24 - 26 | Noviembre 2021 | Madrid
XLVI Reunión de Estudios Regionales
International Conference on Regional Science
Ciudades llenas, territorios vacíos
Universidad Autónoma de Madrid

COMUNICACIÓN

Título: Comparing city size distributions: Population vs. economic activity

Autores y e-mails de todos:
Miguel Puente-Ajovín[a] (mpajovin@unizar.es)
Marcos Sanso-Navarro[a] (marcossn@unizar.es)
María Vera-Cabello[b] (mvera@unizar.es)

Departamento:
[a] Departamento de Análisis Económico
[b] Departamento de Economía y Gestión

Universidad:
[a] Universidad de Zaragoza
[b] Centro Universitario de la Defensa de Zaragoza

Área Temática:
1. Growth, convergence and regional and urban development

Resumen: *(máximo 300 palabras)*

This paper compares the size distribution of cities when they are measured in both demographic and economic terms. In doing so, we have exploited more recent and accurate nighttime lights data than those previously used to proxy urban economic activity. Our results for a sample covering 12,852 urban centres in 100 countries show that economic activity is more unevenly distributed than population. There is weak evidence for Zipf's law in aggregate urban night lights, especially in medium and low income countries. Our findings do not support a Pareto distribution for city sizes measured in economic terms.

Palabras Clave:
City size distribution; Population; Nighttime lights; Zipf's law.

Clasificación JEL:
O10, O18, O57, R12

Comparing city size distributions: Population vs. economic activity[*]

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

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

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

mpajovin@unizar.es, marcossn@unizar.es, mvera@unizar.es

September 29, 2021

Abstract

This paper compares the size distribution of cities when they are measured in both demographic and economic terms. In doing so, we have exploited more recent and accurate nighttime lights data than those previously used to proxy urban economic activity. Our results for a sample covering 12,852 urban centres in 100 countries show that economic activity is more unevenly distributed than population. There is weak evidence for Zipf's law in aggregate urban night lights, especially in medium and low income countries. Our findings do not support a Pareto distribution for city sizes measured in economic terms.

Keywords: City size distribution; Population; Nighttime lights; Zipf's law.

JEL classification: O10, O18, O57, R12.

---

[*] This work has received financial support from Gobierno de Aragón (S39-20R ADETRE Research Group), Ministerio de Economía, Industria y Competitividad (Grant PID2020-112773GB-I00), and Fundación Ibercaja-Universidad de Zaragoza (Grant JIUZ-2020-SOC-11).

[§] Corresponding author. Address: Facultad de Economía y Empresa. Gran Vía 2. 50005 Zaragoza, Spain. Tel.: (+34) 876 554 629.

1 Introduction

There is a well-established link between population, concentration, and economic activity at the urban level that, due to its theoretical and policy-making implications, motivates the study of the city size distribution. Following the seminal contributions in Gabaix (1999) and Eeckhout (2004), the related literature has mostly focused on testing whether the city size distribution fits the rank-size rule, also known as Zipf's law (Rosen and Resnick 1980). This empirical regularity quantifies the concept of urban hierarchy by stating that the size of the N-th city is 1/N times the size of the largest one. As pointed out by Arshad, Hu, and Ashraf (2018), Zipf's law is not universal, even if only the upper tail of the city size distribution is considered. The mixed evidence regarding the rank-size rule becomes especially apparent when the urban structures of different countries are analyzed, see Soo (2005) and Puente-Ajovín, Ramos, and Sanz-Gracia (2020) for international comparisons. Nonetheless, a shortcoming commonly found in these cross-country studies is that the definition of what is considered as a city differs across national data sources. Actually, this issue may lead to conflicting results even within a single country (Fazio and Modica 2015; Ioannides and Skouras 2013; Puente-Ajovín et al. 2020).

