Investigaciones Regionales – Journal of Regional Research
https://doi.org/10.38191/iirr-jorr.24.064

Accepted Article

A test of the relationship between the Pareto exponent and sample size

Rafael González-Val, Fernando Sanz-Gracia***

Received: 01 April 2024
Accepted: 01 October 2024

Abstract:

This paper uses un-truncated city population data from three countries—the United States, Spain and Italy—to empirically test Proposition 1 put forth by Eeckhout (2004 *American Economic Review*, 94: 1429–1451). Eeckhout's hypothesis was that the estimate of the Pareto exponent in a standard Zipf regression decreases with sample size, if the underlying city size distribution is lognormal. Using rolling sample regressions, we find that this proposition is only valid once we enter the lognormal body of the distribution; for the Pareto-distributed upper-tail, the estimated exponent does not vary with sample size.

Keywords: City size distribution; Zipf's law; Pareto exponent; Pareto distribution; lognormal distribution; rolling sample regressions.
JEL Classification: C12; R11; R12.

Una prueba de la relación entre el exponente de Pareto y el tamaño muestral

Resumen:

Este documento utiliza datos de población de ciudades sin restricciones de tamaño de tres países—Estados Unidos, España e Italia—para poner a prueba empíricamente la Proposición 1 presentada por Eeckhout (2004 *American Economic Review*, 94: 1429–1451). La hipótesis de Eeckhout era que la estimación del exponente de Pareto en una regresión Zipf estándar disminuye con el tamaño de la muestra, si la distribución del tamaño de las ciudades subyacente es lognormal. Utilizando regresiones de muestra móvil, encontramos que esta proposición solo es válida una vez que entramos en la parte central lognormal de la distribución; para la cola superior distribuida siguiendo una función de Pareto, el exponente estimado no varía con el tamaño muestral.

Palabras clave: Distribución del tamaño de la ciudad; ley de Zipf; exponente de Pareto; distribución de Pareto; distribución lognormal; regresiones de muestra móvil.
Clasificación JEL: C12; R11; R12.

1. Introduction

Zipf's law is an empirical regularity that has received significant attention in the urban economics and geographic literature. It establishes a linear and stable relationship between the rank and size (population) of cities and is considered to be a reflection of a steady-state situation.

* Universidad de Zaragoza, Facultad de Economía y Empresa, Departamento de Análisis Económico, Zaragoza. España. / Institut d'Economia de Barcelona (IEB), Facultat d'Economia i Empresa, Barcelona. España. rafaelg@unizar.es
** Universidad de Zaragoza, Facultad de Economía y Empresa, Departamento de Análisis Económico, Zaragoza. España. fsanz@unizar.es
Corresponding author: rafaelg@unizar.es

The significance of Zipf's law is difficult to overstate. First, due to its broad applicability across numerous fields, it can be applied to virtually any quantitative phenomenon. For instance, it has been used to study the size distribution of the number of victims in armed conflicts (González-Val, 2016), the frequency of musical notes in famous compositions (Zanette, 2006), the magnitude of migratory movements (Clemente et al., 2011), and, most famously, the frequency of different words in Joyce's Ulysses (Zipf, 1949). Second, it has clear theoretical ties to urban growth theory, as shown by Gabaix (1999). Third, there is an almost esoteric quality to the empirical regularity derived from Zipf's law, which is known as the rank-size rule: the k-th largest city is exactly one k-th the size of the largest city. This empirical regularity has fascinated urban geographers since Auerbach's (1913) seminal work. Finally, in the context of urban planning and demography, Cristelli et al. (2012, p. 7) argued that for Zipf's law to hold, the urban system must be integrated and the sample must be coherent in the sense of being the "result of some kind of optimization in growth processes or of an optimal self-organization mechanism." That situation implies that the urban system shares a common language, culture, history, and set of rules.

