Pareto or log-normal? A recursive-truncation approach to the distribution of (all) cities*

Giorgio Fazio
Università degli Studi di Palermo
University of Glasgow

Marco Modica
IMT Institute for Advanced Studies

July 20, 2012

Abstract

Traditionally, it is assumed that the population size of cities in a country follows a Pareto distribution. This assumption is typically supported by finding evidence of Zipf's Law. Recent studies question this finding, highlighting that, while the Pareto distribution may fit reasonably well when the data is truncated at the upper tail, i.e. for the largest cities of a country, the log-normal distribution may apply when all cities are considered. Moreover, conclusions may be sensitive to the choice of a particular truncation threshold, a yet overlooked issue in the literature. In this paper, then, we reassess the city size distribution in relation to its sensitivity to the choice of truncation point. In particular, we look at US Census data and apply a recursive-truncation approach to estimate Zipf's Law and a non-parametric alternative test where we consider each possible truncation point of the distribution of all cities. Results confirm the sensitivity of results to the truncation point. Moreover, repeating the analysis over simulated data confirms the difficulty of distinguishing a Pareto tail from the tail of a log-normal and, in turn, identifying the city size distribution as a false or a weak Pareto law.

JEL Codes: C46, D30, R12

Keywords: City size distribution; Pareto and Log-normal; Zipf's Law; Kolmogorov-Smirnov; Recursive analysis

*Giorgio Fazio, DSEAF, Facoltà di Economia, Università degli Studi di Palermo, ITA-90128, *Tel.+39 091 23895254, Fax. +39 091 422988*, Email *giorgio.fazio@unipa.it*. Marco Modica, I.M.T. Institute for Advanced Studies Lucca, Email *marco.modica@imtlucca.it*.

The usual disclaimer applies.

1 Introduction

An accurate description of the spatial distribution of population is important for a number of theoretical and policy relevant issues, ranging from a better understanding of firms and people localization decisions to the implementation of national and regional policies in terms of incentives and transport infrastructures. Unfortunately, the literature is still far from reaching consensus on such description. Two specific distributions, however, are the most accredited in the literature: the Pareto and the log-normal. Disentangling between the two has important theoretical implications. For example, a Pareto distribution implies that cities are the result of agglomeration forces and industry specific productivity shocks. A log-normal distribution, instead, implies that cities grow proportionally and independently from the initial city size and their distribution results from city-wide rather than industry specific shocks (see Gabaix, 1999, for a discussion).

Consensus view in traditional studies is in favor of a Pareto distribution with shape parameter equal to one. These studies typically base their conclusions on the evidence of a minus one relationship between the log-rank and the log-size of cities, a regularity known as Zipf's Law. For example, Rosen and Resnick (1980) estimate the value of the Pareto exponent in a sample of 44 countries, finding a mean exponent of 1.136 with most countries falling in the [0.8-1.5] range. They also suggest that larger cities grow faster than smaller cities in most of their sample countries. Soo (2005) updates these results, finding a mean Pareto exponent of 1.105 over a sample of 75 countries, but also concludes for a rejection of Zipf's Law in more than half of cases.[1]

However, these studies usually consider only the upper tail of the data, i.e. the largest cities, with a sample truncation point that is usually arbitrarily chosen.[2] Moreover, the

[1] Other papers search for historical evidence of Zipf's Law concentrating on single countries. Guerin-Pace (1995), for instance, studies Zipf's Law in France between 1831 and 1990 for a sample including cities with more than 2000 inhabitants and shows that the estimated Pareto coefficient may be sensitive to sample selection criteria. Black and Henderson (2003) construct a data set of US metropolitan areas consistently defined over the period 1900-1990 choosing a minimum relative population threshold in each decade, i.e. the 1990 ratio of the minimum to the mean metropolitan area urban population. They find a yearly Pareto coefficient around 0.85. Estimated coefficients are again sensitive to the choice of sample size. Glaeser et al. (2011) study almost 200 years of regional changes in the US and show that the empirical evidence tends to change over time, including Zipf's and Gibrat's Law. See also Gabaix, 1999 and Krugman, 1996.

[2] Gabaix (1999) shows that Zipf Law may result as the steady state distribution from Gibrat's Law, but, once again, concentrates on the upper tail only.

evidence in favor of a Pareto has to be reconciled with other empirical evidence showing that cities grow proportionally, a phenomenon known as Gibrat's Law, which should instead lead to a log-normal city size distribution. Differently from the above studies, Eeckhout (2004) suggests that it should be considered the distribution of all cities, rather than just the upper tail, and proposes an empirical investigation based on the US census dataset of Census Designated Places (CDPs). He shows that if the true underlying distribution is log-normal, then the estimated OLS coefficient of the so-called rank-size rule (i.e. the estimated parameter of the Pareto distribution) varies depending on the truncation city size, i.e. the inclusion of smaller (larger) cities in the sample, leads to a smaller (larger) coefficient. Furthermore, he argues that city growth does not depend on the initial city size, providing evidence in favor of Gibrat's Law. Based on these results, Eeckhout concludes that the size distribution of all cities follows a log-normal, rather than a Pareto.

These results have sparked further investigations on the distribution of cities beyond the upper tail. Using the same data of Eeckhout (2004), Levy (2009) presents a log-log plot of rank and city size and argues that the distribution of city size can be divided into two parts: a power law fits well the upper part, a log-normal fits better the bottom and middle parts. Eeckhout (2009) highlights the caveats of log-log plots. Instead, he proposes looking at the confidence bands of the log-normal estimates generated by a Lilliefors test and argues that the upper tail is also log-normal. Recently, Giesen et al. (2010) look at data for all cities in 8 countries and, using non-parametric and parametric goodness of fitness tests, conclude that the distribution of all cities is a Double Pareto Log-Normal (DPLN), i.e. a distribution that is Pareto in the upper and lower tails and log-normal in between. However, the DPLN distribution uses a larger set of parameters compared to the Pareto or the log-normal, which are definitely more parsimonious with only two parameters. Hence, the improvements of a novel distribution, such as the DPLN, in terms of fitness should be evaluated in relation to their dependence on a larger set of parameters.

When looking at all cities to compare the Pareto and the log-normal distributions, Eeckhout (2004) highlights the sensitivity of the Pareto coefficient to the truncation point. Moreover, Eeckhout (2009) underlines the difficulty of discriminating between a Pareto upper tail and the tail of a log-normal and hints at what may turn out to be a critical, and yet overlooked, point in the literature: *"With all the data available, and given that one nonetheless does not want to use all data, the question arises what the appropriate truncation point is. The choice of the truncation point becomes endogenous and can be chosen subjectively to favor one hypothesis over another"*. (Eeckhout, 2009, pag. 1682). Hence, a particular distribution may be favored in empirical studies depending on the chosen truncation point.

On a more general and methodological note to discriminate between different inverse power laws, Perline (2005) defines different strengths of the Pareto Law: strong, weak and false. A strong Pareto law arises when "*an inverse power law fits the full, untrucated range of the distribution of interest*"; a weak one when "*only some upper portion of the distribution follows an approximate inverse power law*" and a false when *"the largest observations (extremes) of the samples drawn from certain exponential type, and especially log-normal distributions, can closely mimic an inverse power law*" (Perline, 2005, p. 75-76). Hence, the point where the sample is truncated may indeed turn out to be critical in discriminating between alternative distributions. Both traditional and recent studies do not thoroughly address this issue.