Despite the relevance of the city size distribution from an urban economics point of view, and conditioned by data availability, most studies dealing with this topic measure them in demographic terms. The main exception is the analysis carried out by Düben and Krause (2021) who, following Chen and Nordhaus (2011) and Henderson, Storeygard, and Weil (2012), use nighttime lights data compiled by satellites to proxy urban economic activity. The main conclusion drawn by Düben and Krause (2021) is that while the distribution of urban population can be characterized by Zipf's law in most countries, this is not the case of economic activity. To carry out their empirical analysis, these authors use the data set created by Bluhm and Krause (2020) to correct the top-coding problem of the 'stable night light images' collected by the Defense Meteorological Satellite Program (DMSP) Operational Linescan System. At this point, it is also worth noting that these nighttime lights data are also affected by blurring, geo-location errors, lack of calibration, and coarse resolution; see Gibson (2021) and Gibson et al. (2021).

Since April 2012, there are available more precise nighttime lights captured by the Visible Infrared Imaging Radiometer Suite (VIIRS) of instruments onboard the Suomi NPP satellite. The VIIRS Day/Night Band was designed to measure the radiance of lights on earth in a wide variety of lightning conditions and covers a dynamic range of about seven orders of magnitude (DMSP covers less than two), avoiding saturation problems and top-coding. VIIRS nighttime lights are comparable over time and space, do not have blurring or geo-location errors, and display, at least, 45 times greater spatial resolution than DMSP data (Elvidge et al. 2017). For all these reasons, VIIRS data are superior at attributing lights to the place where they are emitted and, therefore, are a better proxy for urban economic activity than DMSP images; see Gibson, Olivia, and Boe-Gibson (2020) for a comparison of these two alternative nighttime lights data sources.

Taking into account previous arguments, the aim of this paper is to contribute to the literature that compares the distributions of urban population and economic activity. Similarly to Puente-Ajovín, Sanso-Navarro, and Vera-Cabello (2021), we will proxy local economic activity using nighttime lights captured by VIIRS. By proceeding in this way, and as a byproduct of our analysis, we will check the suitability of the top-coding correction of DMSP data based on the assumption of a Pareto distribution proposed by Bluhm and Krause (2020). In addition, we will assess how the Pareto assumption affects the evidence regarding the relevance of agglomeration economies at the country level. Finally, we will also explore the distributions of urban population and economic activity by groups of countries classified according to their income level.

2 Data

The first key issue when carrying out cross-country analyses of the distribution of urban size is to adopt a homogeneous definition for cities. For the sake of comparability, we have considered the urban centres delimited by the Global Human Settlement Layer (GHSL) database, provided by the Joint Research Center of the European Commission, as our unit of analysis; see Florczyk et al. (2019a) and Florczyk et al. (2019b). The GHSL defines urban centres consistently across geographical locations as areas with contiguous grid cells of one square kilometer, where each cell has, at least, 1,500 inhabitants or 50 per cent built-up surface. The main limitation of this database is that it only includes areas with more than 50,000 inhabitants. Together with the spatial information of urban centres, we have also extracted the georreferenced population by one square kilometer grid cells for the year 2015 from the GHSL. Given that some areas belong to more than one country, we have proceed in a similar way to Düben and Krause (2021) and assigned an area to a single country when it includes more than 75 per cent of the area. Applying this criterion, as well as only considering countries with more than 10 observations, our sample covers 12,852 urban centres of 100 countries.

The second relevant issue when dealing with urban size is its measurement. The interest in analyzing this topic is motivated by the fact that cities not only concentrate a large share of the population of a given country, but also of its economic activity. Moreover, the urban structure is the outcome of the dynamic interplay between economic activity and the growth process of cities (Arshad, Hu, and Ashraf 2018). Against this background, the great majority of studies about the distribution on city sizes measure them using population data, taking for granted that its location determines the economic landscape. The main reason is that it was difficult to find information about economic outcomes at the urban level and that, when available, it was not comparable across countries. Following Chen and Nordhaus (2011) and Henderson, Storeygard, and Weil (2012), this problem has been circumvented by Düben and Krause (2021) using nighttime lights to proxy economic activity.

In line with Puente-Ajovín, Sanso-Navarro, and Vera-Cabello (2021), and as suggested by Gibson (2021) and Gibson et al. (2021), we have used the current and more precise VIIRS nighttime lights. They have been extracted from the 'vcm-orm-ntl' annual composites[1] for 2015, available at the website of the Earth Observation Group of the National Oceanic and Atmospheric Administration (U.S. Department of Commerce)[2]. This data have been cleaned to exclude background noise, solar and lunar contamination, cloud cover degradation, and features unrelated to electric lighting (Elvidge et al. 2017). At the pixel level, reported radiance values are expressed in nano Watts per square centimeter per steradian, with a resolution of 15 arc seconds (approximately 450 meters at the equator). In the same manner as population, nighttime lights data have been aggregated for all pixels included within urban extents to calculate their size.

