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 GHSL derive the information on built-up area from Landsat image collections, i.e. GLS1975, GLS1990, GLS2000, and ad-hoc Landsat 8 collection 2013/2014, and population data from CIESIN GPWv4. 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] For our application it does, therefore, not matter how exactly the GHSL distribute and disaggregate census population within cities. Potentially incorrect assumptions on intra-urban distributions do not affect the sum of those pixel values. 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 of 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.

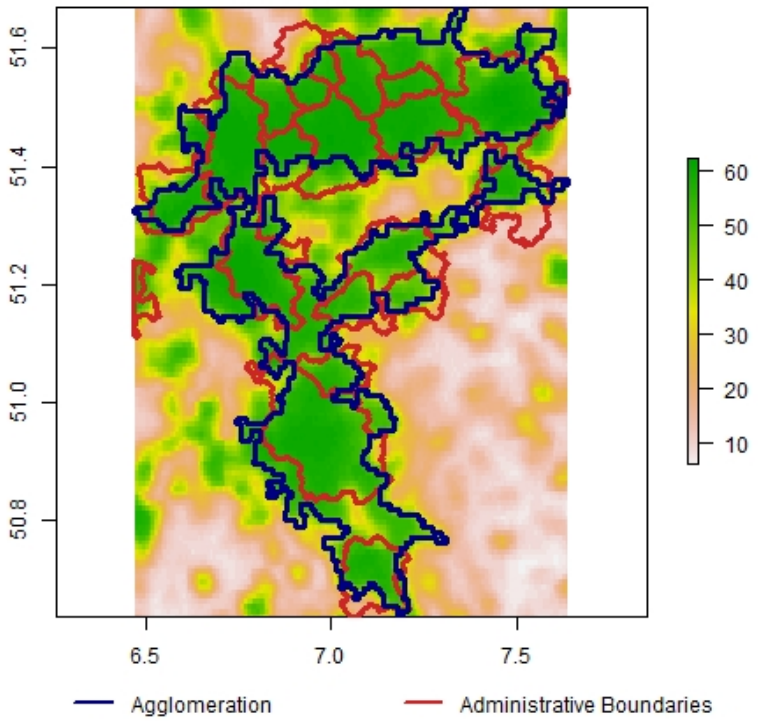
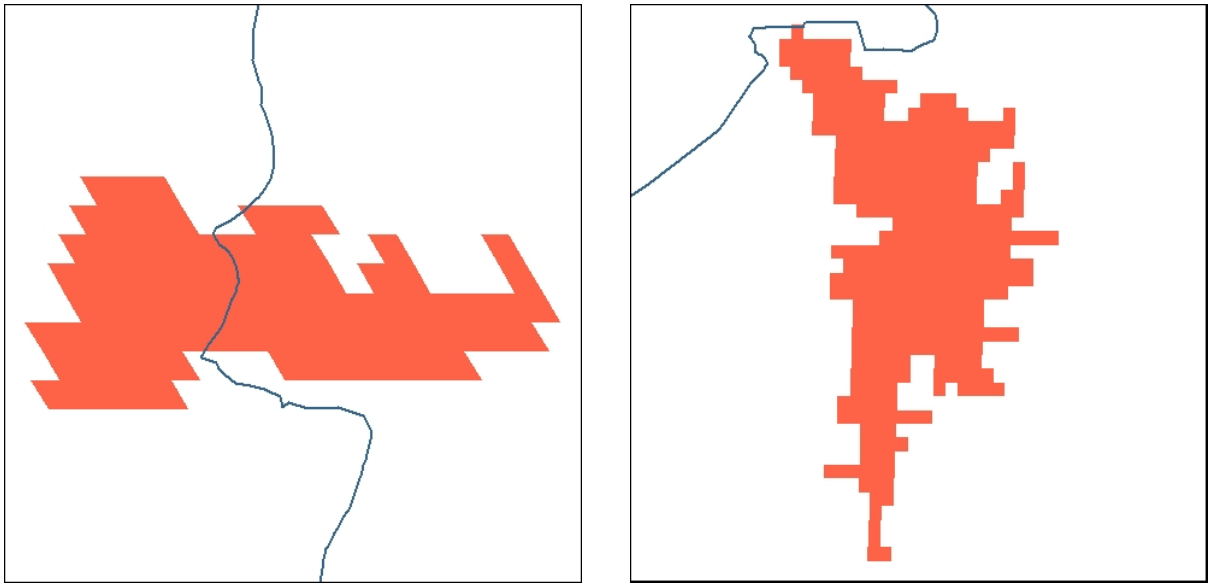


Figure A-1 – The Largest German Agglomeration (Year 2000, Satellite F15, Administrative Boundaries of Urban Municipalities Only)

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.