This paper, then, proposes a reappraisal of the debate on the city size distribution in relation to the specific issue of the truncation point. Similarly to Eeckhout (2004), we do not constrain the investigation to the upper tail, but look at all cities. However, we look at all possible truncation points of the empirical distribution of all cities in order to discriminate between the two most accredited alternative theoretical distributions: the Pareto and the log-normal.

Specifically, we explore the same year 2000 US Census data of Eeckhout (2004) and the 2010 data for comparison, using a recursive approach where we begin by considering the largest cities and then add one smaller city at the time until we consider all cities. Using this approach, we reassess some of the methods used in the literature to investigate the city size distribution. First, we estimate the Pareto exponent from the typical rank-size equation for each possible truncation of data. Collecting the recursive estimates, and respective confidence intervals, we can statistically assess the adherence of Zipf's Law for each truncated sample of the distribution of all US cities. According to Eeckhout (2004), the estimated parameter should be invariant to the truncation point under the Pareto distribution and it should decrease as we extend the sample from the upper tail to the entire distribution (or increase as we move to the top of the distribution) under the log-normal.

Second, we apply the method recently suggested by Clauset et al. (2009) to estimate the lower bound of a Pareto distribution and, using a Kolmogorov-Smirnov test, we compare the relative fitness of the data to the Pareto and the log-normal distribution.

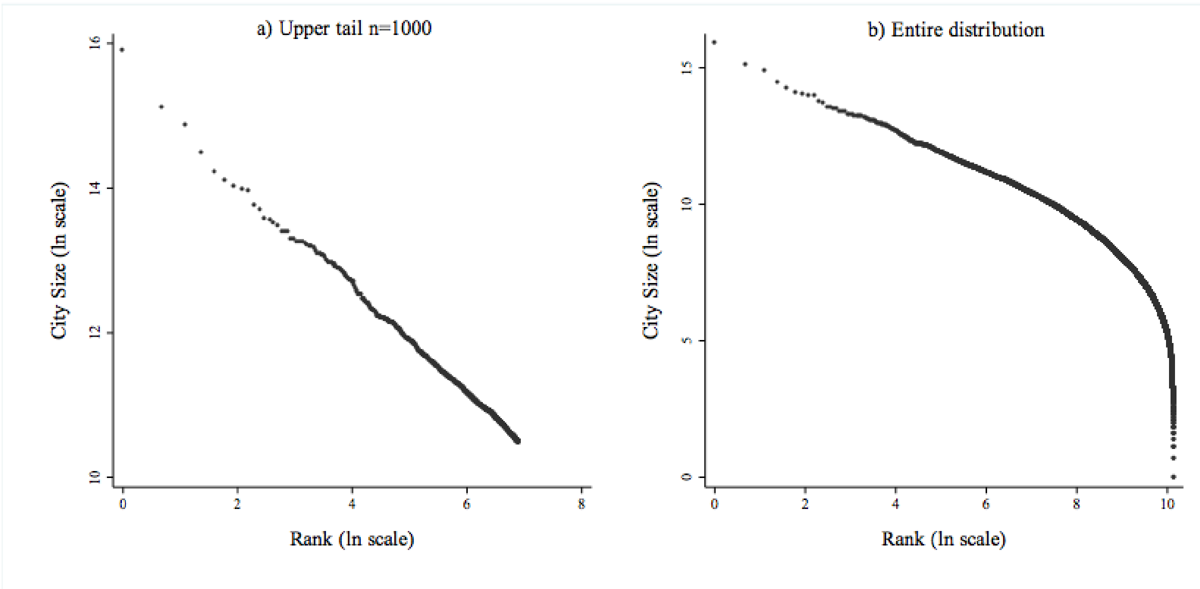
Finally, we reassess the above methods using simulated data of alternative distributions: a Pareto, a log-normal and a mixture of the two, where the upper tail is Pareto and the main body is log-normal. Our results add to the debate on the distribution of city size, highlighting some novel results in terms of the sensitivity of tests to the truncation point and showing some pitfalls of existing parametric and non-parametric methods to distinguish between the two distributions. The arbitrary choice of truncation point may, indeed, lead researchers to incorrectly conclude for a specific distribution. Our tests and simulations seem to confirm the difficulty to distringuish between a Pareto upper tail and the tail of a log-normal and, consequently, identify the city size distribution as a weak or a false inverse power law.

The rest of the paper is organized as follows. The next section presents the empirical strategy and the results of recursive Zipf's Law equations and Kologorov-Smirnov tests. Section 3 replicates the methodology using simulated data. Section 4 concludes.

2 A Recursive approach to the distribution of all cities

A long tradition of papers underlines the difficulty of discriminating between a Pareto tail and log-normal distribution. For example, in reference to the use of log-log plots, which provide a visual assessment of the rank-size rule, Macauley (1922) states that the linearity of the tail of a frequency distribution charted on a logarithmic scale is not informative of a Pareto distribution, as it is a common feature of various types of frequency distributions. Parr and Suzuki (1973) similarly affirm that : "*[...]truncation of the log-normal distribution at an appropriately high level enables the truncated portion to be regarded as not significantly different from the rank size distribution*". This point is illustrated in figure 1 that compares log-log plots for the upper tail and for the entire distribution. While the left quadrant clearly points to a Pareto, the right seems to point to a log-normal.

Figure 1: *Log-log rank-size plots (first 1000 largest cities vs entire distribution).*



More formally, Eeckhout (2004) shows that a variable P obeys a Pareto distribution if its density function, $\phi(P)$, and cumulative density function, $\Phi(P)$, are:

$$\phi(P) = \frac{a\underline{P}^a}{P^{a+1}} \qquad \forall P \geq \underline{P},$$

$$\Phi(P) = 1 - \left(\frac{\underline{P}}{P}\right)^a \; \forall P \geq \underline{P},$$

where a is a positive shape parameter and $\underline{P}$ is the scale parameter or the truncation city size, i.e. the minimum value of population P. The parameter a is also known as the Pareto coefficient and is a tail index. As mentioned above, in a log-log plot the distribution is represented by a straight line and Zipf's law satisifes Pareto with $a = 1$.

According to Clauset et. al (2009), few phenomena seem to obey the Pareto distribution for all values and, as discussed above, most studies on the city size distribution find that the Pareto distribution is a good representation just for the upper tail, i.e. above a minimum threshold. However, even when a researcher intends to investigate just the upper tail, the choice of $\underline{P}$ may be critical, as a truncation point that is too high (low) may shorten (lengthen) the "right" size of the upper tail biasing tests of the appropriate distribution. The identification of the right truncation point may be interesting also for another issue.

