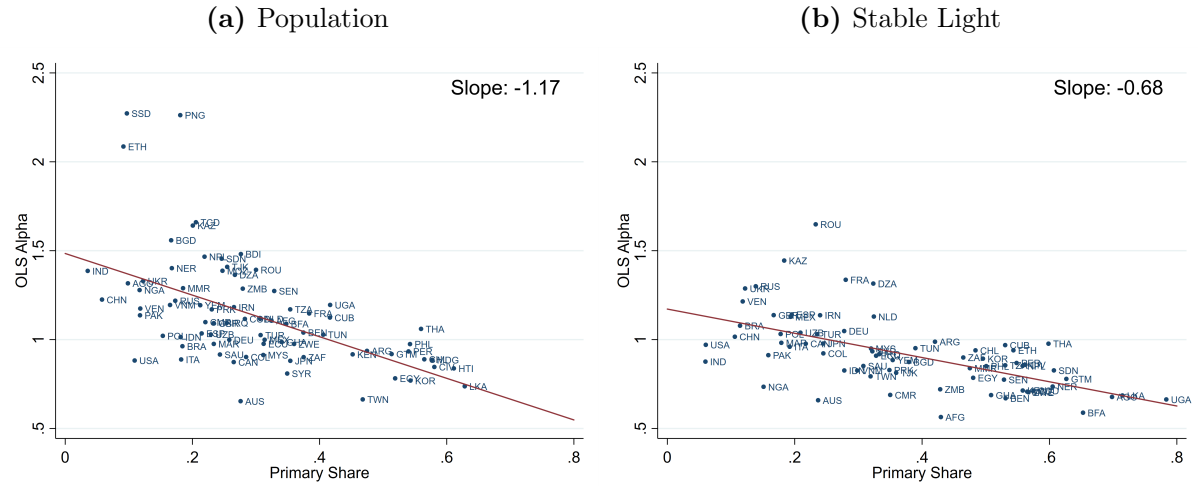
4.2 Explaining patterns along the city size distribution in population and light

Let us now examine which part of the distribution is driving the deviations from Zipf's law. We compute primary shares as the proportion of the distribution's total population or light that is accounted for by the largest city.

Figure 5 show that higher primary shares are strongly correlated with smaller Pareto alphas, both in terms of population and light. The best-fit line for light is flatter, so that a smaller Pareto alpha is associated with an even larger primacy share for light. We see that, overall, countries with a more unequal city size distribution exhibit an undue concentration of population and economic activity in primary cities.

How is this concentration in the primary cities related to the more unequal distribution of light compared to population? Emitted light per capita obviously is not constant throughout the distribution and the process between population, light and economic activity differs between cities of varying size. We therefore run regressions of the logarithm of total light emissions in city i in country j and year t on the logarithm of population,

Figure 5 – Primary Share and Estimated Pareto Alpha in the Year 2000



Notes: The scatter plots show the relation for the estimated Pareto alphas (above median setting, all countries with more than 10 cities) with the primary share, defined as the share of the distribution's population or total luminosity that is due to the largest city. The stable light alphas are based on the satellite F15.

including country- and year-fixed effects, δ_j and γ_t:

$$log(Light_{ijt}) = \beta_0 + \beta_1 \log(Population_{ijt}) + \delta_j + \gamma_t + \varepsilon_{ijt} \tag{3}$$

According to the estimates displayed in the first two columns of Table 3 a one percent surge in population is associated with an increase in light that is significantly more than a one percent.[9] The stronger impact on corrected lights arises from the fact that the total luminosity of larger cities is more severely underestimated by top-coding than it is the case for smaller cities. To see whether the elasticity is different for primary cities compared to the rest of the distribution we add $Primacy_{ijt}$, a dummy that equals one for the largest city in terms of population in a country in a certain year, and its interaction with $log(Population_{ijt})$ to the model. We can see that primary cities are ceteris paribus on average brighter than other cities, irrespective of population, but their light emissions respond less strongly to population level variation.[10] The interpretation of these findings is in line with, inter alia Ades and Glaeser (1995), Henderson and Wang (2007) and Gollin et al. (2017), who find that disproportionately many resources are pooled into primary cities, and that primary cities play an outsized political, social and economic role. Potentially because they are already bright, their emitted luminosity responds less strongly to population growth than that of smaller cities. Regressions depicted in the last two columns of Table 3 show that this effect

[9] As robustness checks we repeat the estimation in a Seemingly Unrelated Regression Equations (SURE) framework, which leads to nearly identical results.

[10] When clustering standard errors at the city rather than country level, the coefficient of the interaction term coefficient is strongly significant for both light measures.

is indeed specific to primary cities and not just a big city effect: Replacing the $Primacy_{ijt}$ dummy with $TopTen_{ijt}$, which equals one for the ten most populated cities per country and year, yields insignificant results for the dummy and the interaction term.

Table 3 – Light-Population Elasticities

Dependent Variable: log(Light)						
	Stable	Corrected	Stable	Corrected	Stable	Corrected
	(1)	(2)	(3)	(4)	(5)	(6)
log(Pop.)	1.098***	1.171***	1.094***	1.160***	1.080***	1.127***
	(0.045)	(0.050)	(0.050)	(0.056)	(0.068)	(0.076)
Primacy			2.419***	1.874**		
			(0.927)	(0.943)		
log(Pop.) x Primacy			-0.155**	-0.106		
			(0.068)	(0.072)		
TopTen					0.865	-0.122
					(0.926)	(1.014)
log(Pop.) x TopTen					-0.048	0.036
					(0.079)	(0.087)
Constant	-8.273***	-9.120***	-8.239***	-9.016***	-8.139***	-8.718***
	(0.600)	(0.662)	(0.659)	(0.728)	(0.839)	(0.931)
Country F. E.	Yes	Yes	Yes	Yes	Yes	Yes
Year F. E.	Yes	Yes	Yes	Yes	Yes	Yes
N	254,689	254,689	254,689	254,689	254,689	254,689
adj. R^2	0.730	0.755	0.730	0.755	0.731	0.756

Note: Standard errors clustered at the country level in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Primacy is a dummy that equals one for the largest city in terms of population in the respective year and country. TopTen is a dummy that equals one for the 10 largest city in terms of population in the respective year and country.

That primary cities in countries around the world behave so differently from their second- and third-largest peers is worth a closer look. With their outsized position, they benefit from agglomeration effects particularly strongly: They are home to many firms and skilled workers, are often the center of political power, and offer a vibrant social and cultural life. This adds to their productivity, which, in turn makes the city more attractive to newcomers, further propelling its role. The literature on agglomeration effects and positive externalities splits these factors into a scaling and a density effect: Larger cities offer larger markets with more opportunities for both firms and workers (Krugman, 1991, Fujita et al., 1999), while denser cities facilitate the interactions of individuals (Bettencourt, 2013) and make them benefit from human capital externalities (Moretti, 2004, Diamond, 2016).

We investigate the relative size of these effects using our data set. Let us keep in mind that city size - be it in terms of population or light - of city i at time t is the product of area and density

$$Size_{it} = Area_{it} \cdot Density_{it}. \tag{4}$$

This means that primary cities can be bigger than secondary cities for two reasons: either they extend to a larger area or they have a higher density of population (or light). Obviously, we often have a combination of both contributors, but they are not equally important. We divide city size, area and density for the primary and secondary city of each country and summarize the mean values across measures and continents in Table 4.[11]

Table 4 – Comparing Primary and Secondary Cities (Countries with 10 or more cities), Years 2013/15

		World	Africa	Americas	Asia	Europe
Population	Size	4.322	4.661	5.011	4.192	2.871
		(4.032)	(2.924)	(3.545)	(5.456)	(2.216)
	Density	1.324	1.118	1.459	1.547	1.082
		(0.905)	(0.691)	(0.544)	(1.255)	(0.377)
	Area	4.241	5.651	3.849	3.559	2.854
		(4.021)	(4.764)	(3.078)	(3.944)	(1.914)
Stable Light	Size	4.666	6.819	4.095	3.555	2.704
		(4.561)	(6.195)	(3.201)	(3.066)	(1.665)
	Density	1.255	1.333	1.099	1.335	1.021
		(0.590)	(0.478)	(0.261)	(0.833)	(0.121)
	Area	3.792	4.930	3.604	3.181	2.674
		(3.119)	(3.707)	(2.387)	(2.994)	(1.604)
Corrected Light	Size	5.507	7.904	5.462	4.240	2.642
		(5.453)	(6.819)	(4.677)	(4.363)	(1.574)
	Density	1.530	1.522	1.472	1.703	1.141
		(0.941)	(0.526)	(0.791)	(1.298)	(0.789)
	Area	3.865	4.950	3.642	3.314	2.768
		(3.163)	(3.691)	(2.465)	(3.139)	(1.641)

The values are computed as $\frac{1}{N}\sum_{i=1}^{N}\frac{PrimaryCitySize_i}{SecondaryCitySize_i}$, $\frac{1}{N}\sum_{i=1}^{N}\frac{PrimaryCityDensity_i}{SecondaryCityDensity_i}$, $\frac{1}{N}\sum_{i=1}^{N}\frac{PrimaryCityArea_i}{SecondaryCityArea_i}$ with country i and N as the total number of countries on the respective continent. The respective standard deviations are denoted in parentheses. Asia includes Oceania.

Several observations can be made: (i) Across the world, a country's primary city is on average 4.3 times as populous and 4.6-5.5 times as bright as the city at rank two. These are sizable proportions, given that the Zipf's Law predicts a factor of two, albeit with large confidence bands (Gabaix and Ioannides, 2004). We conclude that primary cities are disproportionately large, underlining their special role (Ades and Glaeser, 1995, Fetzer et al., 2016, Storeygard, 2016). (ii) In Africa, primary cities are more than 4 times as populous and about 7-8 times as bright as secondary cities, a higher factor than on any other continent. This goes in line with the well-known high primacy share in Africa (Henderson and Wang, 2007, Junius, 1999). Europe, by contrast, has the lowest proportions between primary and secondary cities and they are of a similar magnitude for population and light. (iii) The size differences between primary and secondary cities can mostly be attributed to differences in area rather than density. Across all continents it

[11] These results are for the latest available years, 2013 for light and 2015 for population. They include the countries with 10 or more cities for which we calculate the Pareto alphas. Pooling the data across all available years as well as including countries with fewer than 10 cities does not qualitatively change the results, see Online Appendix F.

holds that the largest cities are only slightly denser - in terms of both light and population - but much more extended than secondary cities. Hence, area is driving inequality at the top of the city size distribution. This suggests that agglomeration effects work through scaling rather than through density, which holds lessons for policy makers. A more extended city brings economies of scale and larger product and labor markets, and it can also ease congestion away from packed areas (Rosenthal and Strange, 2004). By contrast, density is associated with more frequent interactions of its inhabitants and with pooled resources. In the model of Bettencourt (2013), the cost of human interactions increases with the traverse dimension, which is the area of the city. We interpret our results in the way that innercity infrastructure is the key. We see an outsized concentration of economic activity in the largest cities, compared to what is predicted by Zipf's law. But the agglomeration benefits will be limited if cities are so extended and fragmented that they fail to connect their inhabitants. For example, the typically insufficient public infrastructure in many African cities, as inter alia remarked by Castells-Quintana (2017), Lall et al. (2017) and Bluhm and Krause (2018), can be an obstacle on the way of channeling their outsized primary cities into hubs of innovation.

5 Determinants of the city size distribution and time variation

5.1 Determinants of cross-country variation

What are the underlying factors that engender such a city size distribution that is more unequal in some countries than in others? Earlier papers have linked the city size distribution to various institutional and geographic factors, such as total area and population (Rosen and Resnick, 1980), trade openness (Moomaw and Shatter, 1996), infrastructure (Junius, 1999), autocracy (Ades and Glaeser, 1995), government expenditure (Small et al., 2011), fiscal decentralization (Davis and Henderson, 2003), and ethnic fractionalization (Mutlu, 1989).

While all of these factors have been shown to play a role, we are here going to test them in connection with an overarching theory on the evolution of the spatial distribution of economic activity put forward by Henderson et al. (2018). Their key concept is the time of development. According to this argument, the larger spatial equality in economic activity in early developed countries can be traced back to cities' formation in agricultural regions at times when transport costs were still high. This is also in line with Motamed et al. (2014) who use historic population data to show that places with good agricultural quality urbanized earlier. As agglomeration patterns exhibit strong persistence, these structures are still visible today. By contrast, in countries which

developed later, when transport costs were already low, fewer and larger cities were built, often in strategically important coastal locations. While Henderson et al. (2018) provide evidence for this theory in terms of the spatial variation of nighttime lights across the total area of countries, our data set puts us in a position to test it using countries' actual city size distribution. As variables proxying for early and late development, we use education, urbanization and GDP per capita of countries in 1950, just as Henderson et al. (2018).

Table 5 – Time of Development as Determinants of the City Size Distribution

	Stable Lights		Corrected Lights		Population	
	Alpha	Pr. Sh.	Alpha	Pr. Sh.	Alpha	Pr. Sh.
Education in 1950	-0.010	0.009	-0.005	0.002	-0.012	-0.001
	(0.025)	(0.013)	(0.018)	(0.013)	(0.021)	(0.011)
Urbanization in 1950	0.006***	-0.003*	0.004***	-0.001	-0.004*	0.003**
	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
GDP p.c. in 1950	-0.000	-0.000**	-0.000	-0.000**	0.000	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.758***	0.543***	0.740***	0.567***	1.149***	0.341***
	(0.032)	(0.035)	(0.024)	(0.034)	(0.048)	(0.029)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	1358	1742	1358	1742	255	324
adj. R^2	0.217	0.198	0.100	0.150	0.142	0.047

Clustered standard errors in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Education: average years of schooling. Urbanization: percent urbanized. GDP p.c.: dollars (in 2005 PPP). Alpha: OLS Alpha Estimate. Pr. Sh.: Primary Share.

Table 5 shows the result of the panel regression with year fixed effects of

$$ineqcitysize_{it} = \beta_0 + \beta_1 educ1950_i + \beta_2 urban1950_i + \beta_3 gdp1950_i + \gamma_t + \varepsilon_{it}, \qquad (5)$$

where inequality in the city size distribution of country i at time t is expressed either by the Pareto alpha estimate (based on the above-median distribution) or the primary share.[12] According to Table 5 countries which were highly urbanized in 1950, this is early developed, tend to exhibit greater Pareto alpha coefficients in their city size distributions nowadays, both in terms of light and population. The negative coefficient

[12]Running a cross-sectional regression yields very similar results. To increase the sample size we use the panel regression (5) and cluster standard errors at the country level, while taking into account year fixed effects.

estimate of early development on primary cities' relative size supports that notion.

We now investigate the connection between this time of development theory of the city size distribution and other variables, including, but not limiting ourselves to, those that have been used in previous studies. We look at a total of 36 possible explanatory variables ranging from population structure (such as total population, fertility, migration) over physical geography (such as terrain ruggedness, coastal border, continent), institutions (such as time of independence, political rights, fiscal centralization), economic structure (such as agricultural share, energy use, patent applications) and international connectedness (such as exports and interstate war).[13]
Due to the collinearity of this huge number of possible determinants, we employ a model selection approach: We regress the country's Pareto alpha coefficient on up to 7 out of the 36 determinants X_{it} at a time, including year fixed effects:

$$\hat{\alpha}_{it} = \beta_0 + \sum_{i=1}^{I} \beta_i X_{it} + \delta_t + \varepsilon_{it} \qquad with \quad I \in \{1, ..., 7\} \tag{6}$$

This amounts to 10,739,175 regressions to compare, out of which we select the best ones based on AIC and BIC. Table 6 shows the coefficient estimates of the determinants included in the selected models to explain the size distribution of, respectively, 'stable' light, corrected light, and population. Even using such a purely algorithmic approach, we see the importance of historic variables, such as the year of independence as well as population in 1400, which are clearly in line with the time of development framework by Henderson et al. (2018). The continent dummies also play a large role in all three selected models, capturing in particular a more egalitarian city size distribution in (early developed) Europe. Also, current urbanization rates predict a higher Pareto alpha. The effect of trade-related variables on city size is ambiguous, in line with the literature (Duranton, 2008, Fujita and Mori, 1996). However, the association of coastal proximity with a more unequal city size distribution can be seen in light of the recent works by Bonfatti and Poelhekke (2017) and Jedwab and Moradi (2016), who argue that outward-oriented colonial infrastructure in many developing countries still has an effect on their distribution of economic activity. Taken together, our results suggest to a much stronger extent than previous Zipf-related studies that historical determinants can explain a lot of the cross-country variation of the city size distribution.

[13] A complete list of the 36 variables including their sources is contained in Appendix G. We do not use transportation variables, such as the road and rail network, which we consider endogeneous to the city size distribution even though they have been used in some other studies, such as Rosen and Resnick (1980).