In an influential article, Eeckhout (2004) stimulated an academic debate about what distribution better fits city size distributions. He justified the use of un-truncated city size data, showing that the underlying statistical distribution has strong implications for the fulfilment of Zipf's law. Traditionally, due to data limitations, most studies have considered only the largest cities. However, Eeckhout (2004) demonstrated the statistical importance of considering both large and small cities because truncated samples lead to biased results. In particular, he found that the Zipf exponent declines systematically as the sample size increases if the underlying distribution is lognormal rather than Pareto (i.e., Proposition 1 in Eeckhout (2004)).

However, that proposition has remained largely untested (a few exceptions include Luckstead and Devadoss (2014) and Peña and Sanz-Gracia (2021)), probably because the reaction of the literature was a search for the best distribution to fit un-truncated data and the lognormal distribution was soon replaced by other more convoluted ones. The current consensus is that the best city size distribution may be a mixed distribution, separating the lognormal body of the distribution from the upper Pareto tail (Giesen et al., 2010; Ioannides and Skouras, 2013; Puente-Ajovín et al., 2020). However, the method for defining the Pareto upper-tail is still subject to debate (Fazio and Modica, 2015; Schluter, 2021). Therefore, the possible relationship between Zipf's law and the sample size for both the upper-tail and the whole distribution is still a relevant issue that we aim to clarify in this study, which constitutes our primary innovative contribution to the existing literature.

The remainder of this paper is organised as follows. Section 2 briefly reviews the literature. Section 3 presents the data and the methodology that we used. Section 4 describes our main results, and Section 5 concludes our work.

2. Literature review

There are excellent surveys on this topic that comprehensively review the literature on city size distribution and Zipf's law up to the time of their publication. In this regard, we can chronologically cite Nitsch (2005), Cottineau (2017), and Arshad et al. (2018). Accordingly, this paragraph lists a selection of recent or relatively recent papers (from 2018 onwards) that address the topic of Zipf's law for cities: Arshad et al. (2019), Hackmann and Klarl (2020), Düben and Krause (2021), Fernholz and Kramer (2024), and González-Val et al. (2024).

Our contribution to the existing literature lies in explicitly testing Proposition 1 of Eeckhout (2004), which establishes a direct relationship between the behavior of the Pareto exponent and sample size. While many studies have traditionally estimated the Pareto exponent for different sample sizes (e.g., Eeckhout (2004), González-Val (2010)), analyses that examine variations in the parametric estimate of the Pareto exponent across all possible sample sizes are less common. To the best of our knowledge, only Peng (2010), Fazio and Modica (2015), and Peña and Sanz-Gracia (2021) have used a recursive procedure (i.e., rolling sample regressions) to test the relationship between the Pareto exponent and sample size. An alternative branch of the literature estimates the local Pareto exponent by sample size using nonparametric methods (Ioannides and Overman, 2003; Luckstead and Devadoss, 2014).

3. Data and methods

We used settlement size data from the decennial censuses of three countries: the United States (US; 2000 and 2010 censuses), Spain (2001 and 2011 censuses) and Italy (2001 and 2011 censuses). The data were obtained from the national official statistical services and included un-truncated city population data without any size restrictions.

For the US, our sample for the year 2000 is the same as that used by Eeckhout (2004). The spatial units are what the US Census Bureau calls 'places,' which include both incorporated places (i.e., administrative cities legally incorporated under the laws of their respective states) and Census Designated Places (i.e., a concentration of population, housing and commercial structures that is identifiable by name, but is not within an incorporated place). We considered 25,358 places in the US in the 2000 dataset and 29,461 places in the 2010 dataset.

The geographical unit of reference in Spain and Italy is the municipality. Municipalities are the smallest spatial units (local governments); they are the administratively defined "legal" cities, comprising the whole territory and population of both countries. For Italy, the number of cities by period is 8,100 municipalities in 2001 and 8,081 in 2011. For Spain, our samples include 8,108 municipalities in 2001 and 8,074 in 2011.