If, as sustained in some literature, the upper tail is Pareto and the entire distribution is log-normal, is there a switching point between the two distributions? This issue has received some attention in physics and statistics (see, among the others, Mitzenmacher, 2004; Perline, 2005; Clauset et al., 2009), but, with the exception of Eeckhout (2004), it has been largely ignored in economics, where the choice of threshold is usually arbitrary.[3]

In order to investigate the sensitivity of the distribution to the truncation point, we apply a recursive approach to the distribution of all cities. Following Eeckhout (2004), to consider "all" cities we use US Census data covering almost all the US population in "incorporated" and "unincorporated" places in the years 2000 and 2010.[4] For the year 2000, the dataset covers 25,359 places and 208 millions US residents of the total 281 millions and, for the year 2010, 29,494 places and 230 millions US residents of the 308 millions total. The difference in number of places is due to changes introduced by the US Census Bureau: 24,841 are identical in the two years. Even though they may not coincide with the economically more meaningful definition of city, and previous work has considered Metropolitan Areas as the reference unit (see Gabaix, 1999; Ioannides and Overman, 2003), we prefer "places" as reference units in order to make our result comparable with Eeckhout (2004) and account for a larger population size. Here, for robustness we replicate the analysis for the two years.

2.1 Recursive Zipf's Law

As mentioned above, we use a recursive approach to observe the adherence of the data to Zipf's Law for all possible truncation points of the distribution of all cities. As standard in the literature, we estimate the Pareto coefficient using simple rank-size OLS regressions where, following Gabaix and Ibragimov (2007), the rank is shifted by 0.5 to correct for the potential bias in small samples highlighted by Gabaix and Ioannides (2003), so that the estimating equation is:

[3] Bee et al. (2011) consider some related methodological issues. First, different tests often provide different results. Second, sample size may matter as well as truncation. Finally, under the hypothesis of log-normal distribution, when the threshold is high, i.e. when we use few observations, the tail seems to follow a Pareto.

[4] An incorporated place is an entity (populated area) with its own municipal government (city, town, village, borough and so on). Unincorporated places are, instead, areas lacking of own municipal government. In the US Census, these take the name of Census Designated Places (CDPs). The CDPs have been included for the first time in the year 2000.

Table 1: Estimated coefficients for chosen truncation thresholds. Dependent variable ln (Rank-0.5).

	2000 Census Data				2010 Census Data			
$\underline{N}$	$\underline{P}$	$\hat{k}$(s.e.)	$\hat{a}$(s.e.) [GI s.e.]	R^2	$\underline{P}$	$\hat{k}$(s.e.)	$\hat{a}$(s.e.) [GI s.e.]	R^2
135	155,554	21.955	1.423	0.992	178,395	22.532	1.460	0.993
		(0.137)	(0.011) [0.173]			(0.139)	(0.011) [0.178]	
2,000	19,383	20.747	1.322	0.997	21,039	20.870	1.322	0.995
		(0.045)	(0.004) [0.042]			(0.057)	(0.005) [0.042]	
5,000	6,592	18.623	1.129	0.984	7,273	18.721	1.137	0.983
		(0.052)	(0.005) [0.023]			(0.057)	(0.006) [0.023]	
12,500	1,378	15.954	0.864	0.960	1,556	16.064	0.866	0.961
		(0.036)	(0.004) [0.011]			(0.037)	(0.004) [0.011]	
25,000	42	13.187	0.553	0.875	193	13.899	0.630	0.922
		(0.021)	(0.003) [0.005]			(0.021)	(0.003) [0.006]	
29,000	_	_	_	_	35	13.136	0.538	0.882
						(0.018)	(0.003) [0.005]	

Dependent variable ln (Rank-0.5). $\underline{N}$ is the number of cities above the truncation threshold. $\underline{P}$ is the relative population, $\hat{k}$ is the constant, GI stands for Gabaix-Ibragimov

$$ln(rank - 0.5) = k - a\,lnP, \tag{1}$$

where k is a constant and P is the population size. Standard errors are given by $(2/n)^{0.5}\,\hat{a}$. The parameter $\hat{a}$ is estimated for recursively truncated samples of the city size distribution, starting with the ten most populated cities and then adding one (less populated) city at the time until, like Eeckhout (2004), we consider all cities. Collecting the estimates of the Pareto exponent together with the respective 95% confidence interval, we can statistically assess the validity of Zipf's Law for each truncated city size distribution. In particular, while the estimated Pareto coefficient should be invariant to the truncation point, it should increase under the log-normal (Eeckhout, 2004).

Table 1 extracts the recursive OLS estimates of equation (1) for the six truncation points reported in Eeckhout (2004). These results seem consistent with previous work.[5] The estimated Pareto coefficients seem, indeed, "threshold sensitive": the "longer" the upper tail, the lower the estimated coefficient. As already indicated in Eeckhout (2004), the coefficients decrease together with the truncation point. It is also interesting to compare the estimated parameters for the two different census years. The 2010 Census contains a larger number of observations mostly thanks to the improved accuracy in the definition of unincorporated places, as many CDPs present in the Census 2000 dataset have been split into two or more CDPs especially in the middle and in the lower tail of the distribution. The presence of these

[5] Notice that for the year 2000 we use the same data of Eeckhout (2004), but employ the Gabaix and Ibragimov (2007) correction. Standard OLS results seem to be, indeed, downward biased in smaller samples.

new observations does not seem to affect our results.[6]

In figures 2a and 2b we present the full recursive estimates. Figure 2a focuses on the largest 1,000 cities to look more closely at the upper tail. Over this range, the estimated Pareto coefficient looks quasi-constant, indicating a potential Pareto distribution.

The coefficient shows some degree of fluctuation in the very first observations, probably due to the influence of individual observations in a smaller sample, and then increases. In terms of statistical significance, estimates are indifferent from one for the first observations and are then statistically different from one, settling around the average estimated parameter of 1.4. Hence, Zipf's Law seems to be rejected, if not for the very first observations.[7]

Interestingly, this information was not evident by looking at table 1, where it was only possible to see the rank-size rule as a diminishing threshold process, but it was not possible to fully gauge the adherence of the data to Zipf's law for the upper tail.

What happens if we extend the analysis from the upper tail (here, the first 1000 cities) to all cities in the sample? Figure 2b shows all the estimated Pareto coefficients (and 95% confidence intervals) against each recursive truncation threshold. A number of results are worth mentioning. First, the coefficient clearly diminishes (increases) as we include smaller cities (larger cities), a result that, contrary to figure 2a, corroborates the evidence of log-normality. Second, in terms of statistical significance, the recursive coefficients seem to display non-monotonic behavior. The Pareto coefficient is not statistically different from one in the very upper tail, where researchers typically set their cut-off point to estimate Zipf's Law (see Black and Henderson, 2003; Soo, 2005), but also for a second range of cities (between the 7,116th and 8,773rd in the year 2000 and 7,066th and 8,763rd in the year 2010). Hence, the Pareto exponent is not statistically different from 1 for two samples of the same distribution. Clearly, this result could only emerge by looking at all possible truncation points. Figure 2b confirms how picking an arbitrary $\underline{P}$ may (mis)lead researchers to conclude in favor of a specific distribution. Finally, comparison of the left and right panels shows that results are robust to the use of different census years and are stable over time, with similar patterns and hierarchy.[8]