Table 6 – Coefficient Estimates of Selected Models

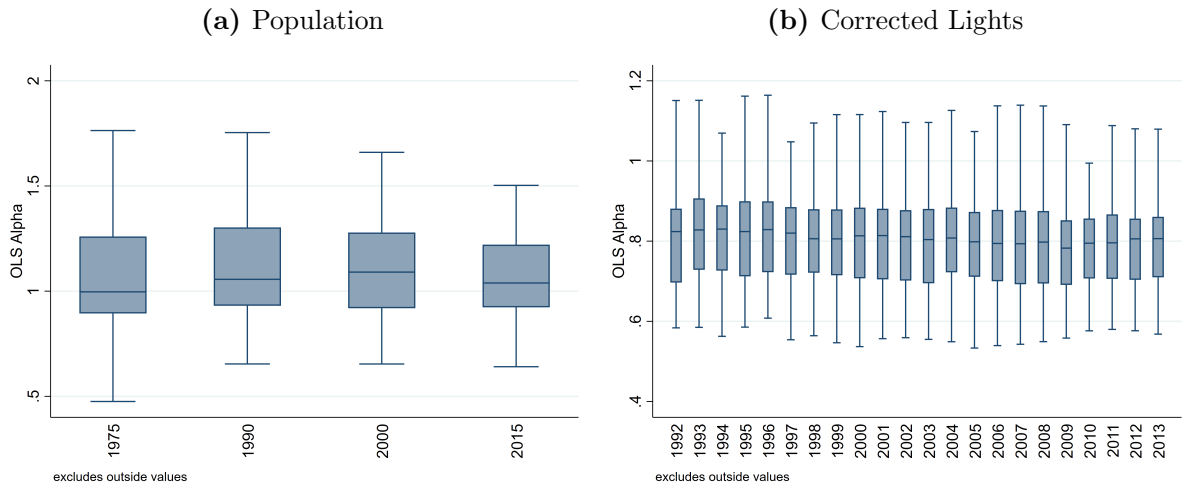
	Stable Light Alpha	Corrected Light Alpha	Population Alpha
Coastal Proximity	-0.002***	-0.001***	
	(0.000)	(0.000)	
Independence			
before 1914	0.049***	0.070***	
	(0.010)	(0.011)	
between 1946 and 1989	0.060***	0.047***	
	(0.009)	(0.009)	
after 1989	0.233***	0.203***	
	(0.012)	(0.011)	
Continent Dummy			
Americas	-0.031***	-0.019*	0.038
	(0.010)	(0.010)	(0.025)
Asia	0.051***	0.084***	0.121***
	(0.009)	(0.008)	(0.027)
Europe	0.166***	0.223***	0.272***
	(0.013)	(0.011)	(0.033)
Oceania	-0.395***	-0.187***	-0.062
	(0.021)	(0.020)	(0.058)
Trade	-0.002***	-0.000***	0.000
	(0.000)	(0.000)	(0.000)
Urbanization	0.004***	0.002***	
	(0.000)	(0.000)	
Fertility	-0.048***		
	(0.002)		
Exports	0.004***		
	(0.001)		
Agriculture		-0.001	
		(0.000)	
Population in 1400		0.000***	0.000***
		(0.000)	(0.000)
Interstate War			0.042
			(0.059)
Patent Applications			-0.000
			(0.000)
GDP p.c.			-0.000***
			(0.000)
Ethnic Fractionalization			0.255***
			(0.038)
Constant	0.894***	0.655***	0.742***
	(0.024)	(0.023)	(0.035)
Satellite-Year F.E.	Yes	Yes	
Year F.E.			Yes
N	2225	2165	123
adj. R^2	0.705	0.482	0.713

Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The base categories for the two categorial variables are non-colonies and Africa respectively. The category "independence 1914 and 1945" drops out due to the lack of observations.

5.2 Changes over time and outlook

What do these results mean in a dynamic perspective? If the city size distribution in terms of light and population can be explained to a large degree by historical factors, it will change only very slowly. It has been shown for particular countries, such as the U.S. (Black and Henderson, 2003), France (Duranton, 2007), and Japan (Eaton and Eckstein, 1997), that the city size distribution has been rather persistent across decades despite structural economic change and city growth. Here we exploit the panel structure of our global data set to verify this hypothesis for countries around the world: Figure 6 shows that across the available years for population (1975-2015) and light (1992-2013), the alpha coefficients and their cross-country distributions exhibit little variation. In terms of population, the range of Pareto alpha coefficients around 1 seems to have narrowed so that some countries are getting a bit closer to Zipf's law from either side. In terms of corrected light, the cross-country distribution of Pareto alphas remains centered around a value of 0.8 for the entire period. Tracing the coefficients for each individual country over time shows similarly little movement, underlining the strong persistence in the city size distribution for most countries.

Figure 6 – Stability of estimated Pareto alpha coefficients over time



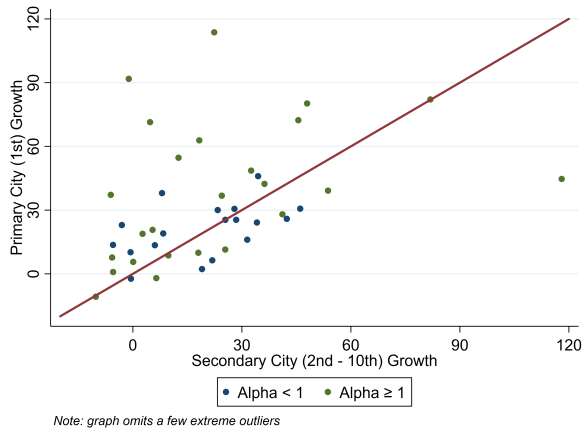
Notes: The estimated Pareto alphas of all countries with more than 10 cities, based on the above-median distribution, are shown in the boxplots. The boxes indicate the 25% to 75% percentiles with the median in between. The range of the plots go up the adjacent values, omitting outliers. In years where more than one satellite per year is available for the lights data, values are averaged.

How can we then expect the future size distribution to look like? Extrapolating the results from our analysis over time would suggest that most countries will remain close to Zipf's law in population and keep their more unequal distribution in terms of light. However, the growth rates of the biggest cities add a twist here. In the long run, cities of different size have to grow at the same average rate for the distribution to remain stable, according to Gibrat's Law. Figure 7 compares the population growth rates of

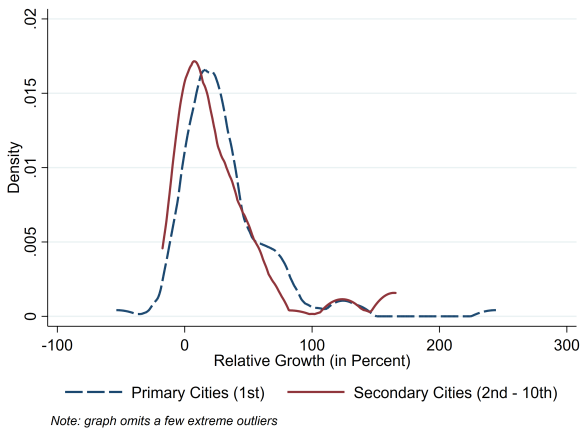
primary and secondary (2nd to 10th largest) cities in each country from 2000 to 2015. In the scatter plot of growth rates (Figure 7a), we see countries distributed along the main diagonal, but a larger number are located above than below it. Hence, in more countries, primary cities have outgrown secondary cities, pointing to a relative consolidation rather than a catch-up. This holds both for countries with both egalitarian and inegalitarian city size distributions, as the different colors of the dots show. Figure 7b compares the cross-country growth rate distributions: The growth rate distribution of primary cities is a bit to the right of that of secondary cities and has its mode at a slightly higher level; the Kolmogorov-Smirnoff test rejects equality of the distributions at the 10% level.

Figure 7 – Population Growth Rates of Primary and Secondary Cities (2000 - 2015)

(a) Growth Rates by Pareto alpha



(b) Growth Rate Densities



This tentative evidence of higher population growth rates of primary cities fits in with other results from the literature. Bluhm and Krause (2018) find that in Sub-Saharan Africa, primary cities are growing significantly faster than secondary cities in terms of light once the top-coding correction is applied to the data. United Nations (2018) forecast that in 2030, there will be 43 mega-cities with more than 10m inhabitants, most of them in developing countries. Given our result that large primary shares drive inequality of the total size distribution of cities, this increasing concentration can have wide-ranging effects. In terms of Zipf's law, we can expect existing deviations not only to persist but to become larger rather than smaller, if the largest cities draw away from the rest. For policymakers, managing the growth of such evolving mega-cities is vital, in particular in developing countries. There is a debate whether overall poverty is lower in larger or smaller cities (Ferre et al., 2012, Christiaensen and Todo, 2014). But growing primary cities will struggle to reap the benefits of agglomeration if (i) living conditions remain bad (Castells-Quintana, 2017, Glaeser, 2014), and (ii) they are disconnected neighborhoods with poor infrastructure (Lall et al., 2017, Bluhm and Krause, 2018).

6 Concluding remarks

We revisit the discussion about Zipf's Law in a cross-country setting by exploiting recent geo-spatial data. We use a consistent city identification scheme, provide a rigorous treatment of the threshold issue, and compare the city size distribution in each country based on both population and light proxying for economic activity. In total, the data set contains 13,844 cities in 194 countries.

The main insight from our analysis is that Zipf's law is an adequate characterization for the size distribution of cities for many, but not for all, countries. Light, however, is typically distributed more unequally, so that Zipf's law does not hold for most countries in terms of light. Such deviations can be explained to a large extent by an undue concentration of resources in the largest cities. We also note that the size effect is mainly driven by area rather than density, so that agglomeration effects are more likely to work through economies of scale and market access rather than human capital interactions.

To explain the cross-country heterogeneity in the size distribution of cities, we make use of recent theories about the time of development. Factors related to economic history turn out to be robust explanatory factors in our model selection procedure. Despite this persistence of the city size distribution, recent growth rates of the largest cities lead us to suggest that we might see a further move away from rather than to Zipf's law in several countries.

Of course, the question remains which distribution of city size is optimal for a given country. Despite the strong theoretical arguments leading to Zipf's law, Henderson (2003) claims that country-specific factors, such as the current development level, might make a different distribution more appropriate. In particular, a stronger concentration of resources at earlier stages of development is said to be beneficial, but later on a more balanced size distribution should emerge (Hansen, 1990, Junius, 1999, Davis and Henderson, 2003). Our results about a relatively persistent size distribution, which might, however, become more unequal rather then equal in the future, suggest that a number of countries are moving in the opposite direction. More research on the interactions between city size, population, economic growth, and welfare is needed here.

References

Ades, A. F. and E. L. Glaeser (1995). Trade and Circuses: Explaining Urban Giants. *Quarterly Journal of Economics 110*(1), 195–227.

Auerbach, F. (1913). Das Gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen 59*, 74–76.

Barrios, S., L. Bertinelli, and E. Strobl (2006). Climatic Change and Rural–Urban Migration: The Case of Sub-Saharan Africa. *Journal of Urban Economics 60*(3), 357–371.

Behrens, K., G. Duranton, and F. Robert-Nicoud (2014). Productive Cities: Sorting, Selection, and Agglomeration. *Journal of Political Economy 122*(3), 507–553.

Bettencourt, L. M. A. (2013). The Origins of Scaling in Cities. *Science 340*(6139), 1438–1441.

Black, D. and J. Henderson (2003). Urban Evolution in the USA. *Journal of Economic Geography 3*(4), 343–372.

Bluhm, R. and M. Krause (2018). Top Lights - Bright Cities and their Contribution to Economic Development. UNU-MERIT Working Paper 2018-041.

Bonfatti, R. and S. Poelhekke (2017). From Mine to Coast: Transport Infrastructure and the Direction of Trade in Developing Countries. *Journal of Development Economics 127*, 91–108.

Brakman, S., H. Garretsen, C. V. Marrewijk, and M. van den Berg (1999). The Return of Zipf: A Further Understanding of the Rank-Size Distribution. *Journal of Regional Science 39*(1), 183–213.

Brinkhoff, T. (2017). City population. http://www.citypopulation.de [Accessed: December 2017 - February 2018].

Castells-Quintana, D. (2017). Malthus Living in a Slum: Urban Concentration, Infrastructure and Economic Growth. *Journal of Urban Economics 98*, 158–173.

Christiaensen, L. and Y. Todo (2014). Poverty Reduction During the Rural-Urban Transformation - The Role of the Missing Middle. *World Development 63*, 43–58.

Cristelli, M., M. Batty, and L. Pietronero (2012). There is More than a Power Law in Zipf. *Scientific Reports 812,2*, 1–7.

Davis, J. and J. Henderson (2003). Evidence on the Political Economy of the Urbanization Process. *Journal of Urban Economics 53*(1), 98–125.

Desmet, K. and E. Rossi-Hansberg (2013). Urban Accounting and Welfare. *American Economic Review 103*(6), 2296–2327.

Diamond, R. (2016). The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980-2000. *American Economic Review 106*(3), 479–524.

Donaldson, D. and A. Storeygard (2016). The View from Above: Applications of Satellite data in Economics. *Journal of Economic Perspectives 30*(4), 171–198.

Duranton, G. (2007). Urban Evolutions: The Fast, the Slow, and the Still. *American Economic Review 97*(1), 197–221.

Duranton, G. (2008). Viewpoint: From Cities to Productivity and Growth in Developing Countries. *Canadian Journal of Economics 41*(3), 689–736.

Eaton, J. and Z. Eckstein (1997). Cities and growth: Theory and evidence from france and japan. *Regional Science and Urban Economics 27*(4-5), 443–474.

Eeckhout, J. (2004). Gibrat's Law for (All) Cities. *American Economic Review 94*(5), 1429–51.

Elvidge, C. D., F.-C. Hsu, Kimberly, Baugh, and T. Ghosh (2014). National Trends in Satellite-Observed Lighting: 1992-2012. In Q. Weng (Ed.), *Global Urban Monitoring and Assessment through Earth Observation*, Taylor & Francis Series in Remote Sensing Applications, Chapter 6, pp. 97–120. CRC Press.

Fazio, G. and M. Modica (2015). Pareto or Lognormal? Best Fit and Truncation in the Distribution of all Cities. *Journal of Regional Science 55*(5), 736–756.

Ferre, C., F. H. Ferreira, and P. Lanjouw (2012). Is there a metropolitan bias? the relationship between poverty and city size in a selection of developing countries. *The World Bank Economic Review 26*(3), 351–382.

Fetzer, T., V. Henderson, D. Nigmatulina, and A. Shanghavi (2016). What Happens to Cities when Countries Become Democratic? unpublished.

Fujita, M., P. Krugman, and A. Venables (1999). *The Spatial Economy: Cities, Regions and International Trade* . MIT Press Cambridge.

Fujita, M. and T. Mori (1996). The Role of Ports in the Making of Major Cities: Self-Agglomeration

and Hub-Effect. *Journal of Development Economics 49*(1), 93–120.

Gabaix, X. (1999). Zipf's Law for Cities: An Explanation. *Quarterly Journal of Economics 114*(3), 739–767.

Gabaix, X. (2016). Power Laws in Economics: An Introduction. *Journal of Economic Perspectives 30*(1), 185–206.

Gabaix, X. and R. Ibragimov (2011). Rank - 1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents. *Journal of Business and Economics Statistics 29*(1), 24–39.

Gabaix, X. and Y. Ioannides (2004). The Evolution of City Size Distribution. In J. Henderson and J.-F. Thisse (Eds.), *Handbook of Regional and Urban Economics*, Chapter 53, pp. 2341–2378. Elsevier.

Gibrat, R. (1931). *Les Inegalités Économiques: Applications: Aux Inegalités des Richesses, à la Concentration des Entreprises, aux Populations des Villes, aus Statistiques des Familles, etc. D'une Loi Nouvelle, la Loi de l'Effet Proportionel* . Paris: Librairie du Recueil Sirey.

Glaeser, E. L. (2014). A World of Cities: The Causes and Consequences of Urbanization in Poorer Countries. *Journal of the European Economic Association 12*(5), 1154–1199.

Gollin, D., M. Kirchberger, and D. Lagakos (2017). In Search of a Spatial Equilibrium in the Developing World. Working Paper 23916, National Bureau of Economic Research.

Hansen, N. (1990). Impacts of Small- and Intermediate-Sized Cities on Population Distribution: Issues and Responses. *Regional Development Dialogue 11*(1), 60–79.

Henderson, J. (2003). The Urbanization Process and Economic Growth: The So-What Question. *Journal of Economic Growth 8*(1), 47–71.

Henderson, J., T. Squires, A. Storeygard, and D. Weil (2018). The Global Distribution of Economic Activity: Nature, History, and the Role of Trade. *Quarterly Journal of Economics 133*(1), 357–406.

Henderson, J. and H. Wang (2007). Urbanization and City Growth: The Role of Institutions. *Regional Science and Urban Economics 37*(3), 283–313.

Henderson, J. V., A. Storeygard, and D. N. Weil (2012). Measuring Economic Growth from Outer Space. *American Economic Review 102*(2), 994–1028.

Hill, B. (1975). A Simple General Approach to Inference About the Tail of a Distribution. *The Annuals of Statistics 3*(5), 1163–1174.

Ioannides, Y. and S. Skouras (2013). U.S. City Size Distribution: Robustly Pareto, But Only in the Tail. *Journal of Urban Economics 73*(1), 18–29.

Jedwab, R. and A. Moradi (2016). The Permanent Effects of Transportation Revolutions in Poor Countries: Evidence from Africa. *Review of Economics and Statistics 98*(2), 268–284.

Jedwab, R. and D. Vollrath (2018). The Urban Mortality Transition and Poor Country Urbanization. *American Economic Journal: Macroeconomics forthcoming*.

Junius, K. (1999). Primacy and Economic Development: Bell Shaped or Parallel Growth of Cities? *Journal of Economic Development 24*(1), 1–22.