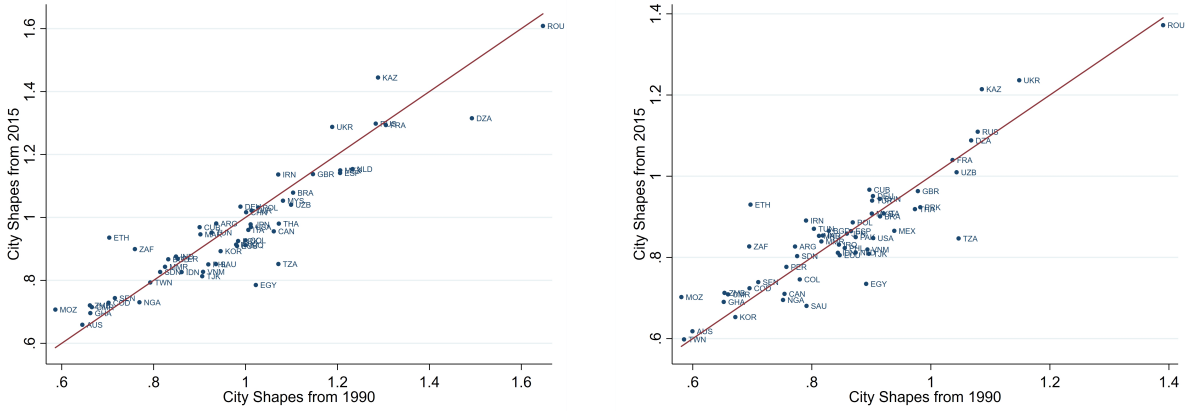
robust. In total, 264 out of the 13,844 cities in the data set are affected by this issue; 106 countries have at least one such border city. 137 out of the 264 cities are assigned to a single country.



(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 respectively 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. Any statistical identification of Zipf's law is misspecified if the underlying sample is not Pareto distributed. Correctly dissecting the lognormal body from the Pareto tail is therefore a necessary first step. 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.

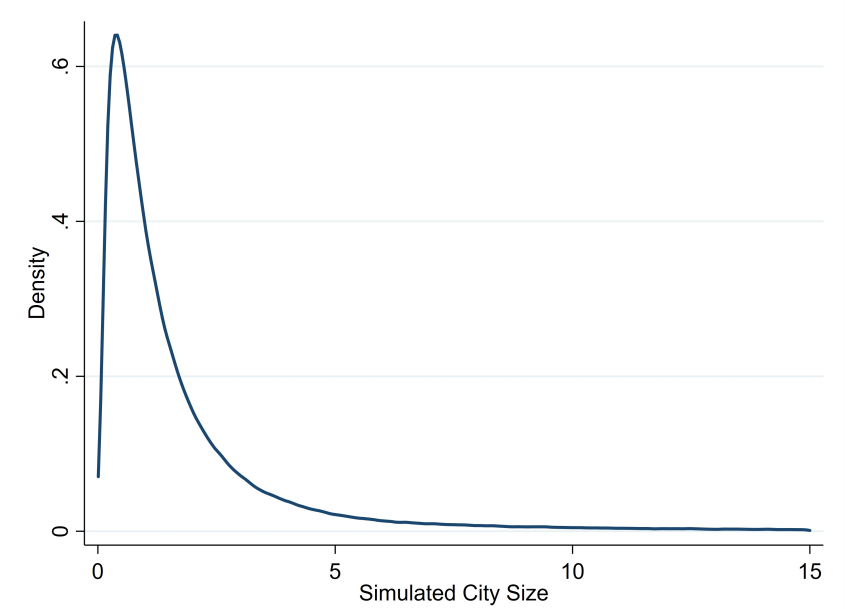


Figure B-1 – Kernel Density Plot of Simulation Outcome (Zoom on $x < 15$)

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?