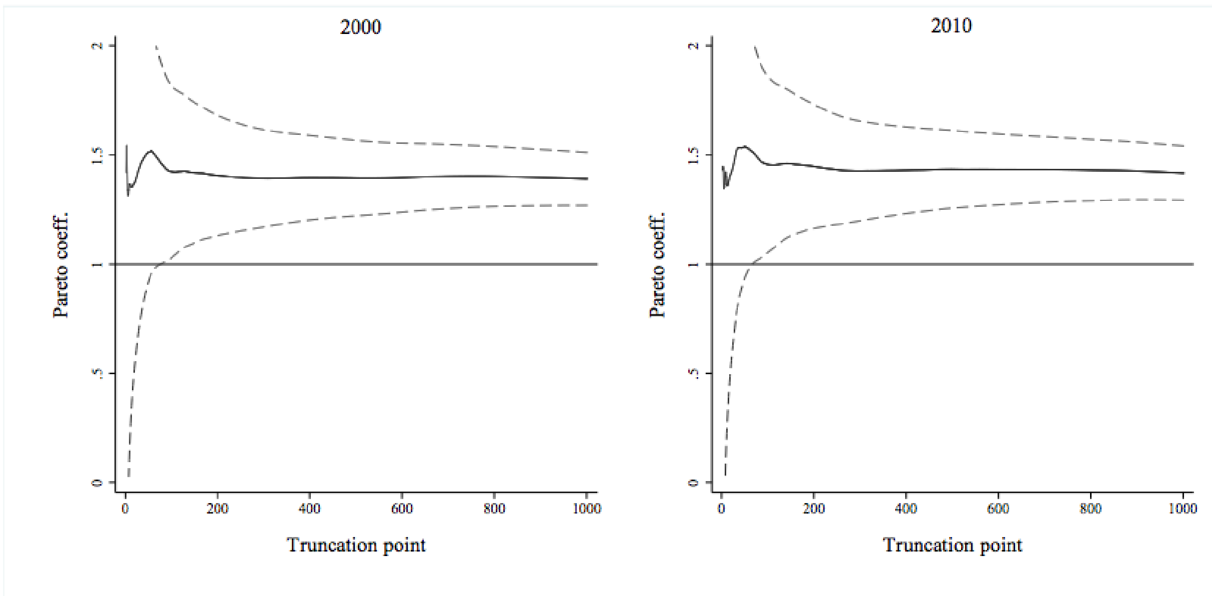
[6] For the full samples, the estimated coefficients are 0.526 and 0.508 for year 2000 and 2010, respecively.

[7] Recall that Gabaix and Ioannides (2003) argue that it is compatible with Zipf's Law a Pareto exponent between 0.8 and 1.2.

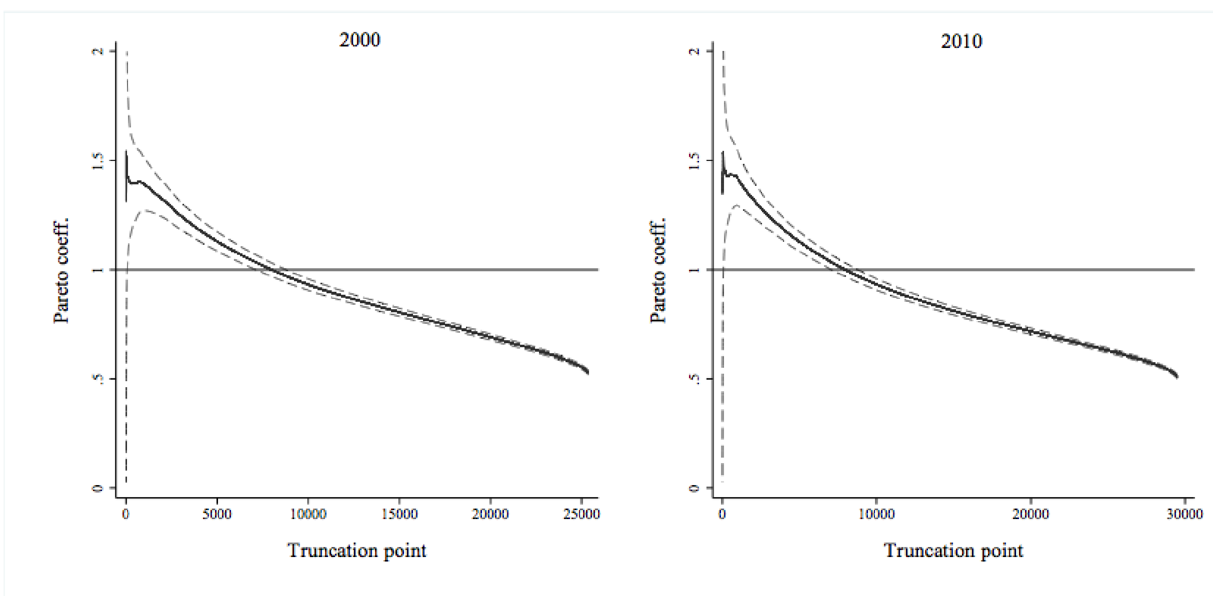
[8] This also suggests that a more precise specification of the medium-small cities in 2010 does not significantly affect the results.

Figure 2: Recursive Pareto coefficient and 95% Confidence Interval (Gabaix-Ibragimov s.e.)

(a) Upper tail (1000 largest cities)



(b) All cities



2.2 A non-parametric test alternative

In this section, we exploit an alternative non-parametric methodology in order to discriminate between a Pareto distribution and a log-normal distribution. As discussed above, few phenomena seem to obey the Pareto distribution for all values and most studies find that the Pareto distribution is a good representation just for the upper tail, i.e. above a minimum threshold. Here we use the method proposed by Clauset et al. (2009) to estimate this minimum threshold, $\hat{\underline{P}}$. The authors suggest testing the equality between the theoretical and empirical density functions using Kolmogorov-Smirnov tests over recursively truncated distributions. Our estimate $\hat{\underline{P}}$ is then the value of $\underline{P}$ that minimizes the "recursive" Kolmogorov-Smirnov statistics, D:

$$D = \sup_{p \geqslant \underline{p}} |\Phi_p(x) - \Phi(x)|,$$

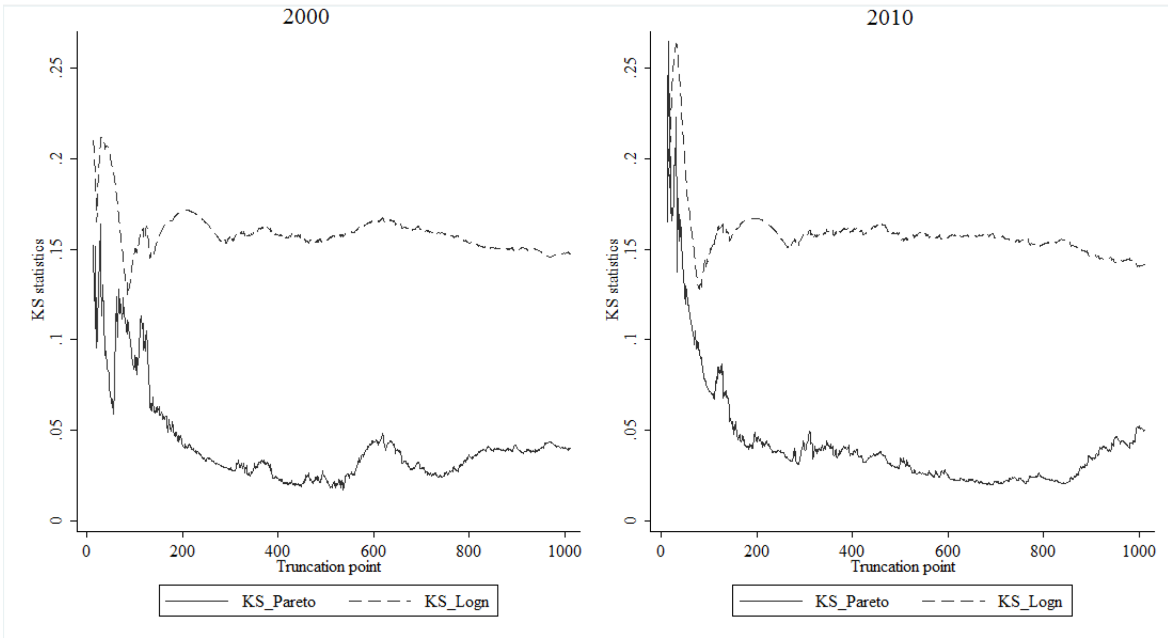
where $\Phi_p(x)$ is the empirical cumulative density function for p i.i.d observations, and $\Phi(x)$ is the theoretical cumulative density function. The KS statistic computes the supremum of the absolute value of the set of distances among the two. Under the null, the difference between the two is zero, i.e. the sample is drawn from the reference distribution. Rejection of the null, however, should be considered carefully, as the KS test tends to over-reject the null when the sample is large. P-values for the KS tests are reported in Appendix A and show clearly that the tests always reject the null as the sample increases. Hence, in order to conclude in favor of one or the other distribution, especially when the sample size is larger, we can compare the size of the KS statistics with smaller statistics denoting a better fit.

Figure 3 reports the recursive Kolmogorov-Smirnov statistics. As before, we begin the recursive analysis with the largest cities and then add smaller ones until we include all cities. In panel a) of figure 3, we first look at the largest 1000 cities. The overall evidence seems to favor the Pareto distribution. Interestingly, however, for the very upper tail (exactly, 93 cities in 2000 and 90 in 2010), the KS statistics are visually too close and do not allow disentangling between the two distributions. For these observations, the p-values of the KS statistics in figure A1 confirm that both theoretical distributions can equally adapt to the empirical. After this portion of the upper tail, however, the Pareto seems to adapt better to the data with p-values rejecting the null of the KS test for a large portion of the upper tail.[9]

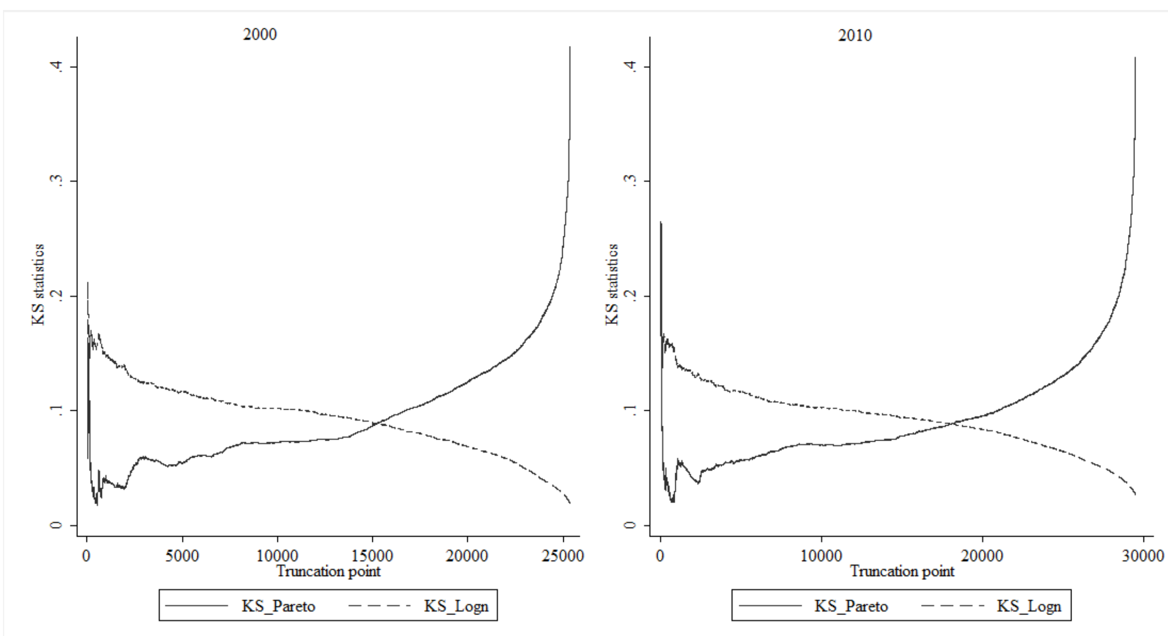
[9] In figure A1, the KS test is rejected for the Pareto up to around the 1500th truncation point in the year 2000 around the 990th truncation point in the year 2010.

Figure 3: Recursive Kolmogorov-Smirnov test

(a) Upper tail (1000 largest cities)



(b) All cities



Following the approach of Clauset et al. (2009), we find the minimum of the KS statistic for the 536th city in the year 2000 (D=0.0173) and 695th city in the year 2010 (D= 0.0198), which implies a minimum population threshold of 57,777 and 55,081 inhabitants, respectively. These tests, then, highlight a Pareto upper tail well before the arbitrary threshold of 100,000 inhabitants typically used by scholars for US data (Soo, 2005). In panel b) of figure 3, we show KS recursive statistics for all truncated samples up to the entire untruncated distribution. Again, the evidence in favor of one or the other distribution changes depending on the truncation point: comparison of the KS statistics shows first the Pareto and then the log-normal as a better fit. Just like for the very upper tail, the two distributions are again indistinguishable half way to the entire distribution. When the distribution of all cities is considered, the log-normal appears as the best fit, as indicated by Eeckhout (2004).These results are in line with those obtained from the Zipf's Law regressions in the previous subsection. If we do not take into account the problem of the correct cut-off, the KS test, also, could lead to concluding for a Pareto, when the true distribution is log-normal, and viceversa. Further, the evidence presented cannot rule out that a portion of the distribution of cities, the upper tail in particular, may be power law distributed and it confirms the difficulty of disentangling a Pareto and a log-normal in a portion of the upper tail.

3 Weak or False Inverse Power Law?

The above analysis seems to confirm the sensitivity of test results with respect to the choice of truncation point. Moreover, it seems to highlight the distribution of cities as potentially as either a weak or a false power law, according to the definitions of Perline (2005). Indeed, it is not clear whether "*only some upper portion of the distribution follows an approximate inverse power law*" (weak power law) or "*the largest observations (extremes) of the samples drawn from certain exponential type, and especially log-normal distributions, can closely mimic an inverse power law*" (false power law).

To further investigate this issue, we reassess the rank-size regressions and the non-parametric alternative against simulated data. In particular, we simulate three different random datasets: a log-normal, a Pareto with shape parameter equal to one (so that Zipf's law holds) and a "mixture" of Pareto upper tail (first 1000 observations) and log-normal body.[10]

[10] In detail, we draw a log-normal dataset with same mean (7.28) and standard deviation (1.75) of the real data in the year 2000. For the "mixture" data we replace the first 1000 observations of the log-normal distribution with a sample where the first observation is twice the second, thrice the third and so on.