Krugman, P. (1991). Increasing Returns and Economic Geography. *Journal of Political Economy 99*(3), 483–499.

Lall, S. V., J. V. Henderson, and A. J. Venables (2017). *Africa's Cities: Opening Doors to the World*. Washington, DC: World Bank.

Lessmann, C. and A. Seidel (2017). Regional Inequality, Convergence, and its Determinants – A View from Outer Space. *European Economic Review 92*, 110 – 132.

Lipton, M. (1977). *Why Poor People Stay Poor: Urban Bias in World Development*. Cambridge, MA: Harvard University Press.

Moomaw, R. and A. Shatter (1996). Urbanization and Economic Development: A Bias Toward Large Cities? *Journal of Urban Economics 40*(1), 13–37.

Moretti, E. (2004). Human Capital Externalities in Cities. In J. Henderson and J. Thisse (Eds.), *Handbook of Regional and Urban Economics*, Volume 4, pp. 2243–2291. Elsevier North Holland.

Motamed, M., R. Florax, and W. Masters (2014). Agriculture, Transportation and the Timing of Urbanization: Global Analysis at the Grid Cell Level. *Journal of Economic Growth 19*(3), 339–368.

Mutlu, S. (1989). Urban Concentration and Primacy Revisited: An analysis and Some Policy Conclusions. *Economic Development and Cultural Change 37*(3), 611–639.

Pesaresi, M. and S. Freire (2016). GHS Settlement grid following the REGIO model 2014 in application to GHSL Landsat and CIESIN GPW v4-multitemporal (1975-1990-2000-2015). Technical report, European Commission, Joint Research Centre (JRC).

Ravallion, M. and S. Chen (2011). Weakly Relative Poverty. *The Review of Economics and*

Statistics 93(4), 1251–1261.

Rosen, K. and M. Resnick (1980). The Size Distribution of Cities - An Examination of the Pareto Law and Primacy. *Journal of Urban Economics 8*(2), 165–186.

Rosenthal, S. and W. Strange (2004). Evidence on the Nature and Sources of Agglomeration Economics. In J. Henderson and J. Thisse (Eds.), *Handbook of Regional and Urban Economics*, Volume 4, pp. 2119–2171. Elsevier North Holland.

Rossi-Hansberg, E. and M. Wright (2007). Urban Structure and Growth. *Review of Economic Studies 74*(2), 597–624.

Rozenfeld, H., D. Rybski, X. Gabaix, and H. Makse (2011). The Area and Population of Cities: New Insights from a Different Perspective on Cities. *American Economic Review 101*(5), 2205–2225.

Small, C., C. Elvidge, D. Balk, and M. Montgomery (2011). Spatial Scaling of Stable Night Lights. *Remote Sensing of Environment 115*(2), 269–280.

Soo, K. (2005). Zipf's Law for Cities: A Cross-Country Investigation. *Regional Science and Urban Economics 35*(3), 239–263.

Storeygard, A. (2016). Farther on Down the Road: Transport Costs, Trade and Urban Growth in Sub-Saharan Africa. *Review of Economic Studies 83*(3), 1263–1295.

United Nations (2018). *World Urban Prospects, the 2018 Revision.* United Nations New York.

World Bank (2009). World Development Report 2009 : Reshaping Economic Geography. World bank.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort.* Addison-Wesley Press.

Online appendix

A Data set construction

This section contains a detailed description of the construction of our data set, including the treatment of border cities and the robustness of our procedure to different shapefiles.

We identify cities based on information provided by the Global Human Settlement Layers. The employed geo-spatial data of a 1 km resolution divide the globe into rural areas, urban clusters and urban centers. The classification of each cell relies on census data and satellite imagery of built-up area. The spatial extent of an urban area in the GHSL data is unconstrained by administrative boundaries such as city, region or country borders. We choose their urban centers as the spatial extent of our cities. These are defined as contiguous cells hosting at least 50,000 inhabitants with a minimum density of 1,500 people per km^2 or a built-up density above 50 percent. The cutoff at 50,000 inhabitants matches exactly the threshold choice recommended by the World Bank (2009). The geo-spatial data provides us with the shape and location of overall 13,844 urban centers which we call cities or agglomerations in a total of 194 countries. All people (or light emissions) aggregated within each of those areas is what we call city size.[1] Not being restricted by administrative city and region borders allows us to measure the contiguous settlements within which agglomeration economies and congestion costs come into play. This means that some what we identify as cities in our data set consist of several cities in the administrative sense. This is illustrated by Figure A-1, a picture of the Ruhrgebiet region in Germany. Although the agglomeration houses a number of different administrative cities (boundaries in red), it is de facto one economic and social contiguum, as the luminosity map indicates. In our data set, it features as one city within the blue boundaries.

[1] Nighttime light images consist of 30 by 30 arc seconds pixel of the globe (about 0.86 square kilometers at the equator). The pixels are thus smaller than GHSL pixels. This does not impede the results since aggregation happens based on GHSL agglomerations derived from the larger GHSL pixels across all data sets.

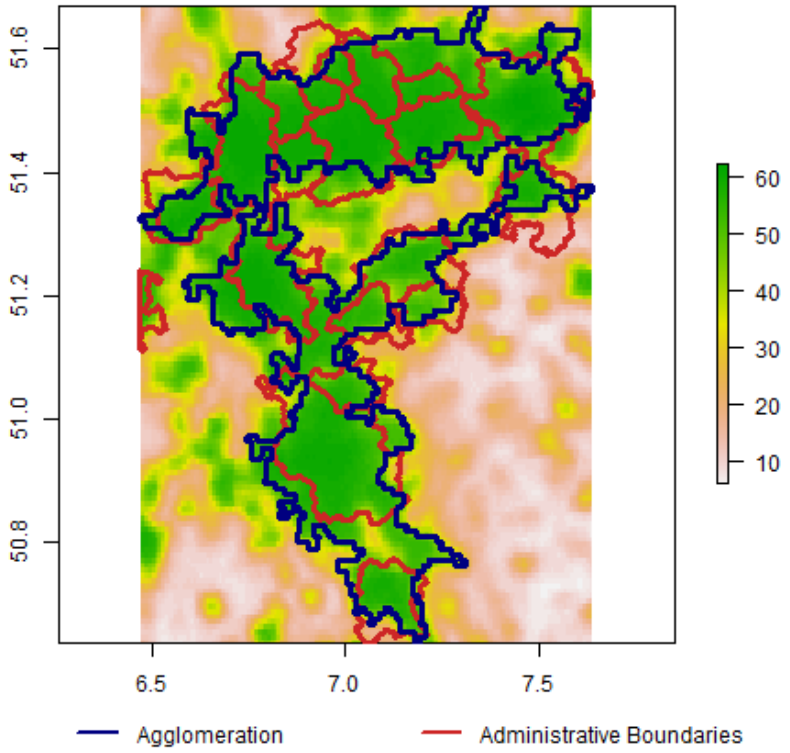
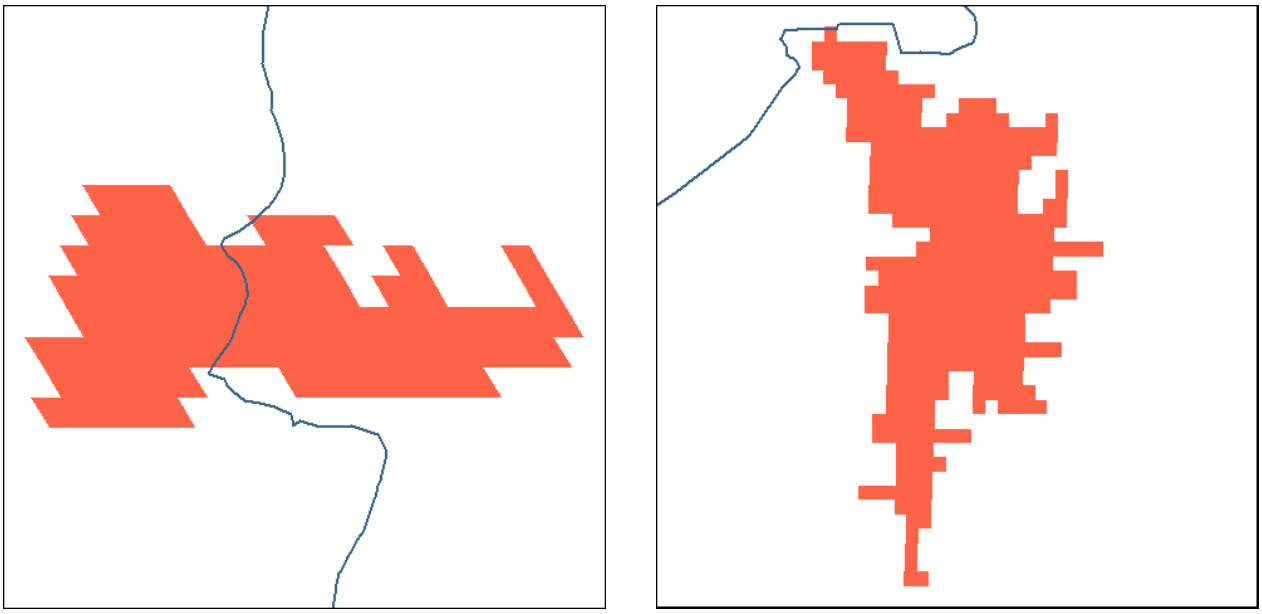


Figure A-1 – The Largest German Agglomeration (Year 2000, Satellite F15, Administrative Boundaries of Urban Municipalities Only)

However, this identification scheme poses a challenge at country borders. Dissecting a city and assigning its fractions to the respective countries would on the one hand violate the lower bound of 50,000 inhabitants and on the other hand underestimate the full area relevant for agglomeration effects. Omitting those border cities biases countries' city size distributions, especially when one of the major metropolises is affected. Assigning the affected cities to all bordering countries at the same time would assign cities to too many countries.[2] In cases with just a few houses on the other side, attributing an entire agglomeration to a country for a few houses seems unreasonable. Therefore, our solution is to assign the full city to one of the countries if more than 75 percent of the urban space are located on one side of the border - and to assign it to all bordering countries otherwise. Figure A-2 illustrates the two cases. The agglomeration Niagara Falls (Figure A-2a) covers areas of similar size in the United States and Canada and is assigned to both countries. Antwerp (Figure A-2b) is a city in Belgium. Our data allocates around 0.04 percent of its area in the Netherlands. The border correction algorithm accordingly assigns it to the city size distribution of Belgium only. We have experimented with other ways to solve the border city issue and find it to be the most

[2]In many cases cities overlap into another country just because of the different format of city shapes and country borders. Our settlements are based on a raster of 1 km grid cells. It can happen that one of the city's cells exceeds the country border, denoted by polygons, simply due to the aggregation.

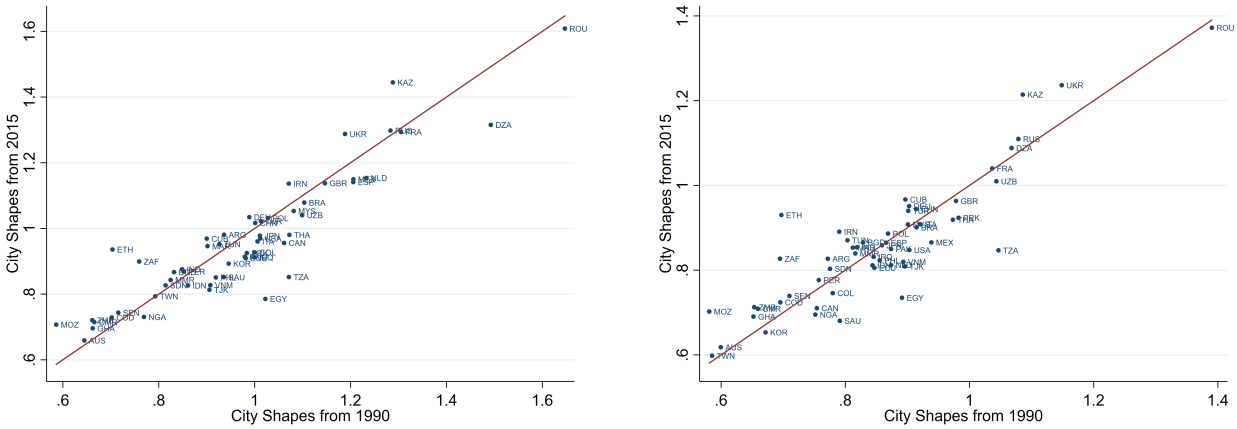
robust. In total, 264 out of the 13,844 cities in the data set are affected by this issue; 106 countries have at least one such border city. 137 out of the 264 cities are assigned to a single country.



(a) Niagara Falls **(b)** Antwerp

Figure A-2 – Border Cities

In our baseline methodology, we use the city shapes derived from observations in 2015 for all years, implicitly assuming the spatial extent of an agglomeration to be time-constant. The advantage is that it allows us to easily identify and track agglomerations over time. However, this approach entails some endogeneity concerns. Using cities' current shapes induces the inclusion of areas that were still sparsely populated in earlier years. In addition, some agglomerations with 50,000 inhabitants in 2015 were considerably smaller in earlier years. To address these concerns, we repeat our analysis with earlier years' agglomeration shapes and check the extent to which the shapes of city size distributions differ. Figure A-3 shows the scatter plots of the Pareto alpha estimates in the year 2000, using repsectively the 1990 and 2015 shape files. Both for stable light (Figure A-3a) and corrected light (Figure A-3b), we see strong correlations between the estimates from different shape files. The estimates for all countries are scattered around the 45 degree line, which is evidence against a systematic bias. We conclude that our city identification procedure is robust to the use of different shape files.



(a) Stable Light **(b)** Corrected Light

Figure A-3 – Comparison of City Shapes (Above Median Setting, Year 2000, Satellite F15)

B Optimal threshold for the Pareto tail

In this section, we provide a rigorous treatment of the threshold selection issue alluded to in the paper. It is a challenge to determine for each country where the cutoff between the lognormal body of towns and smaller cities and the Pareto tail should be, and many previous cross-country papers have largely ignored the issue or used adhoc measures. Here, we (i) conduct a Monte Carlo simulation to illustrate the consequences of using an incorrect threshold, (ii) motivate our threshold choice for the empirical investigation, and (iii) provide results using alternative thresholds.

B.1 Simulation

A crucial prerequisite for empirical tests on Zipf's law in city size distributions is the correct dissection of distributions' lognormal body from the Pareto distributed upper tail. Using Monte Carlo simulations we motivate why this issue deserves more attention than related research has paid to it and highlight the consequences that follow from incorrect assumptions.

The first step of our simulation is the data generation of a stylized city size distribution. We randomly draw 1,000,000 observations from a lognormal distribution with a probability density function of

$$f_X^L(x) = \frac{1}{x\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(lnx-\mu)^2}{2\sigma^2}} \tag{B-1}$$

with $\mu = 0$ and $\sigma = 1$. The threshold T at which the Pareto tail begins is set to $x = 6$. We discard all observations above that cutoff and attach a matching Pareto tail. That tail needs to have the same density as the lognormal body at the cutoff to avoid any discontinuities. Hence, we compute the lognormal distribution's density at the threshold $T = 6$ through (B-1) and set the Pareto distribution's parameters to reach the same density at $x = 6$. The Pareto distribution's probability density function depends on a scale parameter, x_m, a slope parameter, α, and looks as follows:

$$f_X^P(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}} \tag{B-2}$$

Under Zipf's law, it holds that $\alpha = 1$. Plugging this in and using $f_X^P(T) = f_X^L(T)$ determines the scale parameter x_m as follows:

$$x_m = f_X^L(T) \cdot T^2 \tag{B-3}$$

For $T = 6$ we obtain $x_m \approx 0.48$. Accordingly, we draw 1,000,000 random observations from the derived Pareto distribution, discard all values below the threshold and attach

the remainder to the lognormal body. The outcome is a simulated city size distribution with a lognormal body and a Pareto distributed tail as suggested by the literature, see Figure B-1.

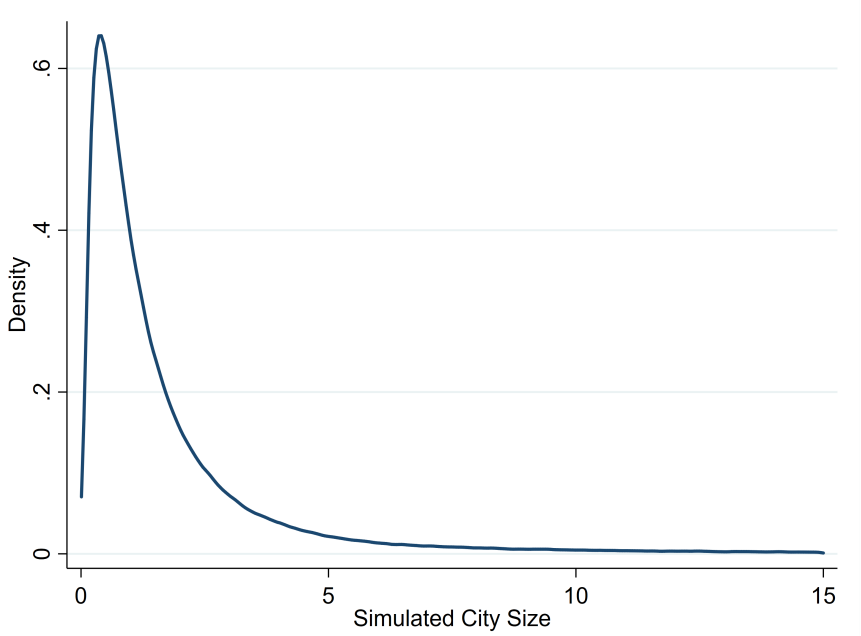


Figure B-1 – Kernel Density Plot of Simulation Outcome (Zoom on $x < 15$)

Overall we draw 1,000 of these distributions. In our Monte Carlo simulations we know the true threshold and obtain an estimated alpha coefficient of approximately unity for the interval $[6, \infty[$.
In real city size distributions the threshold is unknown and has to be placed by assumption. What happens if you set it too low and include smaller cities that should belong to the lognormal body? And, conversely, what happens if you set the threshold too high, hence using too few cities from the Pareto tail for the estimation of the alpha parameter?