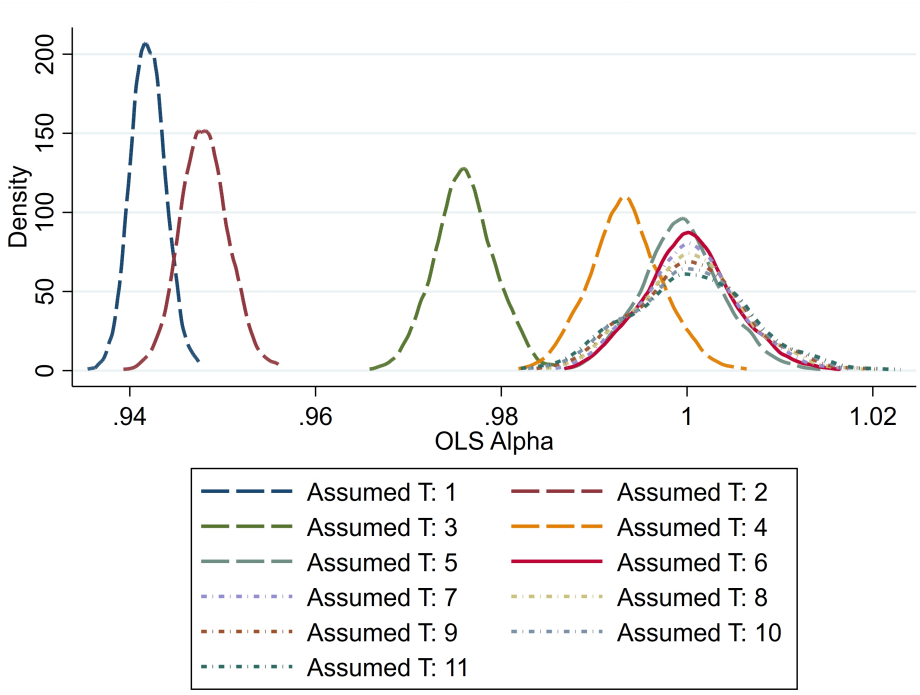


Figure B-2 – Comparison of Assumed Thresholds (True T = 6)

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 of observations 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 identify 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.

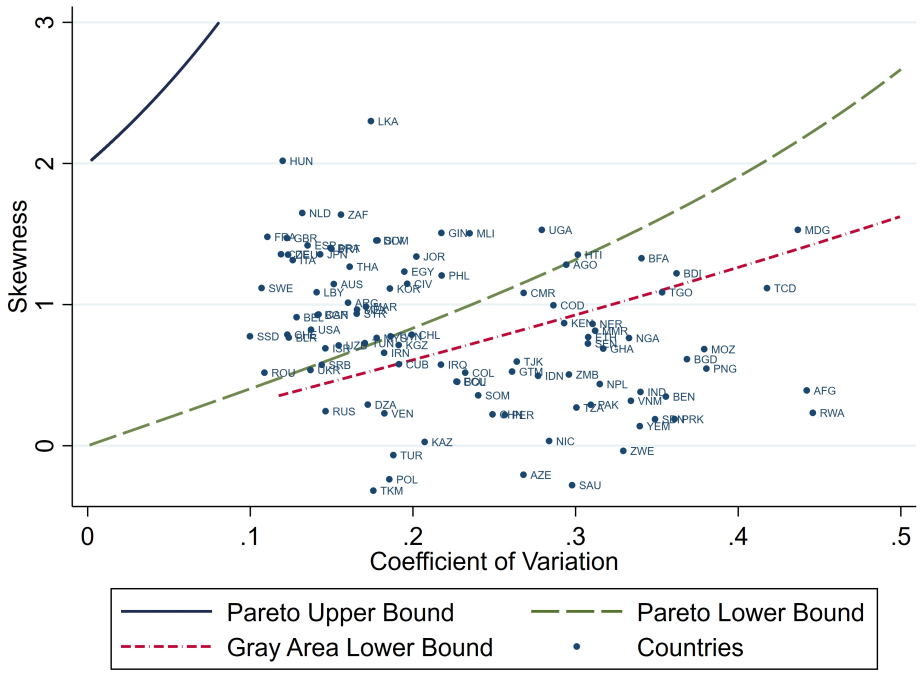


Figure B-3 – Discriminant Moment Ratio Plot (Log Corrected Light, Full Distribution, Year 2000, Satellite F15)

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.