Following the same steps of the previous section, we first report the estimated recursive Pareto coefficients over the range of the 1000 largest cities and then over the entire distribution. Results are presented in panel a) of figure 4. Looking at the upper tail, we notice a quasi-constant behavior of the coefficient (with different means) for all three simulated datasets. As expected, the estimated coefficients are not significantly different from one for the Pareto and the mixture-distributions. Interestingly, the estimated Pareto coefficients are not significantly different from one in the very upper tail to then become different from one for the simulated log-normal data, exhibiting a similar size and statistical significance to the real data.

When we look at the entire distribution (panel b) of figure 4), the estimated Pareto coefficients are, unsurprisingly, constant over the entire distribution. They are flatter for the mixture data, displaying a long Pareto tail. Again, the simulated log-normal displays the same signature of the real data.

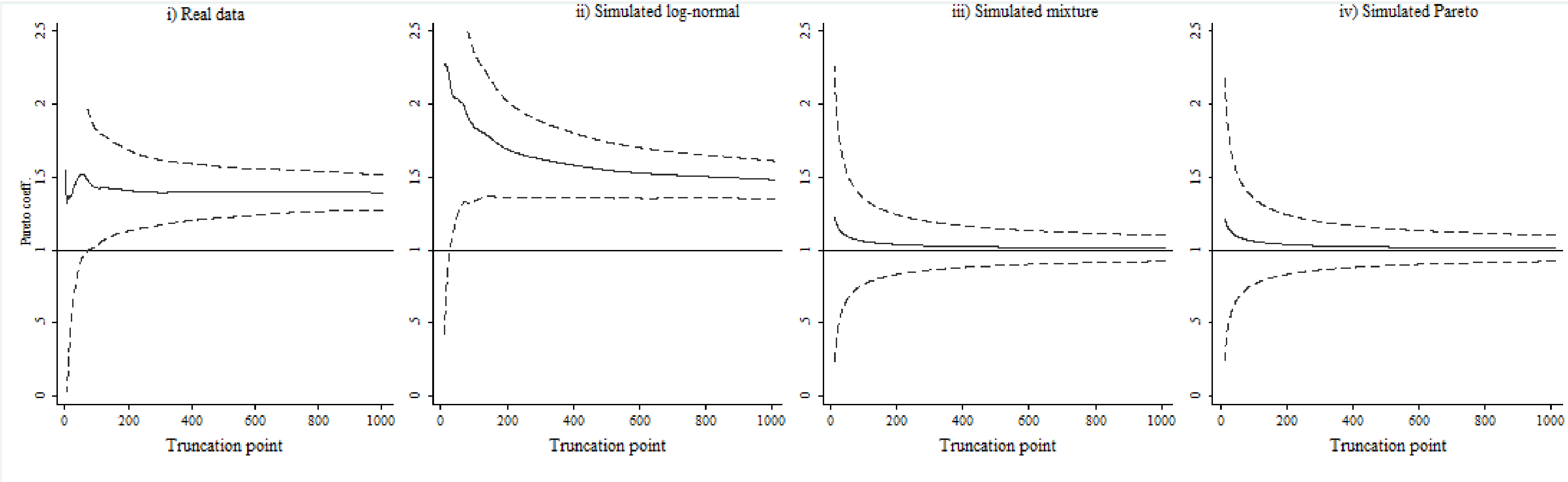
In figure 5, we repeat the recursive non-parametric approach on the simulated data. Overall, for the largest 1000 cities in panel a), the KS statistics seem to indicate that the Pareto distribution is better than the log-normal, irrespective of the type of simulated distribution. However, in the very upper tail the KS statistics for the Pareto and the log-normal are indistinguishable, just like for the real data.[11] When we add smaller and smaller cities beyond the 1000th in panel b) of figure 5, results show great concordance between the real data, the log-normal and the "mixture" distribution, with the latter unsurprisingly exhibiting a longer upper tail.

Both Zipf's Law and Kolmogorov-Smirnov tests seem to highlight the simulated log-normal as the most similar to the real data. This result seems to suggest that for a portion of the upper tail, and especially the distribution of the largest cities, the log-normal may be a close representation of the real data, as well as the Pareto. This evidence, however, has to be combined with that from Kolmogorov-Smirnov tests in the previous section, where the log-normal and the Pareto could both apply to the very first observations of the upper tail (around the largest 100 cities), but when the tail is extended, only the Pareto distribution is statistically indifferent from the real data. Hence, we are unable to unambiguously establish whether the distribution of cities falls in the weak or false power law category.

[11] Of course, for the simulated data the unreported p-values indicate significant KS statistics for both distributions when we consider the entire sample.

Figure 4: Rank-Size Regressions over simulated data

(a) Upper tail (1000 largest cities)