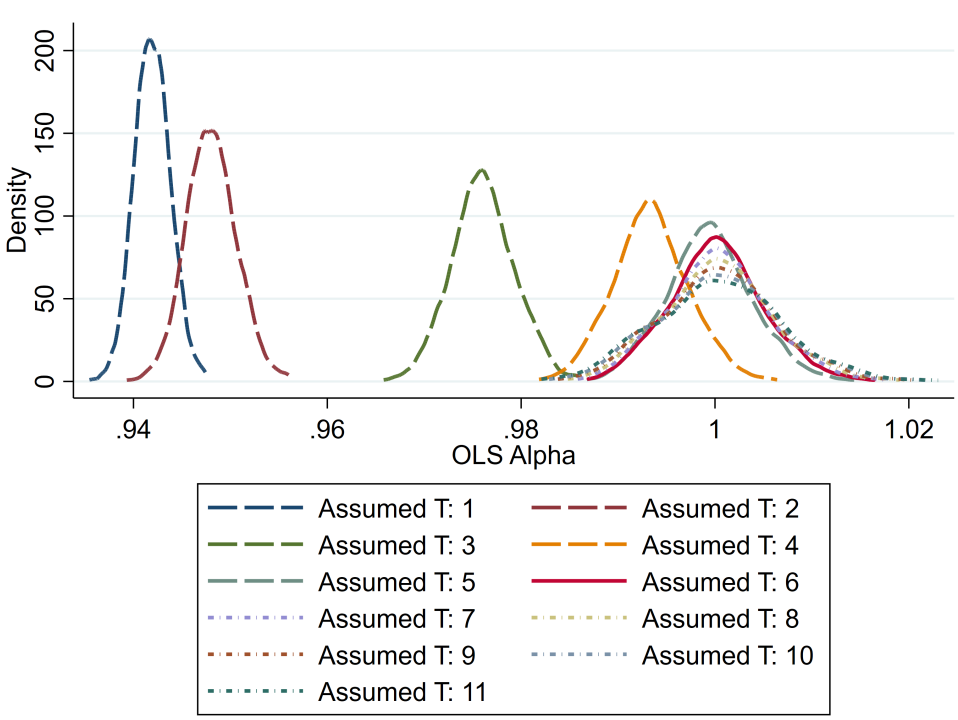


Figure B-2 – Comparison of Assumed Thresholds (True T = 6)

Figure B-2 gives the alpha estimates which are obtained when the threshold is set too low (assumed $T < 6$) or too high (assumed $T > 6$). Dissecting the distribution above the true threshold, i.e. within the Pareto upper tail, does not bias the results but inflates the variance and produces imprecise results because fewer cities are included. The reason behind this is that the assumed shorter tail distribution is still Pareto, hence the parameter estimate is unbiased. The lower number observation is not a major issue in a simulation with more than 1,000,000 observations in each draw but poses a problem in practice, as many countries have much shorter distributions. The standard errors will then be inflated, invalidating inference about Zipf's law.

On the other hand, an assumed threshold which is too low, yields a biased parameter estimate. When observations from the lognormal body are included, the estimation of the Pareto alpha is carried out based on a mixed distribution, leading to results which are centered away from $\alpha = 1$ - even as Zipf's law holds above the correct threshold. We conclude that care should be taken not to set the threshold at too low a value.

B.2 Threshold choice in practice

Section B.1 motivates the sensitivity of our estimates to threshold assumptions. These simulations illustrate the problem but do not help us identify the optimal solution to our real world applications.

Obviously, the threshold between city size distributions' lognormal body and the Pareto distributed upper tail varies between countries. Not only the Pareto tail's length but the

absolute size of its smallest included city should vary between, say, China and Greece. We need to design an identification strategy that can be applied worldwide and which accounts for country-specific heterogeneity.

In a first step we have a minimum city size of 50,000 inhabitants for all countries in our data set as a consequence of our city identification scheme. According to the World Bank (2009), this is the settlement size above which agglomeration effects come into play.

In the second step we need to determine the fraction of those cities with more than 50,000 inhabitants in a given country that belong to the Pareto tail. The threshold should grant comparability across countries and measures, i.e. light and population, and should be set as low as possible to maximize estimation precision but not too low either. One method of assessing whether observations are Pareto distributed is the graphical assessment of the linearity in Zipf Plots, e.g. Figure 3 in the paper. However, Discriminant Moment Ration Plots (Cirillo, 2013) offer a more clear-cut identification and a more aggregate cross-country comparison than is available for Zipf Plots. Discriminant Moment Ratio Plots identifiy a distribution as Pareto based on its skewness and coefficient of variation. Figure B-3 illustrates such a plot for corrected lights in the year 2000. Including all cities with more than 50,000 inhabitants, such as in the displayed scenario, leads to the rejection of the Pareto distribution for many countries as too many observations from the lognormal body are included. This holds for both, light and population, across all years.

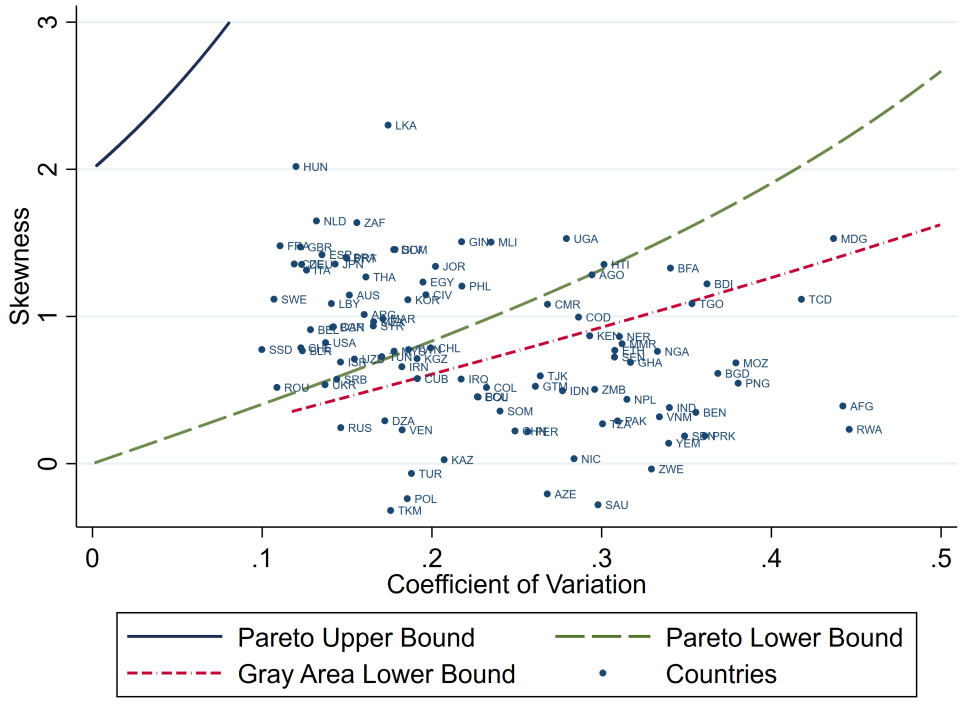


Figure B-3 – Discriminant Moment Ratio Plot (Log Corrected Light, Full Distribution, Year 2000, Satellite F15)

We should therefore use a higher threshold. In order to find the optimal threshold, we

design an algorithm counting the number of countries in the Discriminant Moment Ratio Plot's areas. The aim is to maximize the fraction of countries located in the Pareto area without losing too many observations. Setting the threshold at a (cumulative) percentile of, say, 95% would be a rather safe way of ensuring a Pareto distribution but only 5% of observations could then be used for the estimation of the Pareto alpha. We test all available percentiles of the city size distribution for all three measures. Figure B-4 displays the share of countries falling into the Pareto area for the given threshold setting applied to data from the year 2000. In terms of light, the share in the light data peaks and approximates the share in the population data around the median. But we also note that the graphs are not strictly increasing in the percentile thresholds.

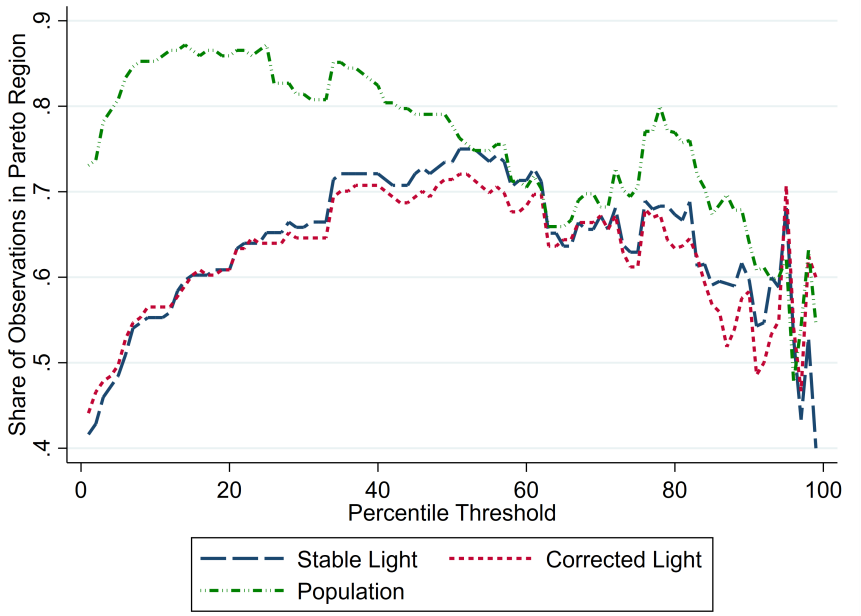


Figure B-4 – Share of Countries in DMRP's Pareto Region (Year 2000, Satellite F15)

However, this picture changes once we restrict the graph to countries with at least 10 cities above the respective percentile cutoff. To avoid our results being driven by outlier countries with, say, three cities, where a Pareto tail cannot be sensibly established, we therefore restrict our empirical Zipf law investigation to countries with at least 10 cities above the threshold. Now, the curves are converging towards unity, in line with the underlying theory about the lognormal body and the Pareto tail. The higher you set the threshold, the more likely you are to obtain a Pareto distribution. Notably, the median threshold is where all three size measures reach equally high levels of about 90% of countries in the Pareto area. Further increasing the threshold hardly brings any improvements in terms of getting more countries into the Pareto area, but it would lead to a loss of observations per country. To illustrate that other years produce the same result Figure B-6 plots the graphs shown in Figure B-4 and Figure B-5 for all years. We note the same pattern. From these findings we conclude that setting a threshold of

including the top 50% of cities above 50,000 inhabitants into the Pareto alpha estimation is the optimal threshold choice.

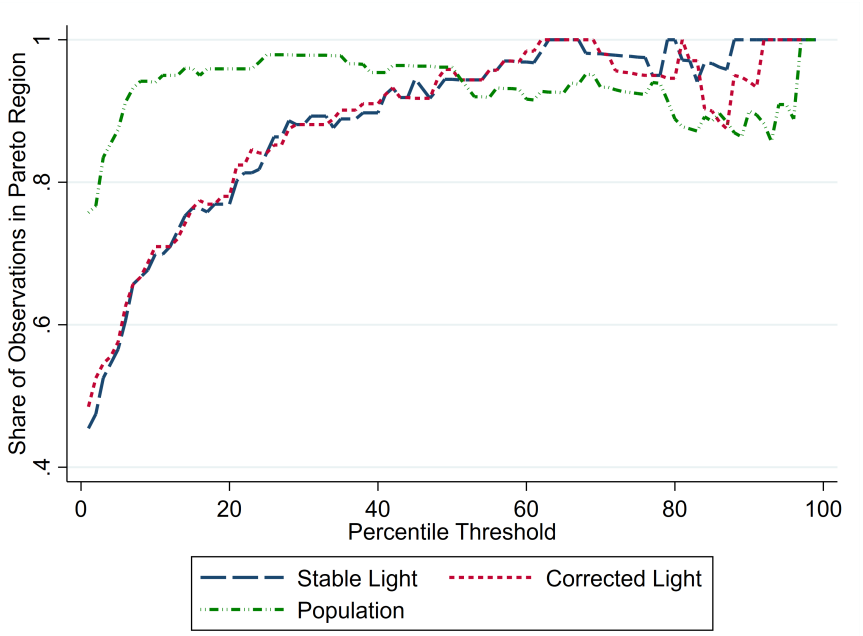
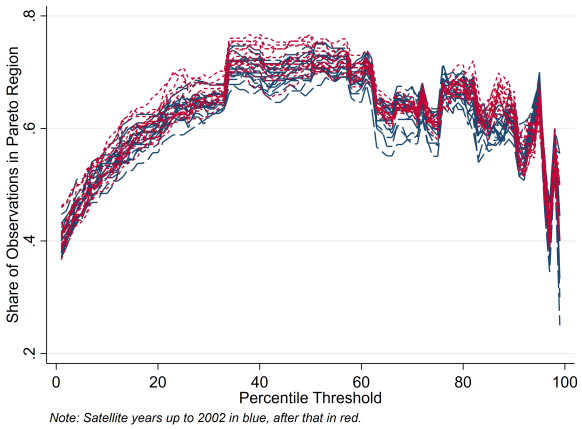
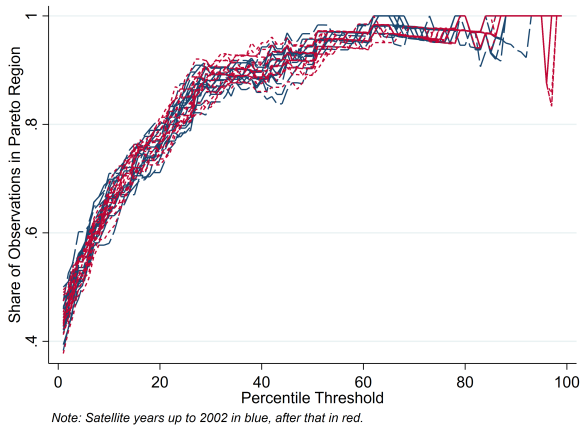


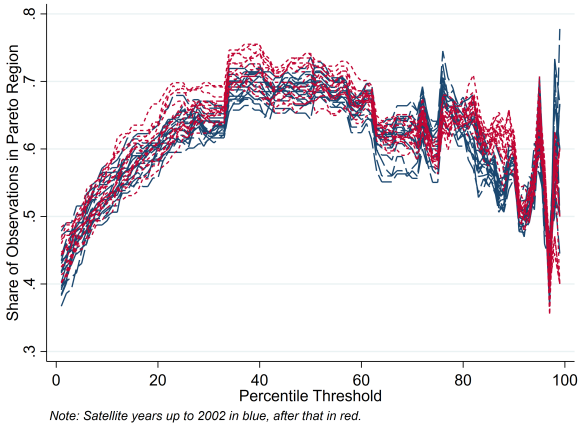
Figure B-5 – Share of Countries in DMRP's Pareto Region (Min. 10 Cities above Threshold, Year 2000, Satellite F15)



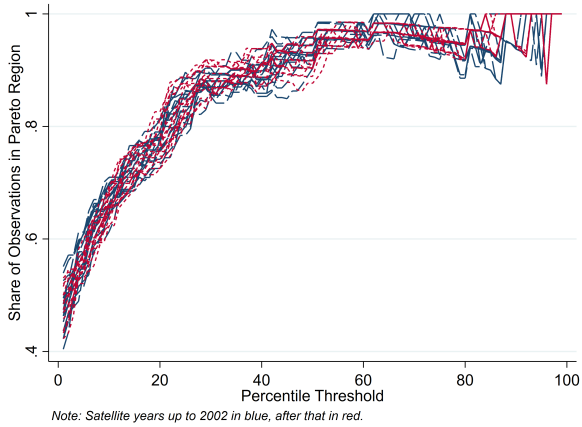
(a) Stable Light



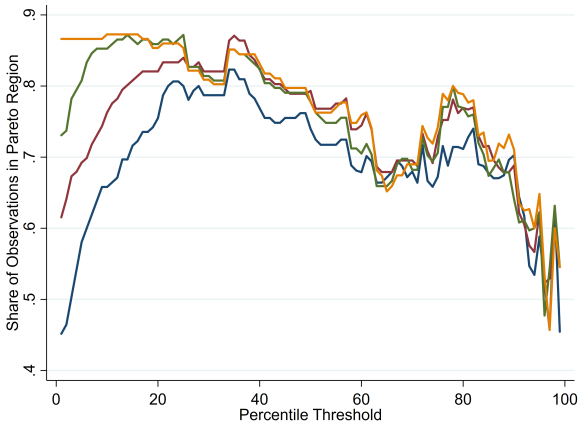
(b) Stable Light (Min. 10 Cities)



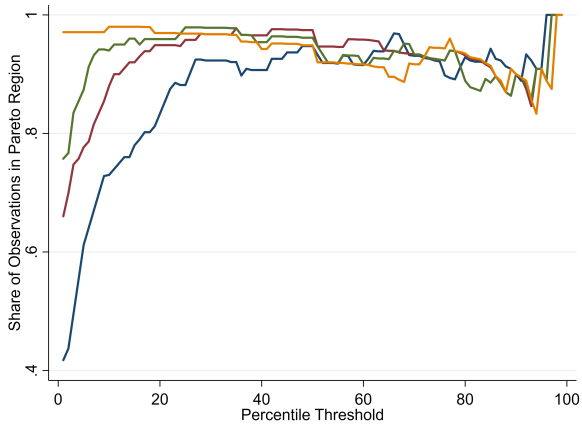
(c) Corrected Light



(d) Corrected Light (Min. 10 Cities)



(e) Population



(f) Population (Min. 10 Cities)

Figure B-6 – Share of Countries in DMRP's Pareto Region

B.3 Alternative thresholds

We also explore other threshold setting mechanisms, but find that they come with serious drawbacks.