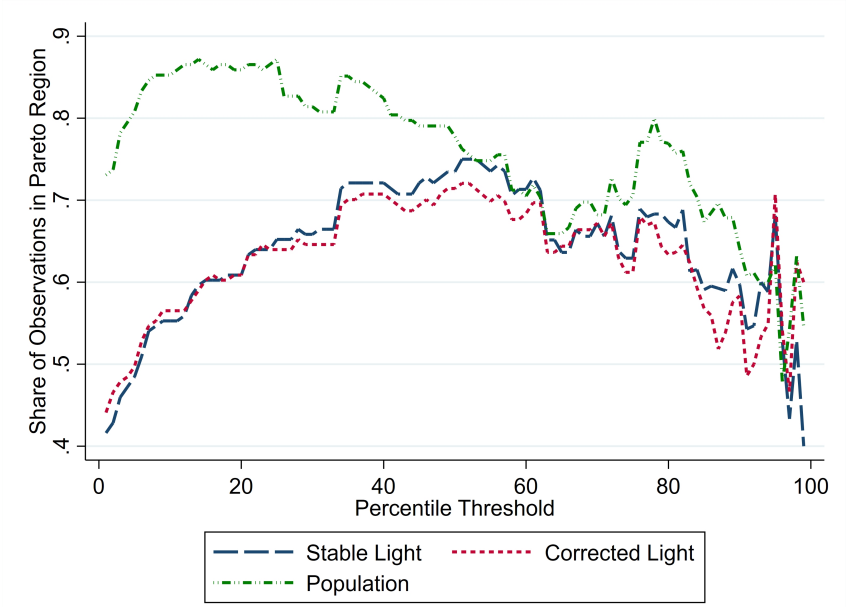


Figure B-4 – Share of Countries in DMRP's Pareto Region (Year 2000, Satellite F15)

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 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 including the top 50% of cities above

50,000 inhabitants into the Pareto alpha estimation is the optimal threshold choice.

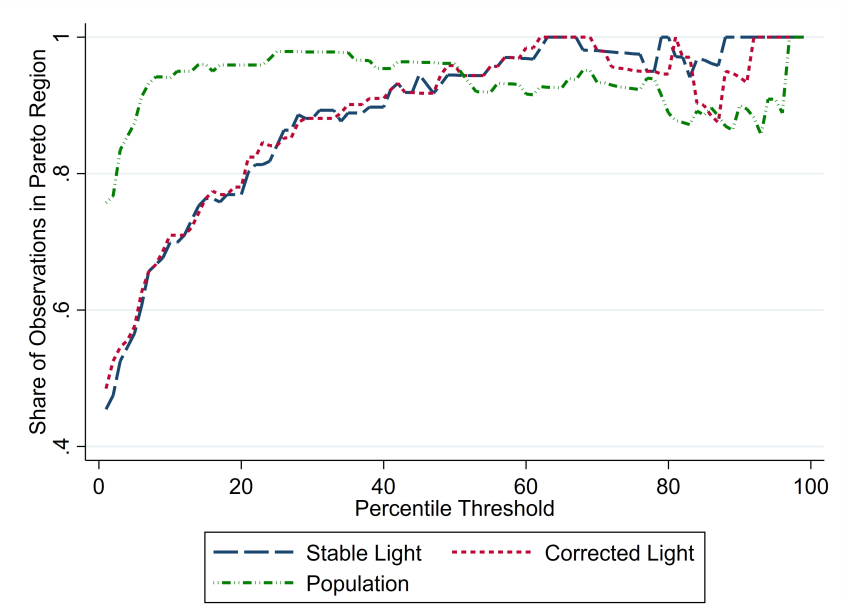
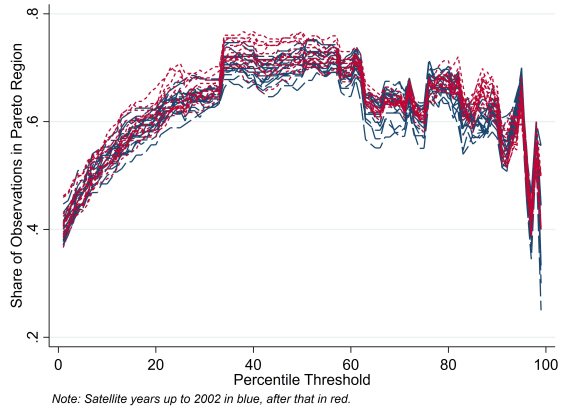
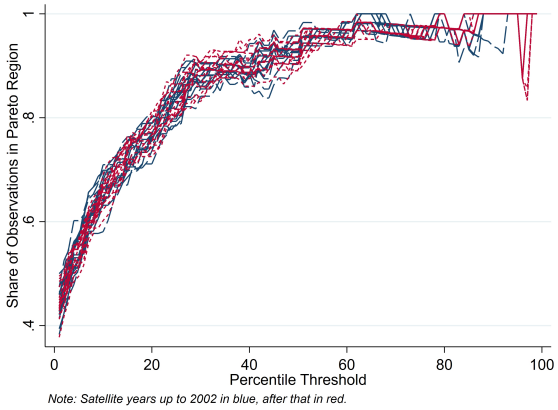