(b) All cities

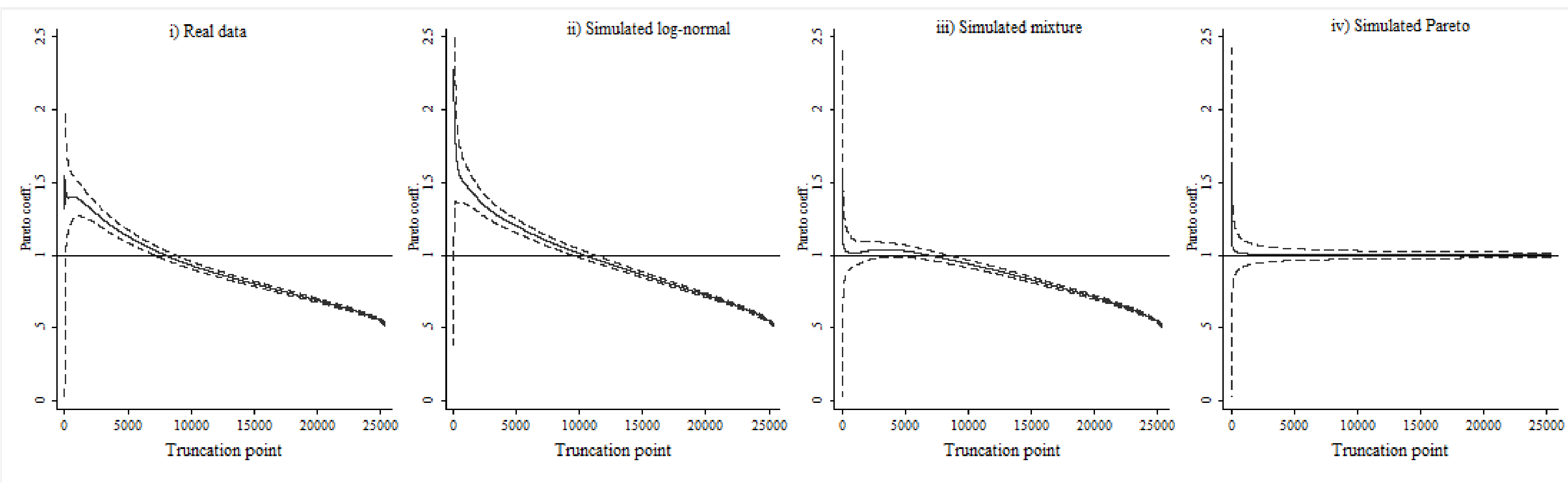
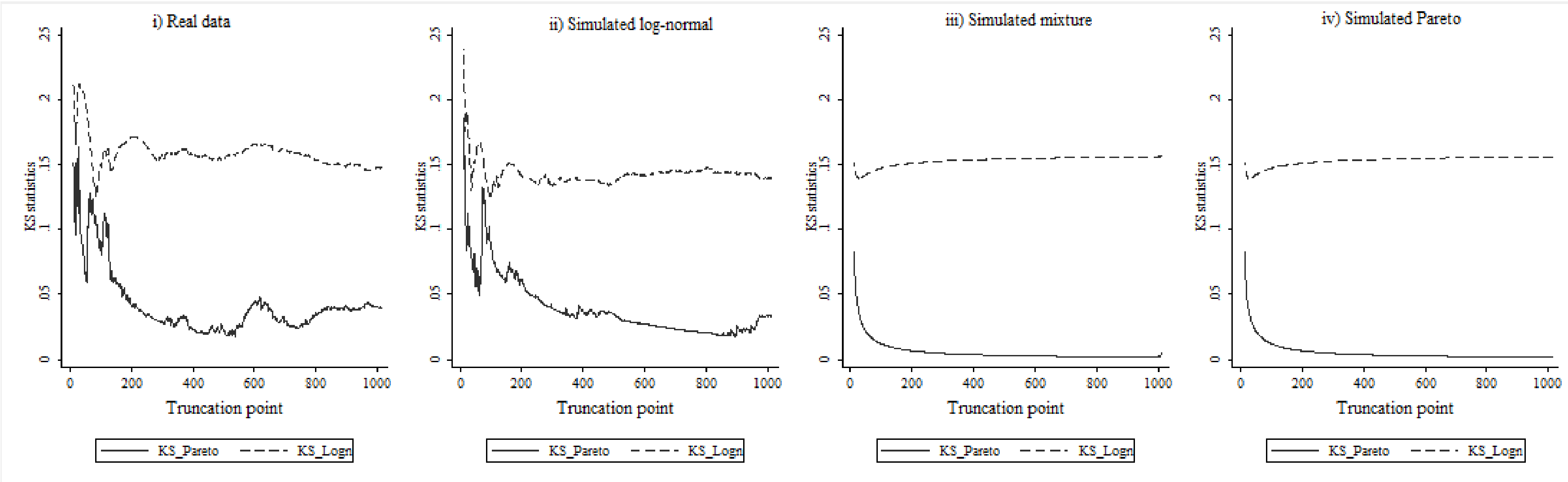
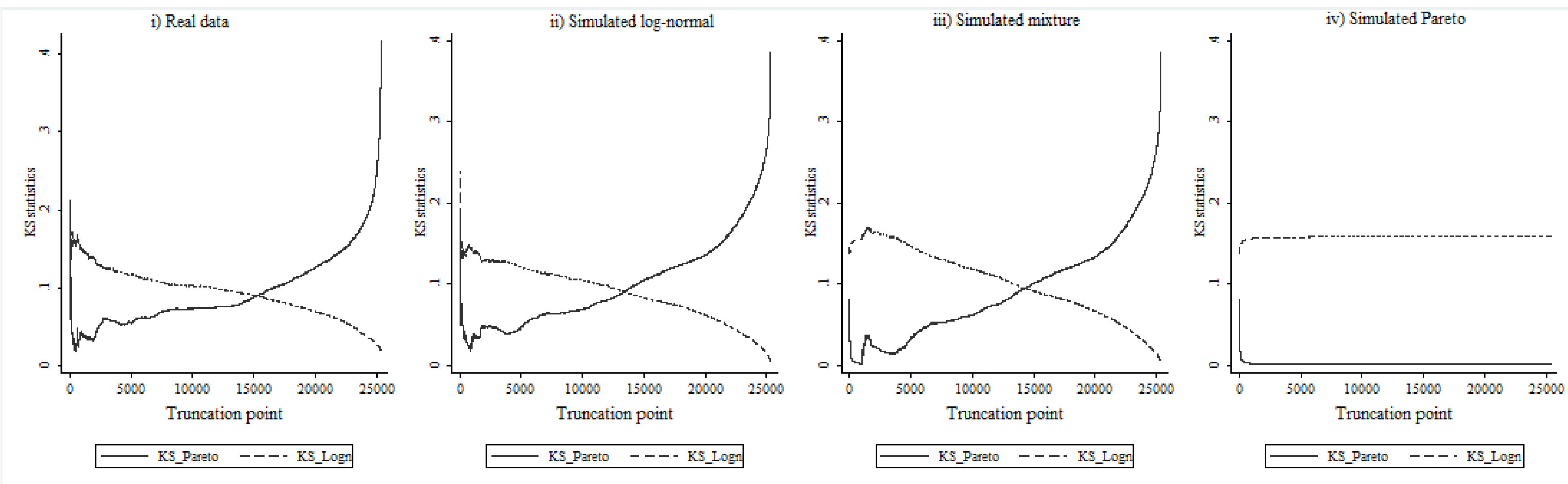


Figure 5: Simulated KS

(a) Upper tail



(b) All cities



4 Conclusions

The identification of the correct city size distribution emerges from the literature as controversy with two most likely candidates: the Pareto and the log-normal distributions. Recently, some commentators (see Eeckhout, 2009, in particular) have suggested the possibility that part of this controversy may be due to the arbitrary choice of truncation of the distribution of all cities. A truncation point that is too high, or too low, may bias tests of the appropriate distribution of the upper tail. Also, a false power law may emerge when the extremes of the samples closely mimic an inverse power law, especially if drawn from a log-normal (Perline, 2005). Yet, this issues has been substantially overlooked in the economics literature.

In this paper, we investigate the city size distribution with respect to the "truncation point". In particular, we propose a recursive-truncation approach to reassess the common Zipf's Law regression and a non-parametric alternative proposed by Clauset et al. (2009) against all possible truncation points of the entire distribution of cities.

Some interesting results emerge from this analysis. First, we find evidence of Zipf's Law with the Pareto exponent equal to one in the very upper tail (above the largest 135 cities). However, when we start lowering the truncation point, adding smaller cities one at the time, we find that the size of Pareto coefficient first becomes statistically different from one and then starts decreasing, showing a non-monotonic behavior with respect to the truncation point. Statistically, the coefficient crosses one for different ranges of truncation points: in the upper tail and much later when the sample is truncated around the 8000th city. Finally, in line with Eeckhout (2004), the log-normal seems the best fit when the entire distribution of cities is considered.

The same recursive approach is also applied using the non-parametric method proposed by Clauset et al. (2009). This analysis shows that the both distribution can apply to the upper portion of the distribution of cities (above around the first 100 cities). Beyond these cities, the upper tail conforms better to the Pareto distribution on the grounds of statistical significance. When the truncation point is extended to include smaller ans smaller cities, it is not possible to disentangle the two on the grounds of statistical significance, but on the grounds of the size of the statistics. In general, the Pareto seems a better fit when we are close to the upper portion of the data and the log-normal seems a better fit when we approach the entire distribution. Also, the Pareto distribution seems to be longer than traditionally postulated by previous studies on the grounds of an arbitrary truncation point.