For example, using relative thresholds, such as the top X cities in each country, is inconsistent with the threshold requirements and country-specific heterogeneity. Selecting, say, the thirty largest settlements in both Greece and China, will include many Greek towns from the distributional body and use fewer cities in China than would be optimal, given its large number of cities in the Pareto tail.

Population-proportional thresholds are another simplistic approach and select cities based on their share of the total population. However, they suffer from sample size sensitivity and a dependence on the degree of urbanization (Cheshire, 1999).

A strategy close but inferior to our baseline approach is an increasing population or lights threshold. The idea is similar to international poverty lines as used in that strand of the literature by, inter alia, Ravallion and Chen (2011). Let us outline the idea using a linear example and the threshold T_{it} of country i at time t be determined by its number of cities N_{it} according to

$$T_{it} = L + (N_{it} - N_L) \cdot \frac{U - L}{N_U - N_L} \tag{B-4}$$

for $N_L < N_{it} \leq N_U$, $T_{it} = L$ if $N_{it} \leq N_L$ and $T_{it} = U$ otherwise. With arbitrary values plugged in, this equation could look as follows:

$$T_{it} = 50,000 + (N_{it} - 20) \cdot \frac{300,000 - 50,000}{2,000 - 20} \tag{B-5}$$

for country i with $20 < N \leq 2,000$ cities at time t. Countries with up to twenty cities are subject to an absolute population threshold of 50,000. Starting with the 21st city, this cutoff increases linearly by $\frac{250,000}{1,980}$ until it reaches a city size of 300,000 with the 2,000th city. Above that the threshold remains constant again.

As a robustness check, we apply this threshold to the city size distribution in terms of population in our data set. The Pareto alpha coefficients using this linear threshold and our above-median threshold are rather similar for most countries (Figure B-7).

One drawback of this approach is the high number of underlying assumptions. The upper bound, U, of 300,000 inhabitants, the switching points at $N_L = 20$ and $N_U = 2,000$ and the linearity of the connection in between are purely arbitrary. However, the largest obstacle to this strategy is that it is unfeasible with the multiple city size measures used in this paper. Given we choose some upper bound or at least a slope for the population data, we would have to identify the corresponding settings for both nighttime light measures. But the luminosity of a city with a given number of inhabitants varies widely across countries. By contrast, our baseline strategy of working with the full- and the above-median distribution is free from such concerns and only relies on the properties of the distributions themselves. It requires little assumptions and enables cross-measure comparisons, which makes it our preferred threshold setting strategy.

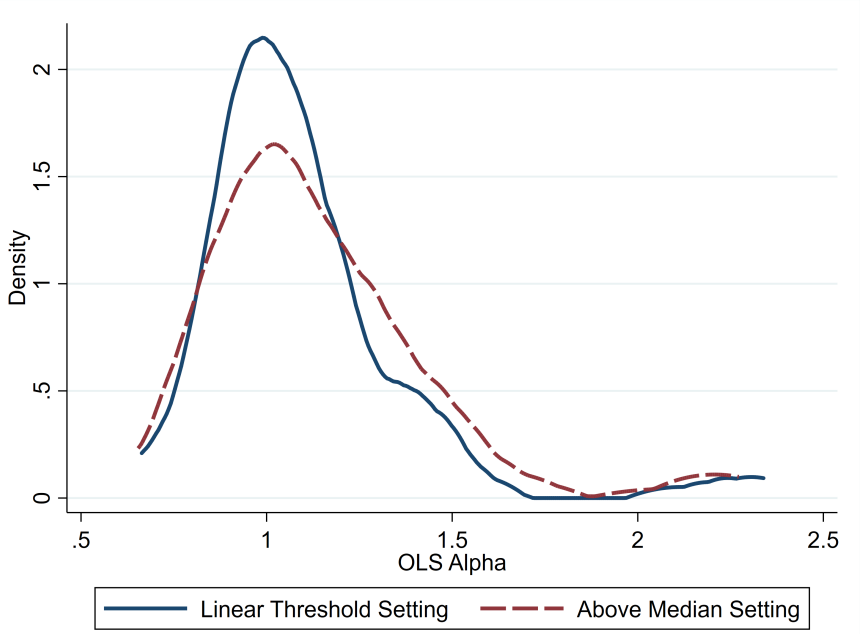


Figure B-7 – Comparison of Threshold Settings (Year 2000)

C Detailed Pareto alpha results for all countries

This section supplements the analysis in the paper by providing the Pareto OLS estimates for all countries in the year 2000. Table C-1 contains all the Pareto estimates for the size distribution of cities in terms of, respectively, population, 'stable' light and corrected light, both for the whole distribution of cities in each countries and only cities above the median. When comparing the coefficient estimates, one should keep in mind that the full distribution might potentially include cities from the lognormal body, leading to a slightly biased estimate, but smaller standard errors. By contrast, the above median distribution contains only observations from the Pareto tails, yielding an unbiased estimate but has larger standard errors due to fewer observations, see also Appendix B.

Table C-1 – OLS Pareto Alpha Coefficient Estimates (Year 2000, Satellite F15)

	Full Distribution			Above Median Setting		
Country	Stable Lights	Corrected Lights	Pop.	Stable Lights	Corrected Lights	Pop.
Afghanistan	0.448	0.447	0.314	0.564	0.561	1.107
	(0.135)	(0.135)	(0.052)	(0.241)	(0.239)	(0.261)
Algeria	0.800	0.748	0.978	1.315	1.088	1.364
	(0.118)	(0.110)	(0.144)	(0.274)	(0.227)	(0.284)
Angola	0.641	0.625	0.682	0.678	0.657	1.317
	(0.165)	(0.161)	(0.139)	(0.247)	(0.240)	(0.380)
Argentina	0.860	0.715	0.908	0.988	0.837	0.937
	(0.149)	(0.124)	(0.157)	(0.240)	(0.203)	(0.227)
Australia	0.704	0.663	0.693	0.659	0.618	0.654
	(0.199)	(0.188)	(0.196)	(0.258)	(0.242)	(0.256)
Azerbaijan	0.486	0.481	0.995			
	(0.172)	(0.170)	(0.341)			
Bangladesh	0.621	0.620	0.742	0.874	0.872	1.559
	(0.042)	(0.042)	(0.047)	(0.084)	(0.084)	(0.140)
Belarus	0.973	0.917	0.931			
	(0.324)	(0.306)	(0.310)			
Belgium	0.985	0.779	0.926			
	(0.338)	(0.267)	(0.317)			
Benin	0.512	0.512	0.721	0.670	0.669	1.040
	(0.158)	(0.158)	(0.200)	(0.286)	(0.285)	(0.408)
Bolivia	0.544	0.486	0.674			
	(0.222)	(0.198)	(0.275)			
Brazil	0.986	0.846	0.983	1.078	0.901	0.963

Table C-1 – OLS Pareto Alpha Coefficient Estimates (Year 2000, Satellite F15)

	Full Distribution			Above Median Setting		
Country	Stable Lights	Corrected Lights	Pop.	Stable Lights	Corrected Lights	Pop.
	(0.075)	(0.065)	(0.075)	(0.117)	(0.097)	(0.104)
Bulgaria	0.894	0.817	1.015			
	(0.400)	(0.365)	(0.454)			
Burkina Faso	0.584	0.582	1.085	0.589	0.586	1.089
	(0.169)	(0.168)	(0.280)	(0.240)	(0.239)	(0.398)
Burundi			0.762			1.481
			(0.204)			(0.560)
Cameroon	0.649	0.647	0.781	0.689	0.683	1.098
	(0.168)	(0.167)	(0.158)	(0.252)	(0.250)	(0.311)
Canada	0.956	0.651	0.869	0.977	0.711	0.874
	(0.193)	(0.132)	(0.175)	(0.276)	(0.201)	(0.247)
Chad	0.563	0.563	0.276			1.660
	(0.205)	(0.205)	(0.059)			(0.500)
Chile	0.686	0.612	0.950	0.940	0.733	0.889
	(0.169)	(0.151)	(0.230)	(0.322)	(0.259)	(0.305)
China	0.595	0.586	1.043	1.016	0.943	1.225
	(0.018)	(0.018)	(0.031)	(0.043)	(0.040)	(0.051)
Colombia	0.655	0.580	0.831	0.923	0.744	0.902
	(0.106)	(0.093)	(0.133)	(0.209)	(0.168)	(0.204)
Congo, D.R.	0.609	0.607	0.661	0.713	0.709	1.116
	(0.140)	(0.139)	(0.078)	(0.231)	(0.230)	(0.186)
Côte d'Ivoire	0.753	0.744	0.951	0.859	0.836	0.846
	(0.213)	(0.210)	(0.269)	(0.337)	(0.328)	(0.332)
Cuba	0.713	0.713	1.023	0.969	0.967	1.124
	(0.210)	(0.210)	(0.302)	(0.396)	(0.395)	(0.459)
Czech Rep.	1.068	0.920	1.095			
	(0.390)	(0.336)	(0.400)			
Dominican Rep.	0.793	0.706	0.845			
	(0.280)	(0.250)	(0.299)			
Ecuador	0.619	0.589	0.899	0.909	0.805	0.976
	(0.160)	(0.152)	(0.232)	(0.332)	(0.294)	(0.357)
Egypt	0.703	0.658	0.794	0.785	0.735	0.782
	(0.100)	(0.094)	(0.113)	(0.157)	(0.147)	(0.156)
El Salvador	0.770	0.706	0.898			

Table C-1 – OLS Pareto Alpha Coefficient Estimates (Year 2000, Satellite F15)

	Full Distribution			Above Median Setting		
Country	Stable Lights	Corrected Lights	Pop.	Stable Lights	Corrected Lights	Pop.
	(0.328)	(0.301)	(0.383)			
Eritrea			1.285			
			(0.486)			
Ethiopia	0.712	0.711	0.377	0.940	0.934	2.086
	(0.113)	(0.112)	(0.027)	(0.210)	(0.209)	(0.214)
France	1.142	0.956	1.050	1.336	1.067	1.147
	(0.181)	(0.151)	(0.166)	(0.299)	(0.239)	(0.256)
Germany	0.997	0.901	0.975	1.048	0.967	0.999
	(0.152)	(0.137)	(0.149)	(0.226)	(0.209)	(0.215)
Ghana	0.535	0.533	1.045	0.688	0.682	0.986
	(0.113)	(0.112)	(0.216)	(0.203)	(0.201)	(0.285)
Guatemala	0.580	0.568	1.075	0.779	0.727	0.919
	(0.150)	(0.147)	(0.269)	(0.284)	(0.266)	(0.325)
Guinea	0.744	0.745	1.014			
	(0.292)	(0.292)	(0.370)			
Haiti	0.608	0.608	0.931			0.838
	(0.230)	(0.230)	(0.287)			(0.357)
Hungary	0.938	0.837	0.888			
	(0.420)	(0.374)	(0.397)			
India	0.558	0.555	0.730	0.876	0.854	1.386
	(0.014)	(0.014)	(0.017)	(0.030)	(0.029)	(0.046)
Indonesia	0.580	0.576	0.850	0.826	0.811	1.014
	(0.043)	(0.043)	(0.060)	(0.088)	(0.086)	(0.101)
Iran	0.801	0.692	1.038	1.137	0.891	1.183
	(0.088)	(0.076)	(0.114)	(0.177)	(0.138)	(0.184)
Iraq	0.666	0.644	0.825	0.918	0.835	1.094
	(0.111)	(0.107)	(0.137)	(0.216)	(0.197)	(0.254)
Israel	0.865	0.673	0.741			
	(0.339)	(0.264)	(0.291)			
Italy	0.936	0.870	0.921	0.960	0.908	0.888
	(0.163)	(0.151)	(0.160)	(0.236)	(0.224)	(0.219)
Japan	0.897	0.754	0.847	0.978	0.859	0.881
	(0.117)	(0.098)	(0.110)	(0.180)	(0.158)	(0.162)
Jordan	0.699	0.599	0.709			

Table C-1 – OLS Pareto Alpha Coefficient Estimates (Year 2000, Satellite F15)

	Full Distribution			Above Median Setting		
Country	Stable Lights	Corrected Lights	Pop.	Stable Lights	Corrected Lights	Pop.
	(0.285)	(0.244)	(0.289)			
Kazakhstan	0.594	0.560	1.088	1.445	1.214	1.641
	(0.138)	(0.130)	(0.253)	(0.469)	(0.394)	(0.533)
Kenya	0.605	0.603	0.960	0.714	0.710	0.917
	(0.147)	(0.146)	(0.215)	(0.232)	(0.231)	(0.290)
Kyrgyzstan	0.714	0.712	1.023			
	(0.270)	(0.269)	(0.387)			
Libya	0.962	0.786	1.125			
	(0.340)	(0.278)	(0.398)			
Madagascar	0.461	0.460	0.530			0.886
	(0.206)	(0.206)	(0.172)			(0.396)
Malaysia	0.687	0.625	0.817	0.950	0.746	0.913
	(0.167)	(0.152)	(0.198)	(0.326)	(0.256)	(0.313)
Mali	0.698	0.696	1.006			
	(0.255)	(0.254)	(0.356)			
Mexico	0.909	0.709	0.800	1.128	0.853	0.999
	(0.101)	(0.079)	(0.089)	(0.177)	(0.134)	(0.157)
Morocco	0.864	0.784	0.999	0.982	0.870	0.976
	(0.158)	(0.143)	(0.182)	(0.254)	(0.225)	(0.252)
Mozambique	0.520	0.519	0.278	0.707	0.702	1.387
	(0.128)	(0.128)	(0.048)	(0.243)	(0.241)	(0.336)
Myanmar	0.652	0.651	0.575	0.839	0.836	1.289
	(0.092)	(0.092)	(0.073)	(0.168)	(0.167)	(0.232)
Nepal	0.585	0.583	0.621	0.852	0.838	1.466
	(0.144)	(0.144)	(0.148)	(0.292)	(0.287)	(0.489)
Netherlands	1.046	0.808	1.106	1.130	0.814	1.117
	(0.247)	(0.190)	(0.261)	(0.377)	(0.271)	(0.372)
Nicaragua	0.493	0.483	0.994			
	(0.193)	(0.189)	(0.390)			
Niger	0.625	0.622	1.029	0.736	0.727	1.402
	(0.188)	(0.188)	(0.206)	(0.314)	(0.310)	(0.396)
Nigeria	0.573	0.560	0.852	0.735	0.699	1.279
	(0.046)	(0.045)	(0.058)	(0.083)	(0.079)	(0.124)
North Korea	0.518	0.517	1.196	0.829	0.817	1.170

Table C-1 – OLS Pareto Alpha Coefficient Estimates (Year 2000, Satellite F15)

	Full Distribution			Above Median Setting		
Country	Stable Lights	Corrected Lights	Pop.	Stable Lights	Corrected Lights	Pop.
	(0.119)	(0.119)	(0.207)	(0.269)	(0.265)	(0.284)
Pakistan	0.532	0.524	0.607	0.913	0.849	1.137
	(0.041)	(0.040)	(0.045)	(0.099)	(0.092)	(0.119)
Papua New Guinea	0.481	0.480	1.718			2.262
	(0.215)	(0.215)	(0.558)			(1.012)
Peru	0.544	0.523	0.634	0.869	0.777	0.933
	(0.122)	(0.117)	(0.140)	(0.275)	(0.246)	(0.288)
Philippines	0.729	0.718	0.940	0.851	0.823	0.975
	(0.105)	(0.104)	(0.136)	(0.174)	(0.168)	(0.199)
Poland	0.622	0.573	0.982	1.032	0.886	1.021
	(0.123)	(0.113)	(0.194)	(0.286)	(0.246)	(0.283)
Portugal	0.822	0.655	0.774			
	(0.351)	(0.279)	(0.330)			
Romania	1.111	1.025	0.366	1.648	1.406	1.393
	(0.282)	(0.260)	(0.093)	(0.583)	(0.497)	(0.492)
Russia	0.850	0.741	1.042	1.300	1.111	1.219
	(0.078)	(0.068)	(0.095)	(0.168)	(0.144)	(0.157)
Rwanda			0.956			
			(0.375)			
Saudi Arabia	0.449	0.359	0.839	0.852	0.681	0.916
	(0.087)	(0.070)	(0.163)	(0.232)	(0.185)	(0.249)
Senegal	0.580	0.579	0.505	0.775	0.769	1.273
	(0.155)	(0.155)	(0.122)	(0.293)	(0.291)	(0.437)
Serbia	0.843	0.802	1.111			
	(0.319)	(0.303)	(0.420)			
Somalia	0.719	0.719	1.009			
	(0.322)	(0.322)	(0.336)			
South Africa	0.842	0.788	0.947	0.899	0.827	0.901
	(0.143)	(0.134)	(0.161)	(0.215)	(0.198)	(0.215)
South Korea	0.789	0.596	0.708	0.893	0.653	0.770
	(0.166)	(0.126)	(0.149)	(0.263)	(0.193)	(0.227)
South Sudan			0.193			2.272
			(0.041)			(0.670)
Spain	1.085	0.803	0.999	1.141	0.865	1.035