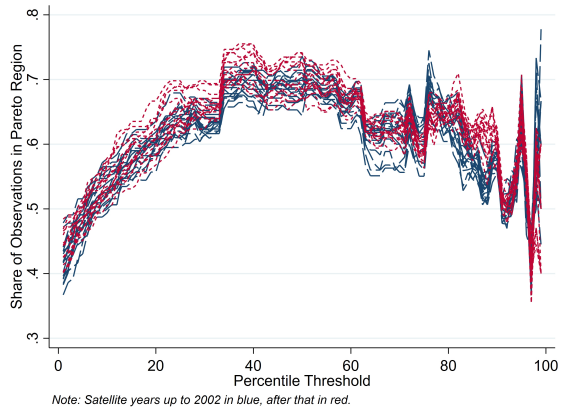
Figure B-5 – Share of Countries in DMRP's Pareto Region (Min. 10 Cities above Threshold, Year 2000, Satellite F15)



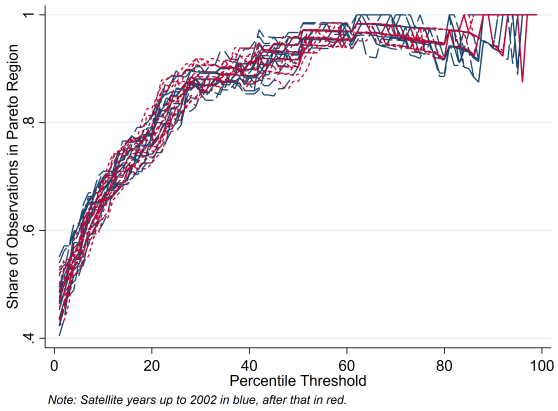
(a) Stable Light



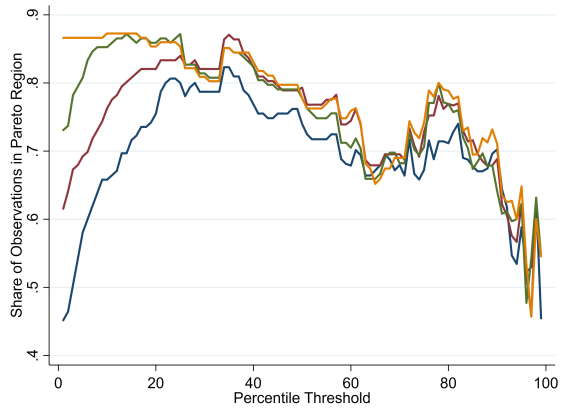
(b) Stable Light (Min. 10 Cities)



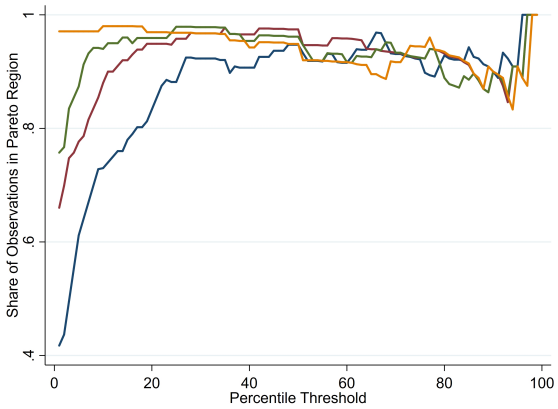
(c) Corrected Light



(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. The threshold T_{it} of country i at time t is 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.