For comparison purposes, we have also calculated city sizes in economic terms using the 'stable night light images' provided by the DMSP, despite their limitations. Given that the production of DMSP lights ended in 2013, we have used the data for this year. In addition, the top-coding correction of DMSP data proposed by Bluhm and Krause (2020) has been used to provide a broad perspective of all nighttime lights data sources available, and to check the robustness of the results about the city size distribution in economic terms to their choice. By proceeding in this way, we will also be able to assess the suitability of the Pareto assumption adopted by Bluhm and Krause (2020) to generate their corrected data set[3].

3 Methodology

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

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

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

Gan, Li, and Song (2006) proposed an alternative approach to investigate urban size distributions based on the implementation of the Kolmogorov-Smirnov (KS) test statistic. This nonparametric method can be used to compare the distribution of city sizes with a function of reference, determining the degree of (dis)similarity. We have considered two references: (i) a Pareto function imposing that the power law exponent is equal to one, i.e. the exact Zipf's law; and (ii) a Pareto function with the estimated β coefficient in expression (1) as the power law exponent.

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

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

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

The Pareto distribution function is given by:

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

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

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

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

4 The distribution of urban population and economic activity

Figure 1 shows kernel densities for the estimated slope parameter in expression (1) at country level, using four alternative ways of calculating city sizes. The density functions of Pareto coefficients when size is expressed in demographic terms (orange) and when it is proxied using VIIRS data (blue) are displayed using solid lines. Estimated slope parameters are centered around values slightly higher than one when city size is calculated using information about the number of inhabitants. However, Pareto coefficients tend to be lower than one when size is expressed in economic terms. Therefore, and corroborating the findings of Düben and Krause (2021) and Puente-Ajovín, Sanso-Navarro, and Vera-Cabello (2021), aggregate urban nighttime lights are more unequally distributed than population. Focusing on the estimation results from the three alternative satellite imagery, it can be observed that density functions for DMSP and VIIRS data are quite similar, while that for the corrected data by Bluhm and Krause (2020) is more leptokurtic. This implies that the top-coding correction has a non-negligible influence on the estimated parameters from rank-size rule regressions.

[Insert Figure 1 about here]

The main aim of the empirical model in expression (1) is to test the null hypothesis that the Pareto coefficient is equal to one; i.e. that Zipf's law holds. Following Gan, Li, and Song (2006), we have implemented the KS test against the null hypothesis that city sizes at the country level are distributed as a Pareto function with power law exponent equal to one. The cumulative distribution function of the p-values that have been obtained for the four measures of city size are plotted in Figure 2. In line with the kernel density estimates of Pareto coefficients shown in Figure 1, the null hypothesis that city sizes adjust to Zipf's law can be more easily rejected when they are measured in economic terms. In fact, rejection rates are similar for the three nighttime lights data sources. The KS test has also been performed using the OLS estimate for the slope parameter in (1) as the Pareto coefficient. The cumulative distribution functions displayed in Figure 3 show that, although there is a slightly higher evidence of a Pareto distribution for aggregate urban night lights, the null hypothesis can be rejected in more than 70 cases at the 1% significance level. These findings suggest that the Pareto assumption established by Bluhm and Krause (2020) to correct for top-coding in DMSP data may not be correct. Nonetheless, this problem mainly affects bigger cities, which tend to be located in more developed countries. For this reason, it seems to be of interest to assess city size distributions by country income group.