Table C-1 – OLS Pareto Alpha Coefficient Estimates (Year 2000, Satellite F15)

	Full Distribution			Above Median Setting		
Country	Stable Lights	Corrected Lights	Pop.	Stable Lights	Corrected Lights	Pop.
	(0.177)	(0.131)	(0.163)	(0.262)	(0.198)	(0.237)
Sri Lanka	0.797	0.798	0.872	0.686	0.686	0.737
	(0.240)	(0.241)	(0.263)	(0.292)	(0.293)	(0.314)
Sudan	0.493	0.490	0.363	0.827	0.803	1.455
	(0.092)	(0.091)	(0.047)	(0.213)	(0.207)	(0.264)
Sweden	1.038	0.903	0.955			
	(0.424)	(0.369)	(0.390)			
Switzerland	1.024	0.892	1.019			
	(0.362)	(0.315)	(0.360)			
Syria	0.827	0.730	0.837	0.934	0.812	0.809
	(0.239)	(0.211)	(0.241)	(0.381)	(0.332)	(0.330)
Taiwan	0.708	0.551	0.654	0.793	0.598	0.664
	(0.230)	(0.179)	(0.212)	(0.355)	(0.267)	(0.297)
Tajikistan	0.616	0.615	1.407	0.813	0.809	1.409
	(0.162)	(0.162)	(0.370)	(0.297)	(0.295)	(0.515)
Tanzania	0.541	0.540	0.718	0.855	0.850	1.170
	(0.129)	(0.129)	(0.157)	(0.285)	(0.283)	(0.361)
Thailand	0.833	0.795	1.159	0.977	0.919	1.060
	(0.147)	(0.141)	(0.205)	(0.244)	(0.230)	(0.265)
Togo	0.530	0.530	0.734			
	(0.208)	(0.208)	(0.245)			
Tunisia	0.796	0.740	1.176	0.952	0.871	1.028
	(0.209)	(0.194)	(0.309)	(0.348)	(0.329)	(0.375)
Turkey	0.668	0.650	0.959	1.030	0.947	1.025
	(0.085)	(0.083)	(0.122)	(0.185)	(0.170)	(0.184)
Turkmenistan	0.630	0.613	1.041			
	(0.282)	(0.274)	(0.465)			
Uganda	0.654	0.653	0.442	0.663	0.662	1.195
	(0.193)	(0.193)	(0.079)	(0.271)	(0.270)	(0.304)
Ukraine	0.876	0.862	1.061	1.288	1.236	1.329
	(0.139)	(0.137)	(0.169)	(0.288)	(0.276)	(0.297)
United Kingdom	1.071	0.907	1.074	1.138	0.963	1.092
	(0.132)	(0.112)	(0.133)	(0.198)	(0.168)	(0.190)
United States	0.852	0.698	0.841	0.971	0.850	0.882

Table C-1 – OLS Pareto Alpha Coefficient Estimates (Year 2000, Satellite F15)

	Full Distribution			Above Median Setting		
Country	Stable Lights	Corrected Lights	Pop.	Stable Lights	Corrected Lights	Pop.
	(0.067)	(0.055)	(0.066)	(0.108)	(0.094)	(0.098)
Uzbekistan	0.839	0.828	0.971	1.039	1.009	1.029
	(0.143)	(0.141)	(0.165)	(0.248)	(0.241)	(0.246)
Venezuela	0.750	0.633	1.000	1.215	0.924	1.175
	(0.138)	(0.116)	(0.184)	(0.314)	(0.239)	(0.303)
Vietnam	0.538	0.538	0.606	0.826	0.819	1.195
	(0.053)	(0.052)	(0.057)	(0.114)	(0.113)	(0.160)
Yemen	0.504	0.499	0.504	0.885	0.831	1.193
	(0.106)	(0.105)	(0.094)	(0.261)	(0.245)	(0.313)
Zambia	0.541	0.540	1.013	0.720	0.713	1.287
	(0.137)	(0.137)	(0.232)	(0.255)	(0.252)	(0.417)
Zimbabwe	0.438	0.438	1.033	0.704	0.704	0.975
	(0.129)	(0.129)	(0.281)	(0.287)	(0.287)	(0.369)

Note: Coefficients are only calculated for distributions with at least ten cities of non-zero city size. A few cities have a non-zero population but no detected light emissions. That can lead to a shorter city size distribution for lights than for population. Corrected standard errors according to $\sqrt{2/N} \cdot \hat{\alpha}$ (Gabaix and Ibragimov, 2011) are given in parentheses.

D Robustness check: Hill estimator

This section discusses an alternative to OLS estimation of the log-rank regression

$$\log \text{rank}(y_i) - \log N \approx \alpha \log y_c - \alpha \log y_i. \tag{D-1}$$

The Hill estimator (Hill, 1975) is given by

$$\hat{\alpha}_{Hill} = \frac{N-1}{\sum_{i=1}^{N-1} log(y_i) - log(y_c)} \tag{D-2}$$

with standard errors as $SE_{Hill} = \hat{\alpha}_{Hill}/\sqrt{N-3}$ (Gabaix, 2009). If the data is indeed Pareto distributed, the Hill estimator is the maximum likelihood estimator and, by consequence, efficient. While the results presented in the paper are based on OLS estimation, we here repeat the analysis with the Hill estimator as a robustness check to confirm that we obtain the same pattern for the cross-country Pareto alpha distribution.

The key motivation to choose OLS over Hill as the baseline methodology is the Hill estimator's sensitivity to violations of the Pareto assumption. To illustrate that point we repeat the Monte Carlo simulation described in Section B.1.

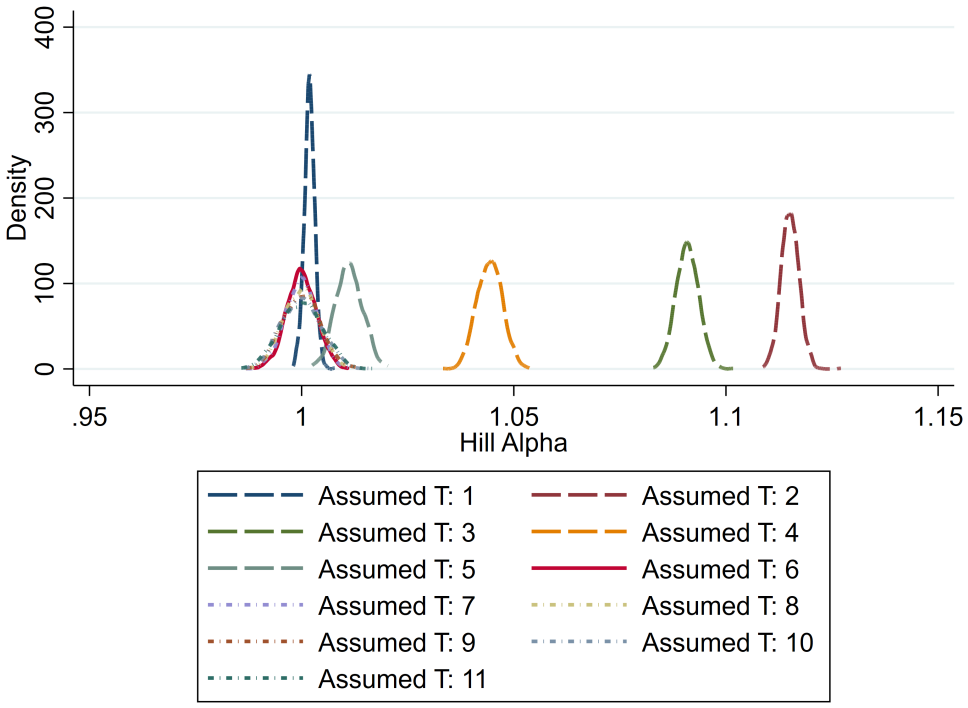


Figure D-1 – Hill Estimator: Comparison of Assumed Thresholds (True T = 6)

The comparison of Figure B-2 and Figure D-1 suggests that the bias of the OLS and the Hill estimator do not only point into opposite directions but that the bias from assuming the threshold within the lognormal body tends to be larger in absolute terms

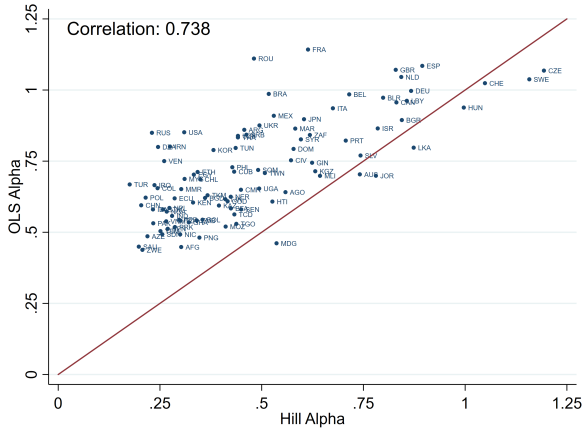
for Hill than for OLS. Hence, although the Hill estimator may be more efficient when the data is indeed Pareto distributed, its larger bias in cases of deviations from the Pareto distribution is a considerable drawback. Still, the Hill coefficient estimates provide a useful robustness check for our main results.

It is known that the two estimates are often highly correlated in empirical studies; for instance, Soo (2005) finds a correlation coefficient of 0.7 between the OLS and Hill estimators on his data set, with larger differences for countries with smaller city samples. In our case the correlation ranges above Soo's coefficient of 0.7: In the baseline sample (above median setting) we obtain correlation coefficients of 0.813 for stable lights, 0.722 for corrected lights and 0.760 for population.

The scatter plots between the OLS and Hill alpha estimates for all other measures (population, 'stable', and corrected light) and both the above-median and full-distribution setting shown in Figure D-2. We note strong correlation coefficients and see that, in particular in the above-median setting, most coefficients are clustered around the main diagonal. The cross-country patterns of the Pareto alpha in the city size distribution does not depend on whether the estimation is carried out by OLS or Hill estimation. However, when the full distribution of cities is used (scatter plots on the right), the majority of countries has a Hill Pareto alpha estimate which is lower than the OLS estimate. In this setting, many countries' distributions still contain cities from the lognormal body and are not purely Pareto. The Hill estimator is then likely to be biased. Still, the correlation coefficients remain high.

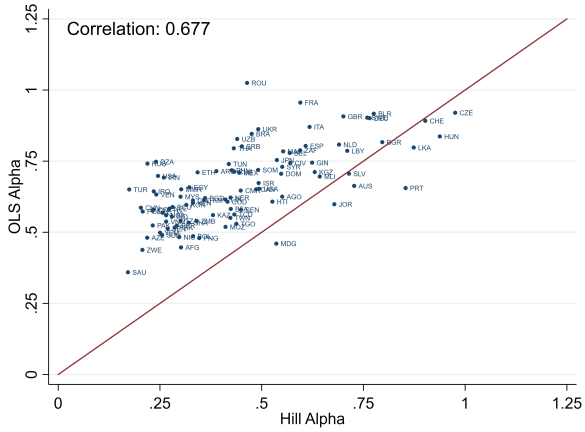
For more detailed insights we report Hill alpha estimates for all countries in the year 2000 in Table D-1 - analogous to Table C-1 for OLS. We see that the Hill coefficient estimates often deviate from Zipf's law more strongly, but the bias may play a role here. When only the above-median distribution is used, which is more likely to be purely Pareto, the Hill estimates are relatively close to the OLS estimates for most countries. Using Zambia as an example, the Hill (OLS) estimate in terms of population is 1.401 (1.287), for 'stable' lights 0.655 (0.720) and for corrected lights 0.652 (0.713). Overall, we conclude that our main insights about the city size distribution of countries around the world are robust to using the Hill rather than the OLS estimator.

(a) Stable Light (Above Median Setting)

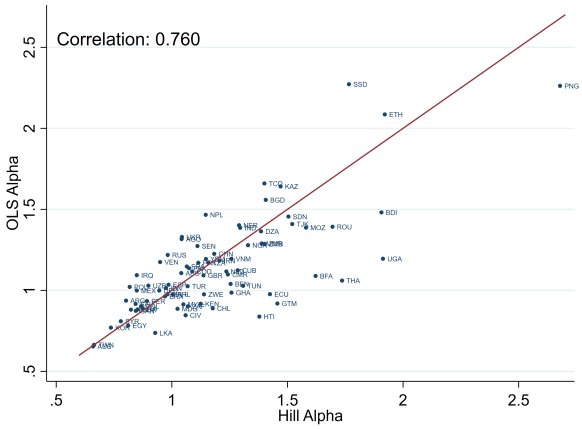


(b) Stable Light (Full Distribution Setting)

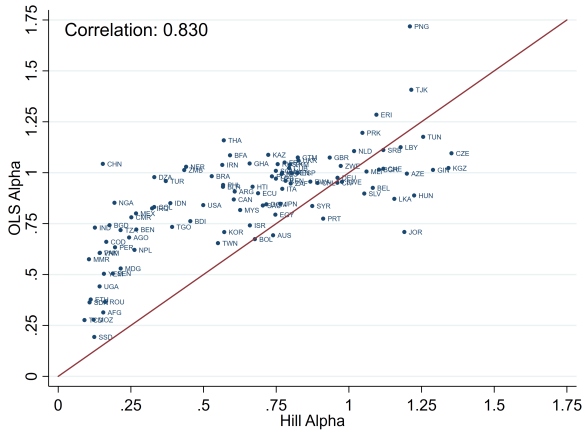
(c) Corrected Light (Above Median Setting)



(d) Corrected Light (Full Distribution Setting)



(e) Population (Above Median Setting)



(f) Population (Full Distribution Setting)

Figure D-2 – Comparison of OLS and Hill Alpha Estimates (Year 2000, Satellite F15)

Table D-1 – Hill Pareto Alpha Coefficient Estimates (Year 2000, Satellite F15)

	Full Distribution			Above Median Setting		
Country	Stable Lights	Corrected Lights	Pop.	Stable Lights	Corrected Lights	Pop.
Afghanistan	0.302	0.302	0.155	0.460	0.458	1.041
	(0.069)	(0.069)	(0.019)	(0.163)	(0.162)	(0.181)
Algeria	0.245	0.240	0.330	1.236	1.064	1.386
	(0.026)	(0.025)	(0.035)	(0.188)	(0.162)	(0.211)
Angola	0.558	0.550	0.244	0.674	0.651	1.042
	(0.107)	(0.106)	(0.036)	(0.195)	(0.188)	(0.227)
Argentina	0.457	0.389	0.607	0.923	0.815	0.802
	(0.057)	(0.049)	(0.076)	(0.166)	(0.146)	(0.144)
Australia	0.741	0.727	0.739	0.687	0.674	0.660
	(0.158)	(0.155)	(0.158)	(0.217)	(0.213)	(0.209)
Azerbaijan	0.219	0.219	1.199			
	(0.061)	(0.061)	(0.321)			
Bangladesh	0.361	0.361	0.176	0.781	0.780	1.406
	(0.017)	(0.017)	(0.008)	(0.053)	(0.053)	(0.090)
Belarus	0.798	0.775	0.567			
	(0.206)	(0.200)	(0.146)			
Belgium	0.715	0.569	1.082			
	(0.191)	(0.152)	(0.289)			
Benin	0.269	0.269	0.269	0.762	0.762	1.254
	(0.063)	(0.063)	(0.056)	(0.270)	(0.269)	(0.397)
Bolivia	0.355	0.331	0.676			
	(0.118)	(0.110)	(0.225)			
Brazil	0.517	0.475	0.528	1.056	0.908	0.970
	(0.028)	(0.026)	(0.029)	(0.081)	(0.070)	(0.075)
Bulgaria	0.844	0.796	1.103			
	(0.319)	(0.301)	(0.417)			
Burkina Faso	0.424	0.424	0.591	0.594	0.593	1.622
	(0.093)	(0.093)	(0.114)	(0.198)	(0.198)	(0.468)
Burundi			0.454			1.907
			(0.091)			(0.575)
Cameroon	0.449	0.448	0.251	0.732	0.729	1.242
	(0.086)	(0.086)	(0.037)	(0.211)	(0.211)	(0.265)
Canada	0.831	0.487	0.604	0.898	0.679	0.844
	(0.123)	(0.072)	(0.089)	(0.192)	(0.145)	(0.180)

Table D-1 – Hill Pareto Alpha Coefficient Estimates (Year 2000, Satellite F15)