The standard Zipf regression equation, including the correction introduced by Gabaix and Ibragimov (2011), is as follows:

$$\ln\left(R-\frac{1}{2}\right)=b-a\ln S+\varepsilon, \tag{1}$$

where R is the empirically observed rank (1 for the largest city, 2 for the second largest, and so on), S is city size (population) and a is the Pareto exponent. If $\hat{a}=1$, Zipf's law holds, meaning that, ordered from largest to smallest, the size of the second city is half that of the first, the size of the third is a third of the first, and so on. In any case, a is interpreted as a measure of the degree of inequality in the city size distribution: large (small) values correspond to more equal (unequal) city sizes. Standard errors are calculated by applying Gabaix and Ioannides's (2004) correction: $\mathrm{GI\,s.e.}=\hat{a}\cdot(2/N)^{1/2}$, where N is the sample size. We use these corrected standard errors to calculate the confidence bands of $\hat{a}$ at the 95% confidence level.

Proposition 1 of Eeckhout (2004, p. 1442) reads as follows: "If the underlying distribution is the lognormal distribution, then the estimate of the parameter $\hat{a}$ of the Pareto distribution is increasing in the truncation city size and decreasing in the truncated sample population." That is, adding increasingly smaller cities to a sample should result in growing inequality in the distribution. This statement is what we sought to test empirically. To do so, we employed rolling sample regressions (Peng, 2010; Fazio and Modica, 2015; Peña and Sanz-Gracia, 2021): we incorporated cities one by one into the sample until we attained the smallest urban unit. We began with the largest two cities, and each time we calculated the estimated Pareto exponent.

4. Results

Figures 1, 2 and 3 show our results for the three countries we studied[1]. In each figure, a graph is shown for the whole city size distribution in any of the two years considered, as well as an additional graph zooming in the upper-tail of the distribution. To define the upper-tail, we followed the procedure set forth by Clauset et al. (2009), which is specifically designed to select an optimal truncation point[2]. Note that this cut-off is only used to delimit the number of estimates shown in the upper-tail graphs. As a robustness check, we estimated alternative thresholds using the procedure of Beirlant et al. (1996) recommended by

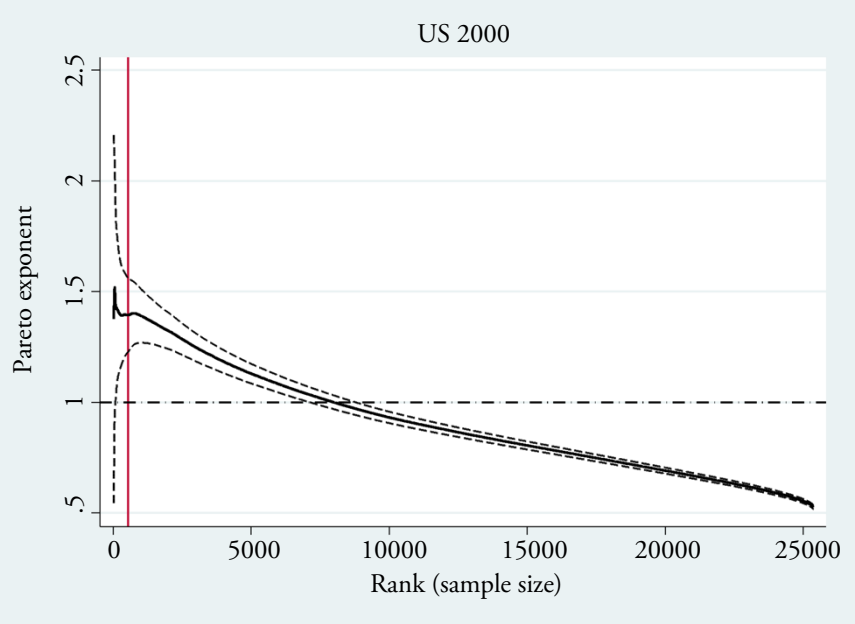
[1] Figure 1 is similar to Figure 2 of Fazio and Modica (2015).

[2] For a review of the various methods available for defining the thresholds and their main properties, see Fazio and Modica (2015) and Schluter (2021).