[Insert Figures 2 and 3 about here]

5 Analysis by country income group

To create country groups, we have used the criterion established by the World Bank for the year 2015. According to this classification, countries are categorized as 'Low income' if their Gross National Income (GNI) per capita is lower or equal than 1,025 U.S. Dollars (22 out of 100 countries in our sample); 'Lower-middle income' if it is between 1,026 and

4,035 USD (29); ‘Upper-middle income’ between 4,036 and 12,475 USD (27); and ‘High income’ if GNI per capita is higher than 12,475 USD (22). Kernel density estimates of Pareto coefficients by country income group are plotted in Figure 4. The higher similarity between the distributions of urban population and economic activity is found for high income countries. Nonetheless, VIIRS nighttime lights and, especially, DMSP-corrected data are more unevenly distributed than population. There is a direct relationship between the similarity between the distributions of population and economic activity and the income level. In particular, estimated Pareto coefficients for population (night lights) tend to increase (decrease) when GNI per capita decreases. The highest similarity for estimated slope parameters of rank-size rule regressions for urban economic activity is found in lower-middle income countries. This reflects that this group is less affected by the top-coding problem of DMSP nighttime lights. However, and even if this is expected to also be the case of low income countries, kernel densities of estimated Pareto coefficients for VIIRS and DMSP-based data are different. This implies that the higher accuracy of VIIRS images also affects the conclusions drawn about the city size distribution in less developed countries.

[Insert Figure 4 about here]

Table 1 reports, at different levels of significance, the percentage of rejections by the KS test of the null hypothesis that the city size distribution is a Pareto function with power law exponent equal to one. Similarly to Figures 2 and 3, there is more evidence against the fullfilment of Zipf’s law in the urban distribution of economic activity than of population when all countries in our sample are analyzed. Broadly speaking, high income countries tend to display lower rejection rates. With the exception of upper-middle income countries, there is higher evidence against Zipf’s law in urban economic activity when it is measured using VIIRS nighttime lights than when it is proxied with DMSP-based data. Table 2 show similar results when the KS test statistic is performed considering that the reference distribution is a Pareto function with the estimated slope coefficient in the rank-size regression as the power law exponent. In this case, and as expected, the evidence of a Pareto distribution is slighlty higher than that for the exact Zipf’s law both for urban population and economic activity. Nonetheless, the rejection rates for aggregate

VIIRS night lights at the city level – higher that 50 per cent – do not support the Pareto assumption established by Bluhm and Krause (2020) to correct DMSP data.

[Insert Tables 1 and 2 about here]

6 Concluding remarks

This paper compares the distribution of urban population and economic activity, proxied using more accurate nighttime lights data than those used to date. The sample that has been analyzed covers 12,852 urban centres in 100 countries of different levels of development. Our results suggest that aggregate urban night lights are more unequally distributed than population. Furthermore, the null hypothesis that city sizes adjust to Zipf's law can be more easily rejected when they are measured in economic terms. We show that there is a higher similarity between the distributions of urban population and economic activity the higher the level of national income per capita. We also provide evidence that casts doubt on the Pareto assumption adopted to correct top-coding problems in some nighttime lights data sources.

The empirical analysis carried out can be extended in several directions. First, alternative definitions of cities, such as functional urban areas, can be adopted. Second, it could be of interest to analyze the influence of the spatial distribution of cities using appropriate – spatial econometrics – estimation techniques (Bergs 2021). Finally, the combination of the urban definitions provided by GHLS with spatial data of the built environment will allow us to study the distribution of other urban dimensions (Lall et al. 2021). These alternative extensions are in the research agenda of the authors.

References

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

Bergs, Rolf. 2021. "Spatial dependence in the rank-size distribution of cities – weak but not negligible". *PLOS ONE* 16 (2): 1–16. doi:`10.1371/journal.pone.0246796`.

Bluhm, Richard, and Melanie Krause. 2020. *Top lights: Bright cities and their contribution to economic development.* SoDa Laboratories Working Paper No 2020-08. Monash University.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lall, Somik, Mathilde Lebrand, Hogeun Park, Daniel Sturm, and Anthony Venables. 2021. *Pancakes to pyramids: City form to promote sustainable growth.* 159710. Washington, DC: World Bank.

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

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

Puente-Ajovín, Miguel, Marcos Sanso-Navarro, and María Vera-Cabello. 2021. “The distribution of urban population and economic activity in the European Union and the United States”. *Letters in Spatial and Resource Sciences* forthcoming. doi:`10.1007/complete`.

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

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

Tables and figures

Table 1: Kolmogorov-Smirnov test: Exact Zipf’s law. Percentage of rejections at different significance levels.