	Full Distribution			Above Median Setting		
Country	Stable Lights	Corrected Lights	Pop.	Stable Lights	Corrected Lights	Pop.
Chad	0.433	0.433	0.090			1.400
	(0.125)	(0.125)	(0.014)			(0.321)
Chile	0.349	0.331	0.892	1.146	1.027	1.177
	(0.064)	(0.060)	(0.160)	(0.306)	(0.285)	(0.314)
China	0.204	0.203	0.153	0.863	0.836	1.183
	(0.004)	(0.004)	(0.003)	(0.026)	(0.025)	(0.035)
Colombia	0.244	0.234	0.330	0.818	0.656	0.867
	(0.028)	(0.027)	(0.038)	(0.136)	(0.109)	(0.145)
Congo, D.R.	0.416	0.416	0.165	0.760	0.758	1.088
	(0.070)	(0.070)	(0.014)	(0.190)	(0.190)	(0.131)
Côte d'Ivoire	0.572	0.570	0.960	0.793	0.787	1.060
	(0.122)	(0.122)	(0.205)	(0.251)	(0.249)	(0.335)
Cuba	0.432	0.432	0.794	1.245	1.246	1.285
	(0.097)	(0.097)	(0.178)	(0.415)	(0.415)	(0.428)
Czech Rep.	1.194	0.975	1.353			
	(0.345)	(0.282)	(0.391)			
Dominican Rep.	0.578	0.548	0.716			
	(0.160)	(0.152)	(0.199)			
Ecuador	0.286	0.281	0.688	0.870	0.784	1.424
	(0.055)	(0.054)	(0.132)	(0.251)	(0.226)	(0.411)
Egypt	0.333	0.322	0.747	0.745	0.731	0.811
	(0.034)	(0.033)	(0.076)	(0.109)	(0.107)	(0.118)
El Salvador	0.743	0.715	1.053			
	(0.263)	(0.253)	(0.372)			
Eritrea			1.095			
			(0.330)			
Ethiopia	0.342	0.342	0.112	0.900	0.899	1.921
	(0.039)	(0.039)	(0.006)	(0.148)	(0.148)	(0.140)
France	0.614	0.595	0.779	1.265	1.146	1.065
	(0.070)	(0.068)	(0.089)	(0.208)	(0.188)	(0.175)
Germany	0.867	0.765	0.961	0.954	0.896	0.946
	(0.095)	(0.084)	(0.105)	(0.151)	(0.142)	(0.150)
Ghana	0.321	0.321	0.658	0.682	0.680	1.258
	(0.050)	(0.049)	(0.099)	(0.153)	(0.152)	(0.275)

Table D-1 – Hill Pareto Alpha Coefficient Estimates (Year 2000, Satellite F15)

	Full Distribution			Above Median Setting		
Country	Stable Lights	Corrected Lights	Pop.	Stable Lights	Corrected Lights	Pop.
Guatemala	0.257	0.257	0.824	0.972	0.946	1.457
	(0.050)	(0.049)	(0.153)	(0.281)	(0.273)	(0.404)
Guinea	0.624	0.624	1.289			
	(0.197)	(0.197)	(0.372)			
Haiti	0.526	0.526	0.669			1.379
	(0.159)	(0.159)	(0.158)			(0.487)
Hungary	0.997	0.938	1.224			
	(0.377)	(0.354)	(0.463)			
India	0.280	0.279	0.126	0.720	0.713	1.295
	(0.005)	(0.005)	(0.002)	(0.018)	(0.017)	(0.030)
Indonesia	0.233	0.232	0.386	0.780	0.768	0.974
	(0.012)	(0.012)	(0.019)	(0.059)	(0.058)	(0.069)
Iran	0.275	0.260	0.564	1.234	0.985	1.205
	(0.022)	(0.020)	(0.044)	(0.138)	(0.110)	(0.135)
Iraq	0.237	0.235	0.323	0.749	0.715	0.849
	(0.028)	(0.028)	(0.039)	(0.130)	(0.124)	(0.146)
Israel	0.785	0.493	0.659			
	(0.248)	(0.156)	(0.208)			
Italy	0.675	0.618	0.771	0.962	0.779	0.874
	(0.085)	(0.078)	(0.097)	(0.176)	(0.142)	(0.160)
Japan	0.604	0.537	0.764	0.933	0.745	0.823
	(0.056)	(0.050)	(0.071)	(0.125)	(0.100)	(0.110)
Jordan	0.781	0.678	1.191			
	(0.260)	(0.226)	(0.397)			
Kazakhstan	0.395	0.380	0.723	0.820	0.718	1.470
	(0.068)	(0.065)	(0.124)	(0.205)	(0.180)	(0.368)
Kenya	0.331	0.331	0.782	0.905	0.903	1.125
	(0.060)	(0.059)	(0.129)	(0.226)	(0.226)	(0.273)
Kyrgyzstan	0.632	0.630	1.342			
	(0.190)	(0.190)	(0.405)			
Libya	0.857	0.710	1.178			
	(0.238)	(0.197)	(0.327)			
Madagascar	0.536	0.536	0.215			1.025
	(0.203)	(0.203)	(0.054)			(0.387)

Table D-1 – Hill Pareto Alpha Coefficient Estimates (Year 2000, Satellite F15)

	Full Distribution			Above Median Setting		
Country	Stable Lights	Corrected Lights	Pop.	Stable Lights	Corrected Lights	Pop.
Malaysia	0.310	0.301	0.626	0.874	0.751	1.050
	(0.056)	(0.054)	(0.112)	(0.234)	(0.201)	(0.281)
Mali	0.643	0.643	1.061			
	(0.186)	(0.186)	(0.294)			
Mexico	0.530	0.445	0.268	0.932	0.752	0.848
	(0.042)	(0.035)	(0.021)	(0.106)	(0.085)	(0.096)
Morocco	0.582	0.553	0.770	0.909	0.779	0.981
	(0.077)	(0.073)	(0.102)	(0.175)	(0.150)	(0.189)
Mozambique	0.411	0.411	0.122	0.636	0.634	1.580
	(0.075)	(0.075)	(0.015)	(0.170)	(0.169)	(0.284)
Myanmar	0.302	0.302	0.106	0.746	0.746	1.389
	(0.031)	(0.031)	(0.010)	(0.109)	(0.109)	(0.181)
Nepal	0.273	0.273	0.262	0.638	0.635	1.147
	(0.050)	(0.050)	(0.046)	(0.170)	(0.170)	(0.296)
Netherlands	0.843	0.690	1.018	1.249	0.978	1.236
	(0.147)	(0.120)	(0.177)	(0.322)	(0.252)	(0.319)
Nicaragua	0.300	0.297	0.769			
	(0.095)	(0.094)	(0.243)			
Niger	0.424	0.423	0.440	0.695	0.692	1.291
	(0.097)	(0.097)	(0.064)	(0.246)	(0.245)	(0.275)
Nigeria	0.267	0.265	0.193	0.602	0.586	1.329
	(0.015)	(0.015)	(0.009)	(0.049)	(0.048)	(0.092)
North Korea	0.287	0.287	1.046	0.654	0.652	1.114
	(0.048)	(0.048)	(0.131)	(0.163)	(0.163)	(0.200)
Pakistan	0.234	0.232	0.144	0.758	0.734	1.074
	(0.013)	(0.013)	(0.008)	(0.059)	(0.057)	(0.081)
Papua New Guinea	0.347	0.347	1.210			2.679
	(0.131)	(0.131)	(0.302)			(1.013)
Peru	0.297	0.291	0.195	0.876	0.785	0.892
	(0.049)	(0.048)	(0.032)	(0.213)	(0.190)	(0.210)
Philippines	0.428	0.426	0.567	0.767	0.758	1.005
	(0.044)	(0.044)	(0.059)	(0.114)	(0.113)	(0.150)
Poland	0.215	0.208	0.735	0.908	0.742	0.818
	(0.031)	(0.030)	(0.106)	(0.189)	(0.155)	(0.171)

Table D-1 – Hill Pareto Alpha Coefficient Estimates (Year 2000, Satellite F15)

	Full Distribution			Above Median Setting		
Country	Stable Lights	Corrected Lights	Pop.	Stable Lights	Corrected Lights	Pop.
Portugal	0.707	0.854	0.912			
	(0.250)	(0.302)	(0.322)			
Romania	0.481	0.464	0.161	1.731	1.706	1.695
	(0.091)	(0.088)	(0.030)	(0.480)	(0.473)	(0.470)
Russia	0.230	0.219	0.755	1.019	0.873	0.982
	(0.015)	(0.014)	(0.049)	(0.094)	(0.081)	(0.091)
Rwanda			0.867			
			(0.274)			
Saudi Arabia	0.198	0.171	0.704	0.810	0.531	0.843
	(0.028)	(0.024)	(0.100)	(0.165)	(0.108)	(0.172)
Senegal	0.450	0.449	0.188	0.531	0.530	1.111
	(0.090)	(0.090)	(0.034)	(0.160)	(0.160)	(0.297)
Serbia	0.462	0.451	1.118			
	(0.139)	(0.136)	(0.337)			
Somalia	0.491	0.491	0.749			
	(0.186)	(0.186)	(0.193)			
South Africa	0.618	0.594	0.800	0.935	0.853	1.069
	(0.076)	(0.073)	(0.098)	(0.165)	(0.151)	(0.189)
South Korea	0.381	0.315	0.571	0.911	0.659	0.736
	(0.059)	(0.049)	(0.088)	(0.204)	(0.147)	(0.165)
South Sudan			0.124			1.766
			(0.019)			(0.395)
Spain	0.895	0.608	0.824	1.137	0.822	0.986
	(0.105)	(0.072)	(0.097)	(0.192)	(0.139)	(0.167)
Sri Lanka	0.874	0.874	1.156	0.916	0.916	0.928
	(0.200)	(0.200)	(0.265)	(0.324)	(0.324)	(0.328)
Sudan	0.256	0.255	0.108	0.699	0.694	1.504
	(0.034)	(0.034)	(0.010)	(0.135)	(0.134)	(0.198)
Sweden	1.158	0.759	0.976			
	(0.386)	(0.253)	(0.325)			
Switzerland	1.049	0.902	1.119			
	(0.291)	(0.250)	(0.310)			
Syria	0.596	0.550	0.874	0.714	0.619	0.779
	(0.130)	(0.120)	(0.191)	(0.238)	(0.206)	(0.260)

Table D-1 – Hill Pareto Alpha Coefficient Estimates (Year 2000, Satellite F15)

	Full Distribution			Above Median Setting		
Country	Stable Lights	Corrected Lights	Pop.	Stable Lights	Corrected Lights	Pop.
Taiwan	0.508	0.423	0.550	0.828	0.583	0.664
	(0.127)	(0.106)	(0.137)	(0.313)	(0.220)	(0.251)
Tajikistan	0.411	0.411	1.215	0.632	0.631	1.521
	(0.081)	(0.081)	(0.238)	(0.182)	(0.182)	(0.439)
Tanzania	0.301	0.301	0.214	0.751	0.750	1.158
	(0.053)	(0.053)	(0.034)	(0.194)	(0.194)	(0.273)
Thailand	0.442	0.431	0.570	1.052	0.972	1.736
	(0.057)	(0.055)	(0.073)	(0.195)	(0.181)	(0.322)
Togo	0.438	0.438	0.392			
	(0.138)	(0.138)	(0.101)			
Tunisia	0.436	0.419	1.255	0.971	0.818	1.308
	(0.086)	(0.082)	(0.246)	(0.280)	(0.247)	(0.378)
Turkey	0.176	0.174	0.370	1.057	0.999	1.068
	(0.016)	(0.016)	(0.034)	(0.138)	(0.130)	(0.139)
Turkmenistan	0.367	0.358	0.798			
	(0.139)	(0.135)	(0.302)			
Uganda	0.494	0.494	0.142	0.948	0.947	1.913
	(0.110)	(0.110)	(0.019)	(0.316)	(0.316)	(0.362)
Ukraine	0.495	0.492	0.826	1.060	1.033	1.044
	(0.057)	(0.056)	(0.095)	(0.174)	(0.170)	(0.172)
United Kingdom	0.829	0.701	0.934	1.126	0.916	1.137
	(0.073)	(0.062)	(0.083)	(0.142)	(0.115)	(0.143)
United States	0.310	0.245	0.499	0.874	0.761	0.852
	(0.017)	(0.014)	(0.028)	(0.069)	(0.060)	(0.068)
Uzbekistan	0.442	0.440	0.749	1.000	0.982	0.899
	(0.054)	(0.054)	(0.092)	(0.177)	(0.174)	(0.159)
Venezuela	0.260	0.241	0.801	0.971	0.807	0.949
	(0.035)	(0.032)	(0.107)	(0.187)	(0.155)	(0.183)
Vietnam	0.265	0.265	0.143	0.690	0.688	1.258
	(0.018)	(0.018)	(0.010)	(0.068)	(0.068)	(0.121)
Yemen	0.251	0.250	0.157	0.758	0.739	1.148
	(0.039)	(0.039)	(0.021)	(0.170)	(0.165)	(0.225)
Zambia	0.341	0.340	0.434	0.655	0.652	1.401
	(0.064)	(0.064)	(0.073)	(0.182)	(0.181)	(0.350)

Table D-1 – Hill Pareto Alpha Coefficient Estimates (Year 2000, Satellite F15)

	Full Distribution			Above Median Setting		
Country	Stable Lights	Corrected Lights	Pop.	Stable Lights	Corrected Lights	Pop.
Zimbabwe	0.207	0.207	0.972	0.746	0.746	1.139
	(0.046)	(0.046)	(0.198)	(0.249)	(0.249)	(0.343)

Note: Coefficients are only calculated for distributions with at least ten cities of non-zero city size. A few cities have a non-zero population but no detected light emissions. That can lead to a shorter city size distribution for lights than for population. Corrected standard errors according to $SE_{Hill} = \hat{\alpha}_{Hill}/\sqrt{N-3}$ (Gabaix, 2009) are given in parentheses.

E Additional data source: Citypopulation

For comparison purposes we also take a look at the often-used data www.citypopulation.de (CP) by Brinkhoff (2017). The website compiles census outcomes and official estimates mostly released by national statistical offices. The data is not geo-referenced and different city definitions are used in each country, so that it cannot be combined with our city identification scheme. We repeat our analysis on the city size distribution with the CP data mainly to replicate the results from the literature (Soo, 2005, Henderson and Wang, 2007) and to compare them with ours.[3] Table E-1 provides an overview of the data used, highlighting the superiority of the population data (and nighttime light data) compared to the CP.

The CP data faces some disadvantages that do not apply to our baseline GHSL data. It is a pure compilation of national census databases, therefore entailing various problems of data availability and comparability. First, the lower bound of cities size distributions varies drastically across countries. The Swiss data for the year 2000 counts 162 cities and towns with the smallest one hosting 5,447 residents. The corresponding table on France for the year 1999 is truncated at a much higher point and only provides information on 39 settlements with at least 85,832 inhabitants. Second, reference years differ. CP provides population information on multiple years. However, there is no standard on how many and which years are reported, impeding comparisons. Third, the CP data is not geo-referenced. Thus, we cannot compute the nighttime lights that fall within the areas of the CP cities. We also do not know to what extent the chosen (administrative) city boundaries cover the economically relevant agglomeration and to what degree variation in these choices drives the resulting Pareto alphas. Consider Spain as an example. The CP city size distribution in the year 2001 begins with a city of 50,096 inhabitants and contains 76 observations in total - close to GHSL with 50,000 and 75 observations. Despite this outstanding similarity the distributional shapes are not similar. With GHSL we obtain a Pareto alpha of 0.999, with CP one of 1.237. City size distributions constructed from administrative boundaries tend to overestimate equality because the largest cities have outgrown their administrative borders.
In Table E-1 we look at some more countries and compare the Pareto alpha based on our data set with the CP Pareto alpha, as well as the size of the smallest CP city. We see the large heterogeneity. For some countries, the CP estimates are similar than for our data set, for others they are not, and it is plausible that all of the factors described above contribute to these differences.

[3]The data on the CP website are continuously updated with information on smaller cities from previous years, leading to an increase in the number of cities. For the same years and countries, CP lists more cities than, for instance, when Soo (2005) carried out his study.

Table E-1 – CP-GHSL Comparison for Selected Countries

Country	GHSL Alpha (Full D.)	CP Alpha	Lower Bound (CP)
United States	0.841	1.231	14,676
Brazil	0.983	1.143	34,552
Bangladesh	0.742	1.118	12,660
Taiwan	0.654	0.629	9,531
France	1.050	1.359	85,832
Kenya	0.960	0.807	2,931
Colombia	0.831	0.872	10,032
Spain	0.999	1.237	50,096
Germany	0.975	1.239	47,382

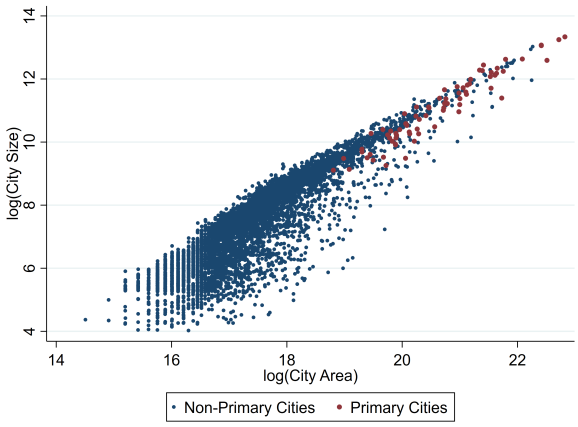
Note: Data refers to the year 2000 or the closest one available in the CP data. CP results are based on CP's "Cities and Towns" tables.