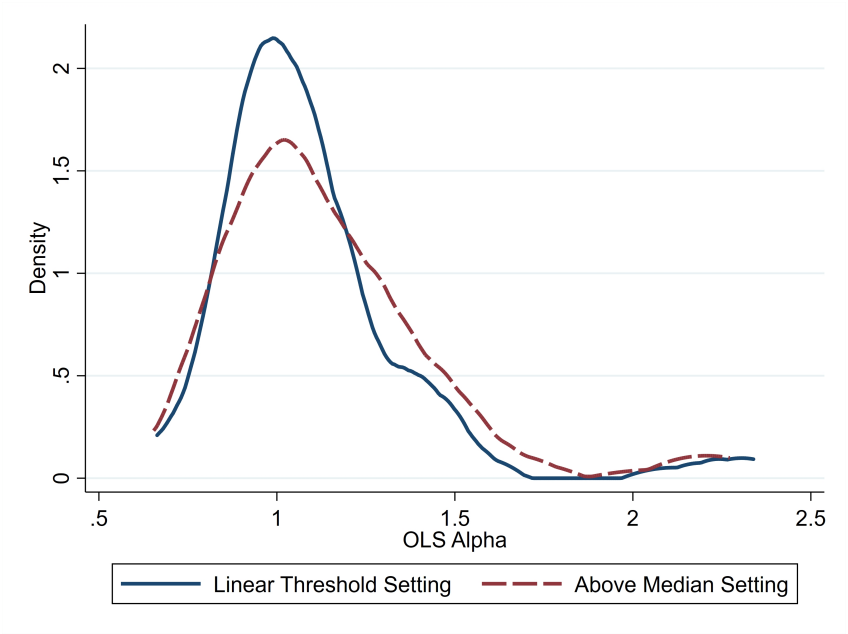


Figure B-7 – Comparison of Threshold Settings (Year 2000)

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 country 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)			

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.
Brazil	0.986	0.846	0.983	1.078	0.901	0.963
	(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)

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.
El Salvador	0.770	0.706	0.898			
	(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)

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.
Jordan	0.699	0.599	0.709			
	(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)

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.
North Korea	0.518	0.517	1.196	0.829	0.817	1.170
	(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)

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.
Spain	1.085	0.803	0.999	1.141	0.865	1.035
	(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)

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.
United States	0.852	0.698	0.841	0.971	0.850	0.882
	(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 & 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.

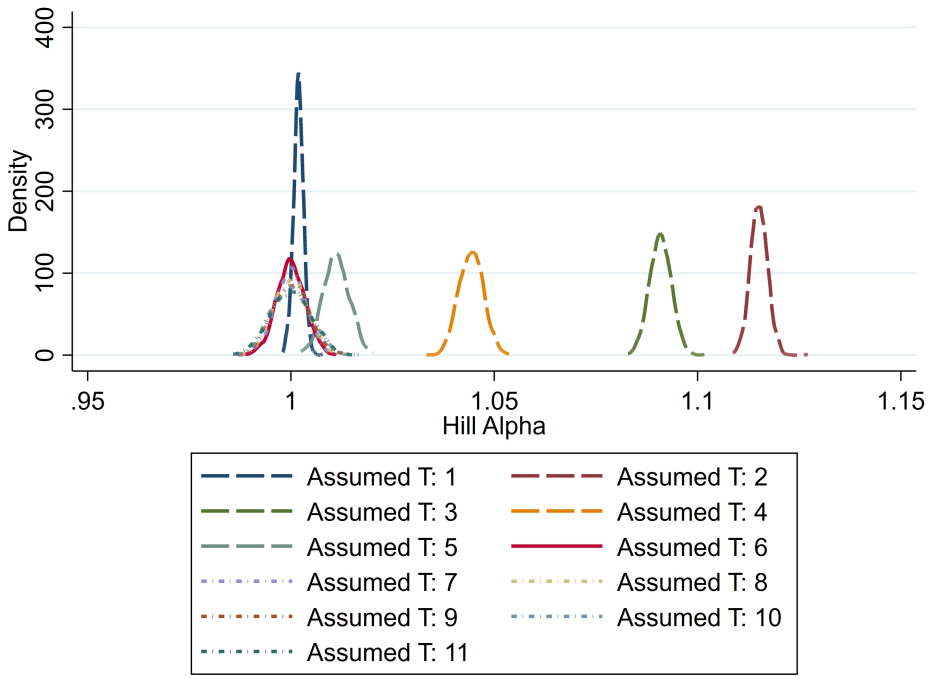


Figure D-1 – Hill Estimator: Comparison of Assumed Thresholds (True T = 6)

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

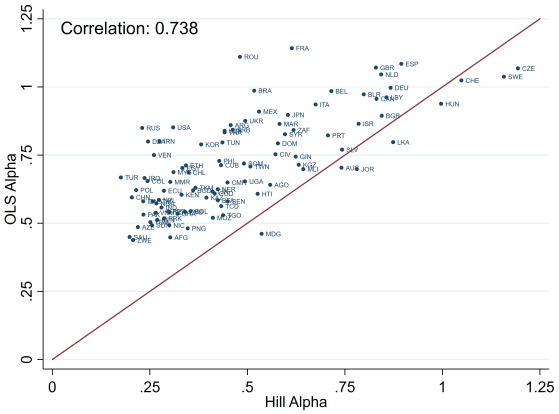
for Hill than for OLS. Hence, although the Hill estimator may be more efficient when the data is indeed Pareto distributed, its larger bias in cases of deviations from the Pareto distribution is a considerable drawback. Still, the Hill coefficient estimates provide a useful robustness check for our main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