	Population			VIIRS		
	1%	5%	10%	1%	5%	10%
All countries	16.00	30.00	37.00	85.00	88.00	92.00
High income	0.00	9.09	18.18	63.64	77.27	81.82
Upper-middle	11.11	22.22	25.93	81.48	81.48	88.89
Lower-middle	17.24	44.83	55.17	96.55	96.55	100.00
Low income	36.36	40.91	45.45	95.45	95.45	95.45
	BK			DMSP		
	1%	5%	10%	1%	5%	10%
All countries	81.00	87.00	89.00	78.00	84.00	88.00
High income	63.64	72.73	72.73	50.00	63.64	68.18
Upper-middle	85.18	88.89	88.89	85.18	85.18	88.89
Lower-middle	86.21	93.10	96.55	86.21	93.10	96.55
Low income	86.36	90.91	95.45	86.36	90.91	95.45

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

	Population			VIIRS		
	1%	5%	10%	1%	5%	10%
All countries	9.00	17.00	23.00	75.00	80.00	84.00
High income	0.00	0.00	4.54	50.00	63.64	68.18
Upper-middle	7.41	18.52	25.93	77.78	77.78	85.19
Lower-middle	13.79	24.14	31.03	86.21	89.66	89.66
Low income	13.64	22.73	27.27	81.82	86.36	90.91
	BK			**DMSP**		
	1%	5%	10%	1%	5%	10%
All countries	72.00	75.00	79.00	72.00	76.00	81.00
High income	45.45	54.54	54.54	45.45	54.54	63.64
Upper-middle	81.48	81.48	88.89	81.48	85.19	88.89
Lower-middle	82.76	82.76	86.21	82.76	82.76	86.21
Low income	72.73	77.27	81.82	72.73	77.27	81.82

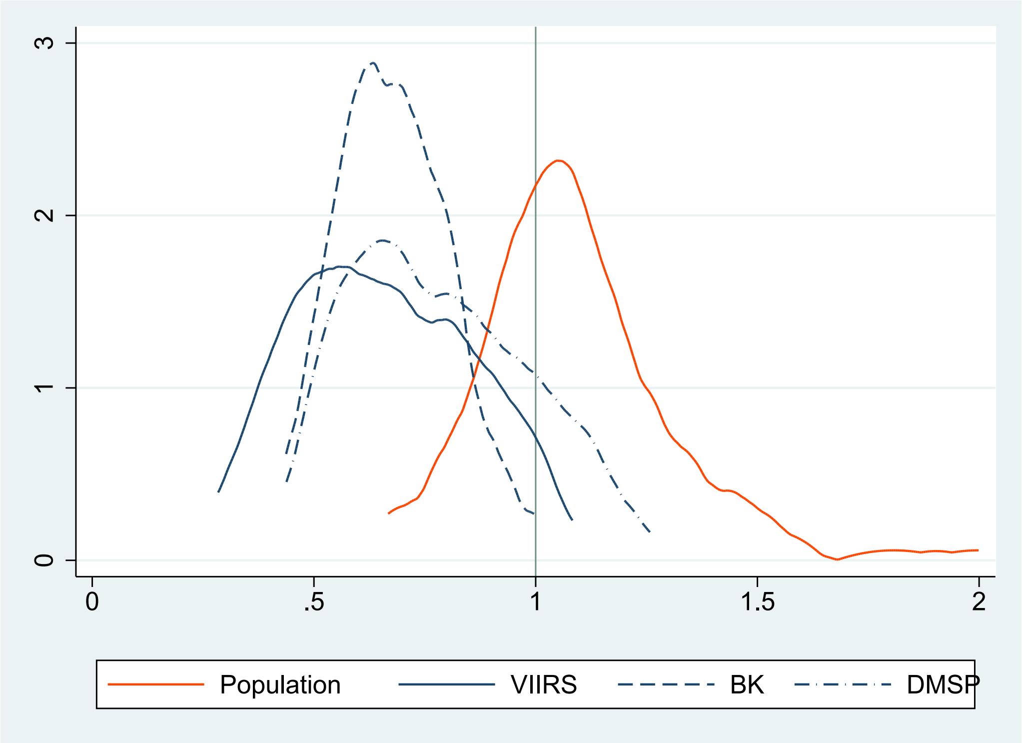

Figure 1: Kernel density estimation of Pareto coefficients at country level.