F Additional results on density and area

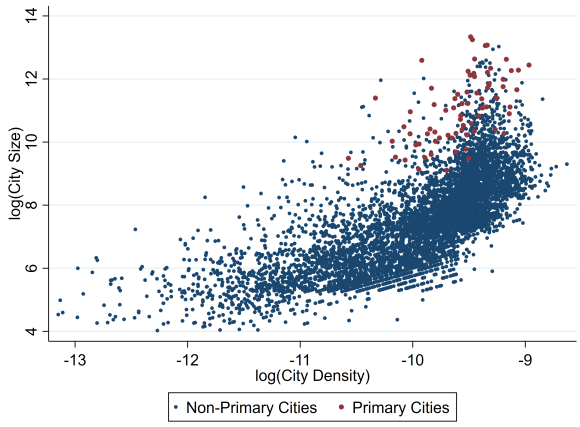
Here we provide some additional results on the decomposition of city size into area and density, supplementing the discussion in Section 4.2 in the paper.

Figure F-1 displays scatter plots about the city size (measured, respectively, in terms of population, 'stable' and correctly lights) and either city area or city density. Cities from countries around the world are pooled, with red dots indicating primaries cities. As regards city area (Figure F-1a, Figure F-1c, Figure F-1e), we note a clear positive association between city size of all three measures and area: More populous and brighter cities are more extended; and primary cities (red dots) are the most extended ones. There is more variation in the relation between size and density (Figure F-1b, Figure F-1d, Figure F-1f): Cities of the same size can have widely different densities, although on average brighter and more populous cities also tend to be denser. But the relation is less clear-cut than for area. This holds in particular when size is measured based on population. Overall, these scatter plots confirm our insights that size differences between cities are driven to a larger extent by area than by density.

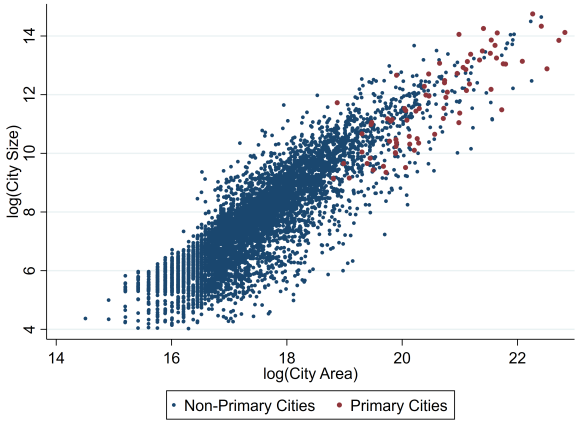
In the following, we take a closer look at the proportions of total size, area and density between the largest and second-largest city in each country. We present robustness checks to the analysis in the paper and show that the result of area rather than density being the dominant factor is not driven by (i) the country sample, and (ii) the year. Table F-1 shows the results using the latest available year, 2013 for light and 2015 for population, but unlike the calculations in the paper, we now take all countries that have two cities, including those countries with less than 10 cities so that we do not have Pareto alpha estimates. We see that the proportions of the largest to the second-largest cities become even more severe and area is the driving factor. The world's smallest countries often have a dominating primary city, so that our baseline sample in the paper provides a lower bound. We also see that on all continents, the proportions for light are larger than for population.
Table F-2 underlines that there is nothing particular about the chosen years at the end of the sample: Pooling the proportions between primary and secondary cities for all available years for all countries with at least two cities yields very similar results to the ones in the paper. This is in line with the large persistence in the city size distribution, which also kept the proportions between primary and secondary cities for most countries rather stable.



(a) Stable Light: Area (Year 2013)

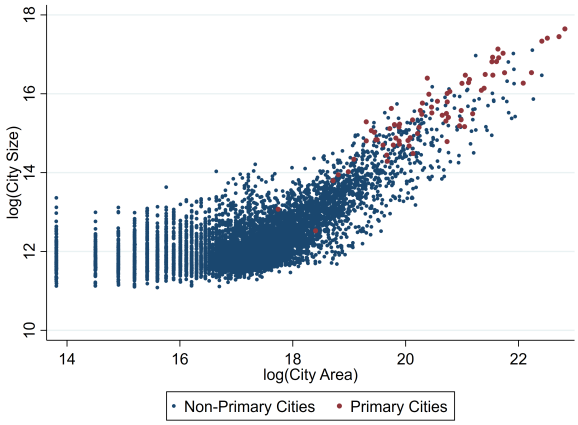


(b) Stable Light: Density (Year 2013)

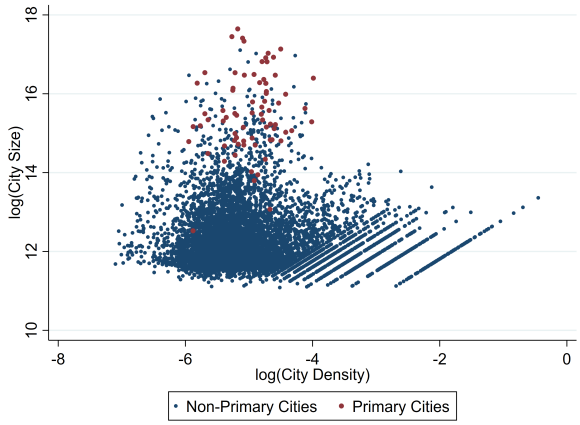


(c) Corrected Light: Area (Year 2013)

(d) Corrected Light: Density (Year 2013)



(e) Population: Area (Year 2015)



(f) Population: Density (Year 2015)

Figure F-1 – City Size Compared to City Area and City Size (Cities Relevant for Above Median Pareto Alpha Estimates Only)

Table F-1 – Comparing Primary and Secondary Cities in the most recent years 2013/2015

		World	Africa	Americas	Asia	Europe
Population	Size	4.754	5.279	5.373	4.673	3.797
		(4.076)	(4.469)	(3.175)	(4.996)	(2.319)
	Density	1.348	1.424	1.375	1.409	1.165
		(1.251)	(1.969)	(0.429)	(0.992)	(0.387)
	Area	4.352	5.413	4.268	3.984	3.476
		(3.639)	(4.367)	(2.936)	(3.917)	(2.076)
Stable Light	Size	5.785	9.729	4.346	4.246	3.500
		(9.910)	(16.503)	(2.771)	(4.308)	(2.152)
	Density	1.286	1.619	1.068	1.242	1.045
		(0.950)	(1.525)	(0.196)	(0.653)	(0.103)
	Area	4.244	5.575	3.985	3.723	3.346
		(3.381)	(4.299)	(2.222)	(3.342)	(1.997)
Corrected Light	Size	7.244	10.418	6.087	6.239	5.156
		(10.706)	(16.572)	(4.256)	(7.643)	(5.098)
	Density	1.657	1.735	1.493	1.842	1.433
		(1.176)	(1.518)	(0.573)	(1.273)	(0.737)
	Area	4.290	5.587	3.973	3.841	3.383
		(3.403)	(4.290)	(2.303)	(3.479)	(1.857)

The values are computed as $\frac{1}{N}\sum_{i=1}^{N}\frac{PrimaryCitySize_i}{SecondaryCitySize_i}$, $\frac{1}{N}\sum_{i=1}^{N}\frac{PrimaryCityDensity_i}{SecondaryCityDensity_i}$, $\frac{1}{N}\sum_{i=1}^{N}\frac{PrimaryCityArea_i}{SecondaryCityArea_i}$ with country i and N as the total number of countries on the respective continent. The respective standard deviations are denoted in parentheses. Asia includes Oceania. The lights data come from 2013, the population data from 2015.

Table F-2 – Comparing Primary and Secondary Cities (All Years)

		World	Africa	Americas	Asia	Europe
Population	Size	4.791	5.407	5.216	4.948	3.547
		(8.705)	(11.565)	(3.087)	(10.188)	(2.319)
	Density	1.879	1.434	4.157	1.823	1.063
		(11.278)	(2.052)	(27.499)	(5.459)	(0.368)
	Area	4.781	5.049	4.200	5.805	3.541
		(9.475)	(4.071)	(3.328)	(16.622)	(2.242)
Stable Light	Size	5.943	9.520	5.046	4.557	3.564
		(8.429)	(13.278)	(4.117)	(4.631)	(2.130)
	Density	1.465	2.097	1.176	1.268	1.067
		(2.722)	(4.762)	(0.445)	(0.632)	(0.137)
	Area	4.215	5.364	4.077	3.799	3.333
		(3.264)	(4.023)	(2.354)	(3.315)	(1.888)
Corrected Light	Size	6.923	9.817	6.626	5.918	4.598
		(8.976)	(13.293)	(5.126)	(6.887)	(3.633)
	Density	1.722	2.152	1.560	1.636	1.375
		(2.753)	(4.756)	(0.664)	(1.014)	(0.519)
	Area	4.459	5.784	4.078	4.105	3.426
		(4.290)	(5.099)	(2.460)	(5.016)	(2.083)

The values are computed as $\frac{1}{N \cdot T} \sum_{i=t}^{T} \sum_{i=1}^{N} \frac{PrimaryCitySize_{it}}{SecondaryCitySize_{it}}$, $\frac{1}{N \cdot T} \sum_{i=t}^{T} \sum_{i=1}^{N} \frac{PrimaryCityDensity_{it}}{SecondaryCityDensity_{it}}$, $\frac{1}{N \cdot T} \sum_{i=t}^{T} \sum_{i=1}^{N} \frac{PrimaryCityArea_{it}}{SecondaryCityArea_{it}}$ with country i, time t, N as the total number of countries in the respective region and T as the total number of time periods. The respective standard deviations are denoted in parentheses. Asia includes Oceania.

G Model selection procedure for determinants

This section provides further insights into the model selection approach discussed in Section 5.1, including an overview of the 36 variables considered together with their data sources. As mentioned in the paper, our simplistic algorithm tests all models with between one and seven regressors drawn from a pool of 36 variables and year fixed effects. These 36 variables originate from a variety of country characteristics. We label them into five categories: population structure, physical geography, institutions, economic structure and international connectedness (see Table G-1). The respective data sources are listed in Table G-2. Unlike earlier research we do not consider man-made transport infrastructure, such as road (Soo, 2005) or railway networks (Rosen and Resnick, 1980), due to simultaneous causality concerns. It is unclear whether an inegalitarian city size distribution, often associated with a dominant (coastal) primate city, is caused by a scant transport network or whether the absence of major hinterland cities renders an extensive transport system unnecessary. Instead of purely man-made infrastructure, we rely on terrain ruggedness and waterway density to account for intra-country connectedness. These factors are also modifiable by humans, though to a much smaller extent than roads and railways.

From our 36 potential regressors we obtain 10,739,175 combinations with between one and seven variables. In the paper, we discuss the specifications resulting in the lowest Akaike Information Criterion (AIC) and the lowest Bayesian Information Criterion (BIC) for the respective data set.

Although our strategy is a considerable improvement to the determinant identification in the existing Zipf literature, it comes with a number of caveats. First, we assume all variables to be linearly related to the outcome which might be inappropriate for some of the regressors. Second, missing values in the given variables do not only render a regression containing all 36 factors impossible but induce variation in the sample employed for our iterative procedure. This could induce sample selection bias and influence the resulting information criteria. Third, the maximum of seven explanatory factors - in addition to the year fixed effects - is chosen arbitrarily. The most suitable specification might be longer than that. Fourth, some variables may be too aggregated. For example, the influence of natural resources rents could vary between resources. Fifth, we count categorial variables as a single variable in the maximum of seven even though they come as multiple dummy regressors costing multiple degrees of freedom. Overall, we could come up with numerous extensions, modifications and robustness checks to explore the determinants of city size distributions even more profoundly. Options vary from designing a theoretical model to the adoption of random forests classification algorithms. However, that is beyond the scope of this paper and left for further research.

Table G-1 – Explanatory Variables in Model Selection

	Population Structure	Physical Geography	Institutions	Economic Structure	International Connectedness
Total population	x				
Population growth	x				
Urbanization	x				
Population in 1400	x				
Fertility	x		x		
Net migration	x		x		x
Ethnic fractionalization	x		x		
Terrain ruggedness		x			
Coastal proximity		x			
Coastal border		x			
Land area		x			
Malaria incidence		x			
Extreme weather		x			
Natural resource rents		x		x	
Border length		x			
Waterway density		x			
Latitude		x			
Continent		x			
Agricultural land		x			
Colonial heritage			x		
Financial development			x		
Fiscal centralization			x		
Government expenditure			x	x	
Democracy			x		
Interstate war			x		x
Political rights, civil liberties			x		
Time of independence			x		
Patent applications			x	x	
Trade				x	x
Exports				x	x
Energy use				x	
GDP				x	
GDP p.c.				x	
Agriculture				x	
Manufacturing				x	
Services				x	

Table G-2 – Explanatory Variables' Data Sources

Variable	Data source
Total population	World Development Indicators
Population growth	World Development Indicators
Urbanization	World Development Indicators
Population in 1400	Nunn and Puga (2012)
Fertility	World Development Indicators
Net migration	World Development Indicators
Ethnic fractionalization	Alesina et al. (2003)
Terrain ruggedness	Nunn and Puga (2012)
Coastal proximity	Nunn and Puga (2012)
Coastal border	CIA World Factbook
Land area	World Development Indicators
Malaria incidence	World Development Indicators
Extreme weather	World Development Indicators
Natural resource rents	World Development Indicators
Border length	CIA World Factbook
Waterway density	CIA World Factbook
Latitude	Nunn and Puga (2012)
Continent	World Development Indicators
Agricultural land	World Development Indicators
Colonial heritage	CEPII GeoDist
Financial development	Global Financial Development
Fiscal centralization	IMF Government Finance Statistics
Government expenditure	World Development Indicators
Democracy	Polity IV
Interstate war	Correlates of War
Political rights, civil liberties	Freedom House
Time of independence	ICOW
Patent applications	World Development Indicators
Trade	World Development Indicators
Exports	World Development Indicators
Energy use	World Development Indicators
GDP	World Development Indicators
GDP p.c.	World Development Indicators
Agriculture	World Development Indicators
Manufacturing	World Development Indicators
Services	World Development Indicators

Additional references

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003). Fractionalization. *Journal of Economic Growth 8*(2), 155–194.

Brinkhoff, T. (2017). City population. http://www.citypopulation.de [Accessed: December 2017 - February 2018].

Center for Systemic Peace (2016). Polity IV Project. http://www.systemicpeace.org/inscrdata.html [Accessed: July 2018].

Central Intelligence Agency (2018). The World Factbook. https://www.cia.gov/library/publications/the-world-factbook/index.html [Accessed: July 2018].

CEPII (2011). GeoDist Database. http://www.cepii.fr/CEPII/en/bdd_modele/presentation.asp?id=6 [Accessed: July 2018].

Cheshire, P. (1999). Trends in Sizes and Structures of Urban Areas. In P. Cheshire (Ed.), *Handbook of Regional and Urban Economics*, Volume 3, pp. 1339–1373. Elsevier.

Cirillo, P. (2013). Are your Data Really Pareto Distributed? *Physica A: Statistical Mechanics and its Applications 392*(23), 5947–5962.

Correlates of War Project (2011). COW War Data. http://correlatesofwar.org/data-sets/COW-war [Accessed: July 2018].

Freedom House (2018). Freedom in the World. https://freedomhouse.org/report-types/freedom-world [Accessed: July 2018].

Gabaix, X. (2009). Power Laws in Economic and Finance. *Annual Review of Economics 1*, 255–294.

Gabaix, X. and R. Ibragimov (2011). Rank - 1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents. *Journal of Business and Economics Statistics 29*(1), 24–39.

Henderson, J. and H. Wang (2007). Urbanization and City Growth: The Role of Institutions. *Regional Science and Urban Economics 37*(3), 283–313.

Hensel, P. R. (2018). Issue Correlates of War (ICOW) Project. http://www.paulhensel.org/icow.html [Accessed: July 2018].

Hill, B. (1975). A Simple General Approach to Inference About the Tail of a Distribution. *The Annuals of Statistics 3*(5), 1163–1174.

International Monetary Fund (2018). Government Finance Statistics. http://data.imf.org/?sk=a0867067-d23c-4ebc-ad23-d3b015045405 [Accessed: April 2018].

Nunn, N. and D. Puga (2012). Ruggedness: The Blessing of Bad Geography in Africa. *Review of Economics and Statistics 94*(1), 20–36.

Ravallion, M. and S. Chen (2011). Weakly Relative Poverty. *The Review of Economics and Statistics 93*(4), 1251–1261.

Rosen, K. and M. Resnick (1980). The Size Distribution of Cities - An Examination of the Pareto Law and Primacy. *Journal of Urban Economics 8*(2), 165–186.

Soo, K. (2005). Zipf's Law for Cities: A Cross-Country Investigation. *Regional Science and Urban Economics 35*(3), 239–263.

World Bank (2009). World Development Report 2009 : Reshaping Economic Geography. World bank.

World Bank (2018a). Global Financial Development. https://databank.worldbank.org/data/source/global-financial-development [Accessed: July - September 2018].

World Bank (2018b). World Development Indicators. https://databank.worldbank.org/data/reports.aspx?source=world-development-indicators [Accessed: July - September 2018].