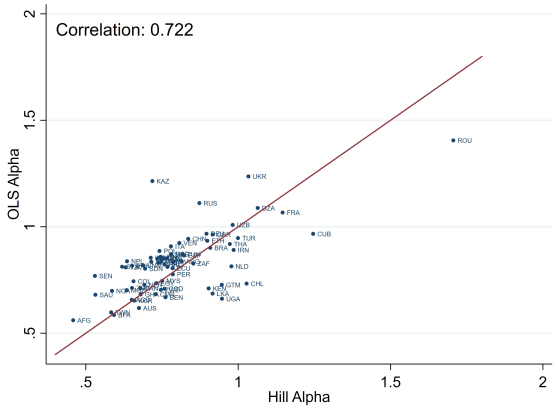
The scatter plots on the relation between the OLS and Hill alpha estimates for all measures (population, "stable", and corrected light) and both the above-median and full-distribution setting are shown in Figure D-2. We note strong correlation coefficients and see that, in particular in the above-median 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)

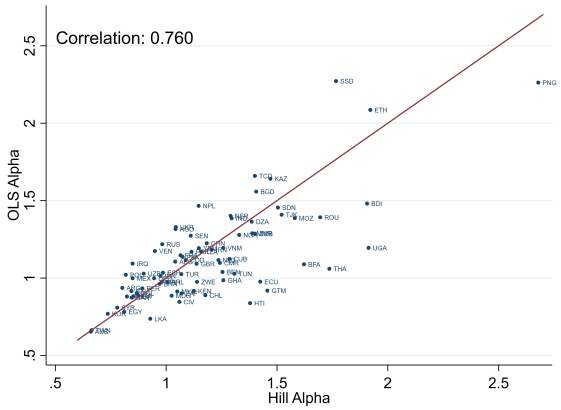


(b) Stable Light (Full Distribution Setting)

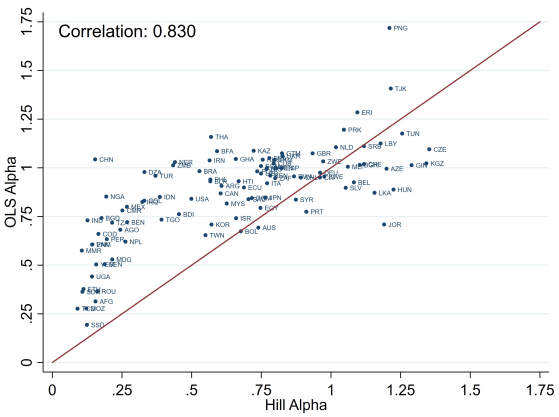


(c) Corrected Light (Above Median Setting)

(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)

Country	Full Distribution			Above Median Setting		
	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 (Henderson & Wang, 2007; Soo, 2005) 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 to our data set, for others they are not, and it is plausible that all of the factors described above contribute

[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.

to these differences.

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

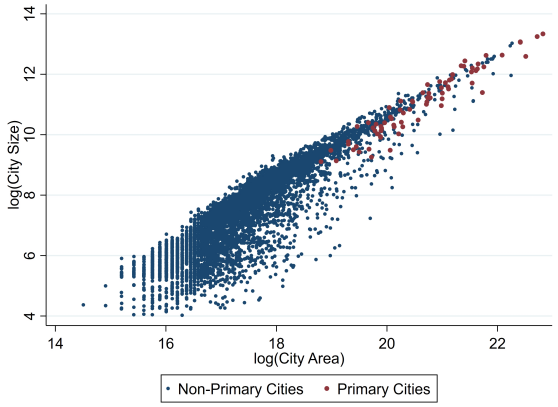
Note: Data refers to the year 2000 or the closest one available in the CP data. CP results are based on CP's "Cities and Towns" tables.

F ADDITIONAL RESULTS ON DENSITY AND AREA

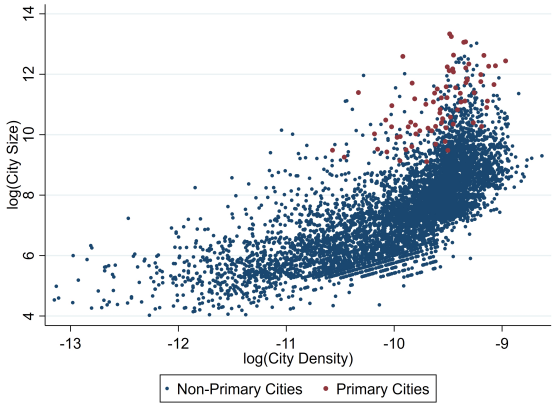
Here we provide some additional results on the decomposition of city size into area and density, supplementing the discussion in Section 4.2 in the paper.