These results seem to support to the claim by Eeckhout (2009) that an arbitrary choice of the cut-off of the distribution may mislead scholars to conclude in favor of one or the other distribution. While the log-normal seems to best fit the entire sample, truncating the distribution may lead to conclude in favor of a Pareto, especially in the upper tail. Even then, however, the analysis returns a kind of non-monotonic behavior indicating that a Pareto might apply over more than one range of city sizes: in the very upper tail and when the sample is truncated mid-way to the distribution of all cities.

Finally, we assess whether the distribution of cities can potentially fall into the weak or false power law categories defined by Perline (2005). To this end, we replicate the proposed recursive rank-size and the non-parametric alternative test of city size distributions using simulated data drawn from a Pareto, a log-normal and a mixture of the two. While the rank size regressions seem to point to the distribution of cities as potentially a false power law with the log-normal simulated data displaying a remarkably similar signature to the real data, the non parametric test seems less conclusive. In this case, the size of the test statistics for the simulated log-normal and the real data are also remarkably similar in the very upper tail. When the upper tail, however, is extended, the test seems to point in favor of the Pareto. Hence, the non-parametric test seems unable to settle whether the distribution of cities is a weak or a false power law.

Overall, the analysis seems to provide methodological insights into the issue of discriminating between alternative city size distributions and the truncation point problem. Moreover, it confirms the difficulty of distinguishing between the tail of a log-normal and a power law tail for the population distribution of cities. More reliable tests are probably necessary to settle this long standing issue.

References

Bee, Marco; Riccaboni, Massimo and Schiavo, Stefano, "Pareto versus lognormal: A maximum entropy test," *Phys. Rev. E*, 2011, *84*, 026104.

Black, Duncan and Henderson, Vernon, "Urban evolution in the USA," *Journal of Economic Geography*, 2003, *3* (4), 343–372.

Clauset, Aaron; Shalizi, Cosma R. and Newman, M. E. J., "Power-law distributions in empirical data," *ArXiv e-prints*, 2007.

Eeckhout, Jan, "Gibrat's Law for (All) Cities," *American Economic Review*, 2004, *94* (5), 1429–1451.

Eeckhout, Jan, "Gibrat's Law for (All) Cities: Reply," *American Economic Review*, 2009, *99* (4), 1676–83.

Gabaix, Xavier, "Zipf'S Law For Cities: An Explanation," *The Quarterly Journal of Economics*, 1999, *114* (3), 739–767.

Gabaix, Xavier and Ibragimov, Rustam, "Rank-1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents," NBER Technical Working Papers 0342, National Bureau of Economic Research, Inc 2007.

Gabaix, Xavier and Ioannides Yannis M., "The Evolution of City Size Distributions," Technical Report 2003.

Gibrat, Robert, *Les inégalités économiques*, Libraire du Recueil Siray, Paris France, 1931.

Giesen, Kristian; Zimmermann, Arndt and Suedekum, Jens, "The size distribution across all cities - Double Pareto lognormal strikes," *Journal of Urban Economics*, 2010, *68* (2), 129–137.

Glaeser, Edward L.; Ponzetto, Giacomo A.M. and Tobio, Kristina, "Cities, Skills, and Regional Change," NBER Working Papers 16934, National Bureau of Economic Research, Inc 2011.

Guerin-Pace, France, "Rank-size distribution and the process of urban growth," *Urban Studies*, 1995, *32* (3), 551–562.

Ioannides, Yannis M. and Overman, Henry G., "Zipf's law for cities: an empirical examination," *Regional Science and Urban Economics*, March 2003, *33* (2), 127–137.

Krugman, Paul, "Confronting the Mystery of Urban Hierarchy," *Journal of the Japanese and International Economies*, 1996, *10* (4), 399–418.

Levy, Moshe, "Gibrat's Law for (All) Cities: Comment," *American Economic Review*, 2009, *99* (4), 1672–75.

Macauley, F., *Pareto's law and the general problem of mathematically describing the frequency distribution of income*, Vol. 2 of *Income of the United States. its Amount and Distribution 1909-1919*, New York: National Bureau of Economic Research, 1922.

Mitzenmacher, Michael, "A Brief History of Generative Models for Power Law and Lognormal Distributions," *Internet Mathematics*, 2004, *1* (2), 226–251.

Parr, John B. and Keisuke Suzuki, "Settlement Populations and the Lognormal Distribution," *Urban Studies*, 1973, *10* (3), 335–352.

Peng, Guohua, "Zipf's law for Chinese cities: Rolling sample regressions," *Physica A: Statistical Mechanics and its Applications*, 2010, *389* (18), 3804 – 3813.

Perline, Richard, "Strong, weak and false inverse power laws.," *Stat. Sci.*, 2005, *20* (1), 68–88.

Rosen, Kenneth T. and Resnick, Mitchel, "The size distribution of cities: An examination of the Pareto law and primacy," *Journal of Urban Economics*, 1980, *8* (2), 165 – 186.

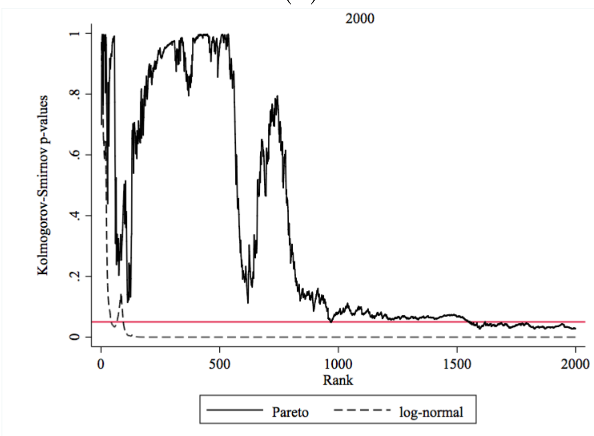
Soo, Kwok Tong, "Zipf's Law for cities: a cross-country investigation," *Regional Science and Urban Economics*, 2005, *35* (3), 239–263.

Zipf, George Kingsley, *Human behaviour and the principle of least effort*, New York, NY, USA: Hafner Pub. Co., 1949.

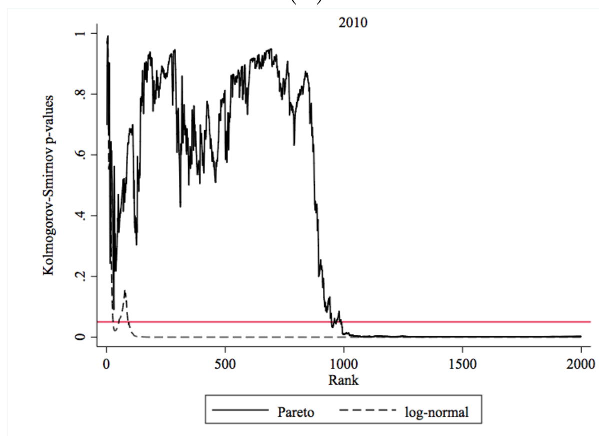
A Additional Figures

Figure A1: Kolmogorov-Smirnov Tests p-values (2000 largest cities)

(a)



(b)



Horizontal line denotes 5% statistical significance