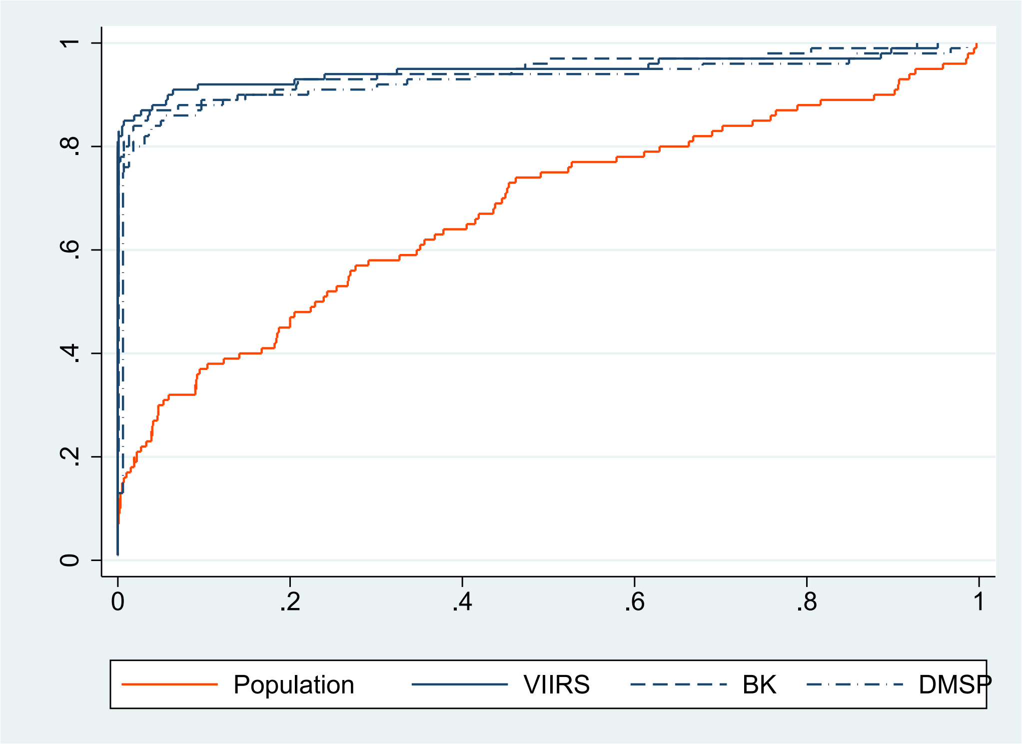

Figure 2: Cumulative density function of Kolmogorov-Smirnov test p-values using exact Zipf's law as a reference.