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 spatially 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.

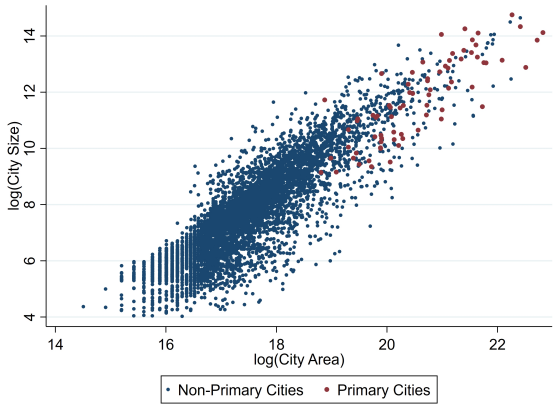
In the following, we take a closer look at the proportions of total size, area and density between the largest and second-largest city in each country. We present robustness checks to the analysis in the paper and show that the result of area rather than density being the dominant factor is not driven 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 in the calculations in the paper, we now take all countries that host at least two cities, including those countries with less than 10 cities for 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.



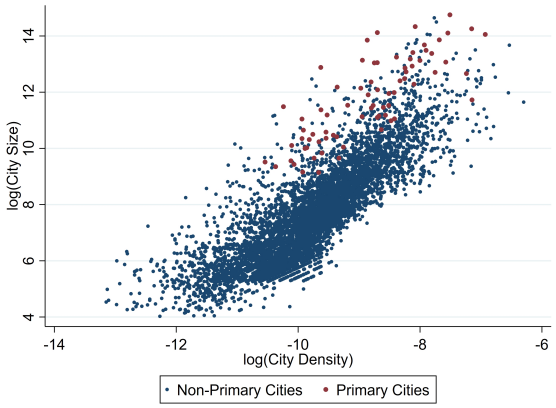
(a) Stable Light: Area (Year 2013)



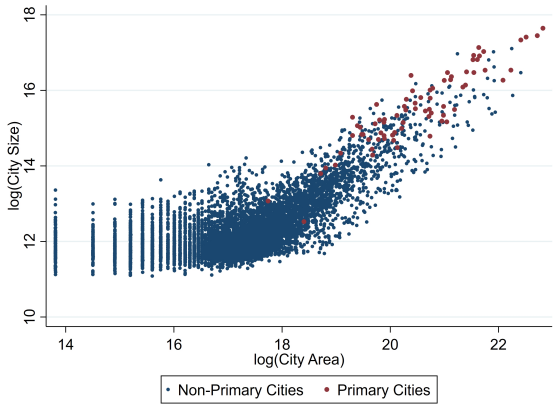
(b) Stable Light: Density (Year 2013)



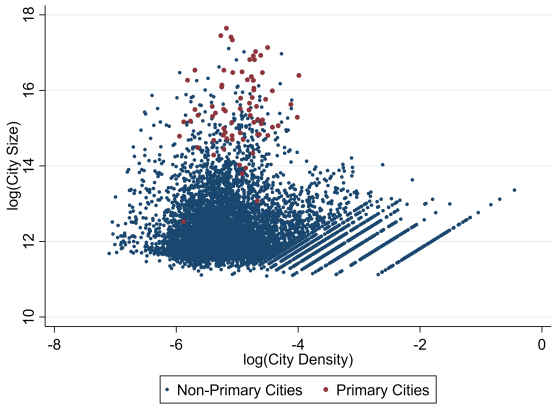
(c) Corrected Light: Area (Year 2013)



(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 Density (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 & 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 categorical 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 lasso regressions and random forests 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, Devleeschauwer, Easterly, Kurlat, and Wacziarg (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., Devleeschauwer, A., Easterly, W., Kurlat, S., & Wacziarg, R. (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* (Vol. 3, p. 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., & Ibragimov, R. (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., & Wang, H. (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., & Puga, D. (2012). Ruggedness: The Blessing of Bad Geography in Africa. *Review of Economics and Statistics*, *94*(1), 20-36.

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

Rosen, K., & Resnick, M. (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])