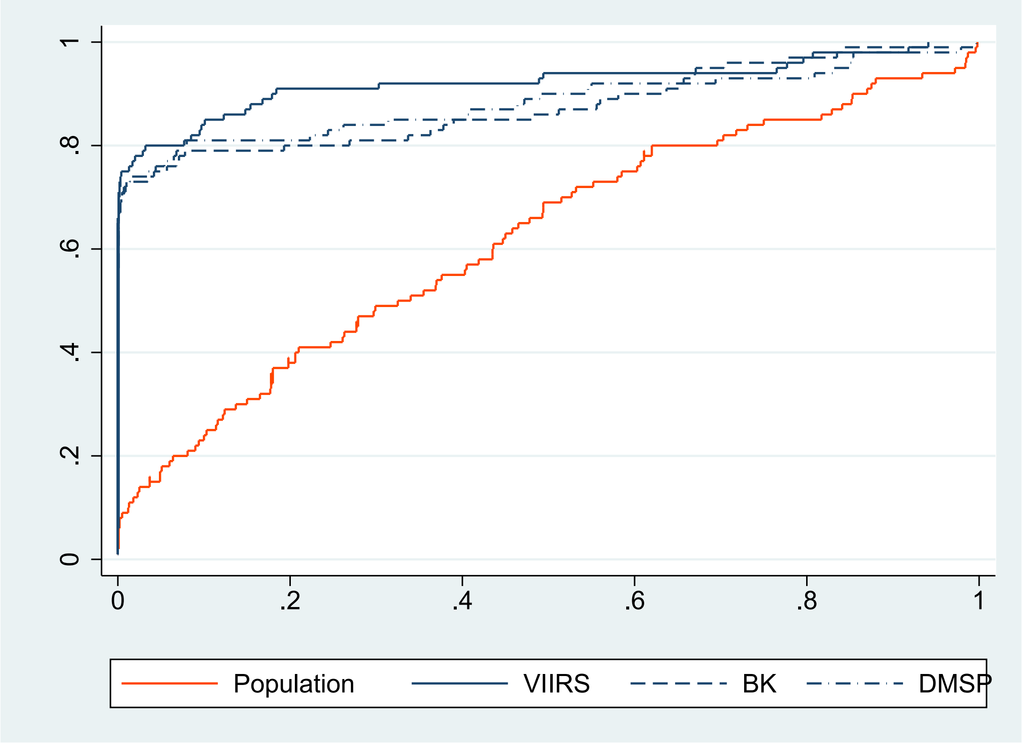

Figure 3: Cumulative density function of Kolmogorov-Smirnov test p-values using a Pareto distribution as a reference.

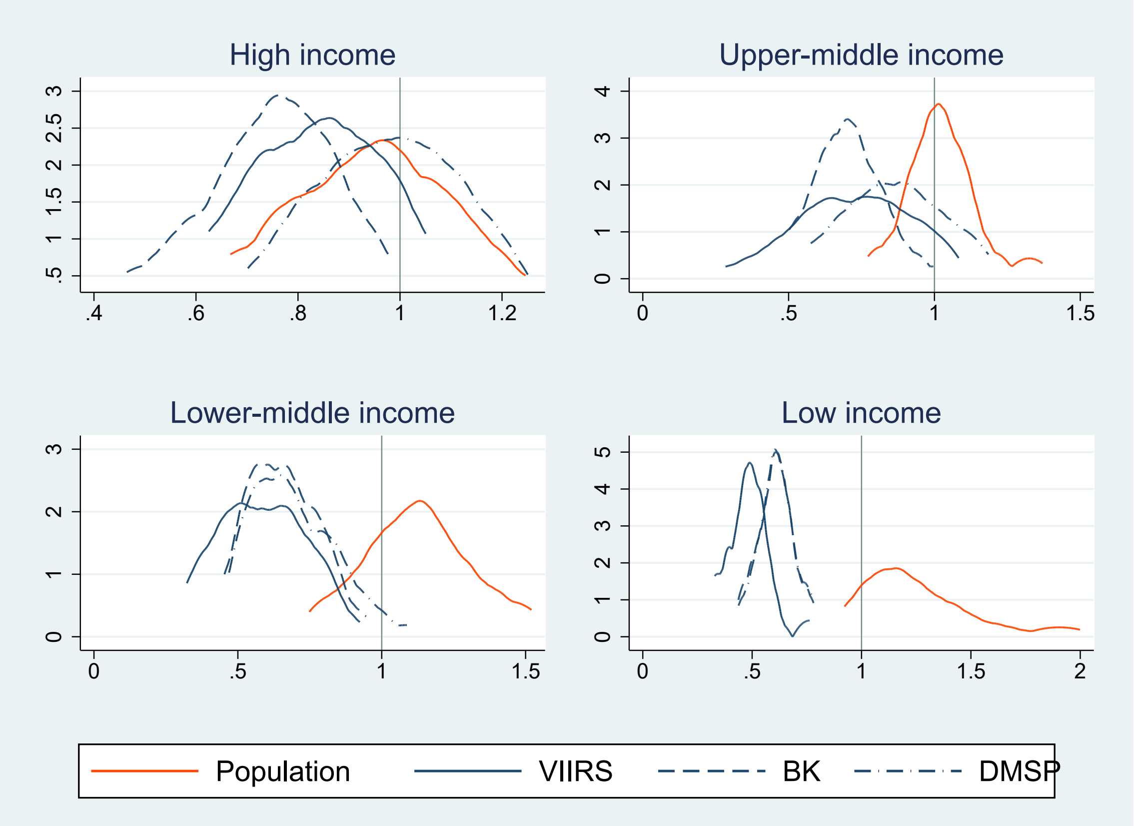

Figure 4: Kernel density estimation of Pareto coefficients at country level by income group, World Bank classification 2015.