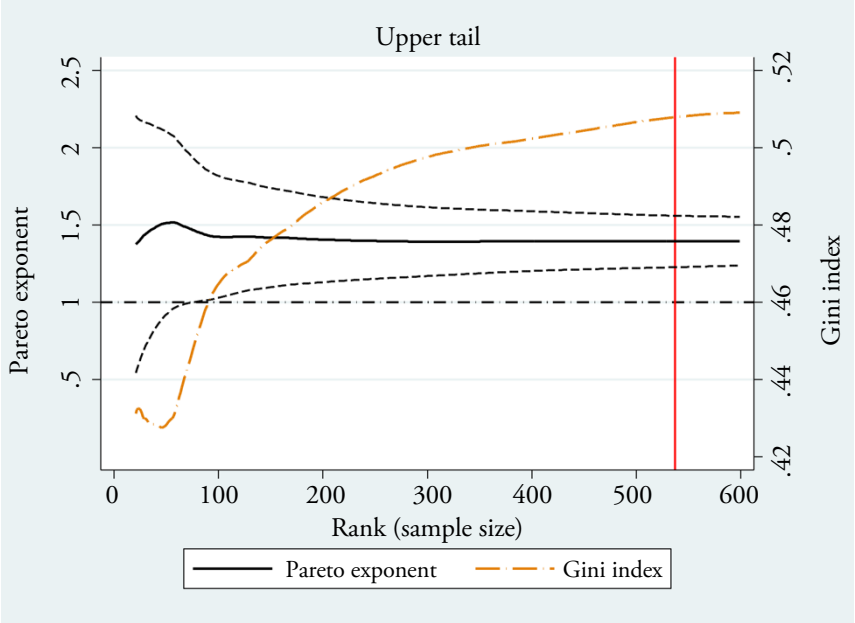
Schluter (2021); we also used the method of Ioannides and Skouras (2013) to estimate the threshold of the distribution that switches between a lognormal and a power-law distribution, and results hold[3]. Finally, the upper-tail graphs also include the calculation of the Gini index[4] by sample size—which does not impose a specific size distribution (Pareto for Zipf regressions)[5].

FIGURE 1.
Pareto exponent by sample size, US places in 2000 and 2010

a) US places in 2000 (whole sample)



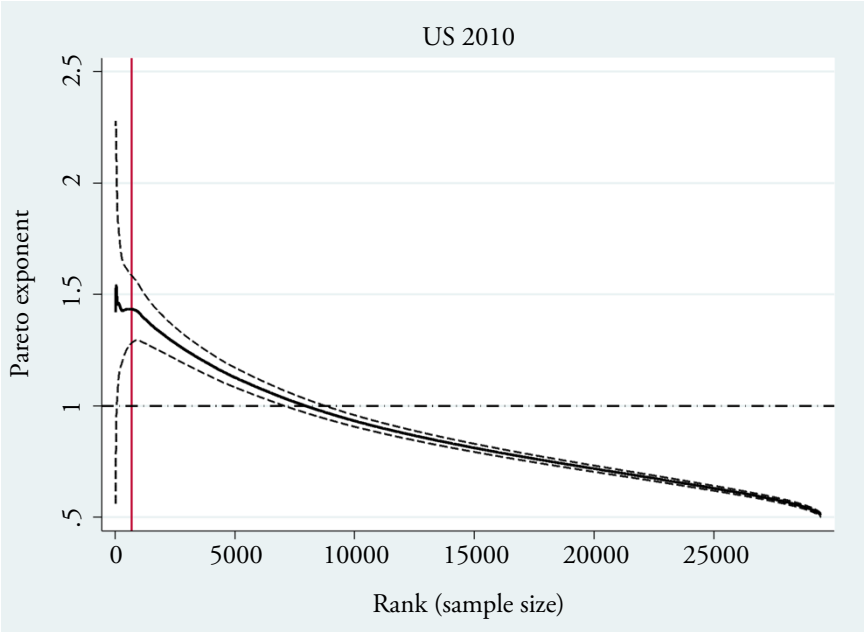
b) US places in 2000 (zoom)



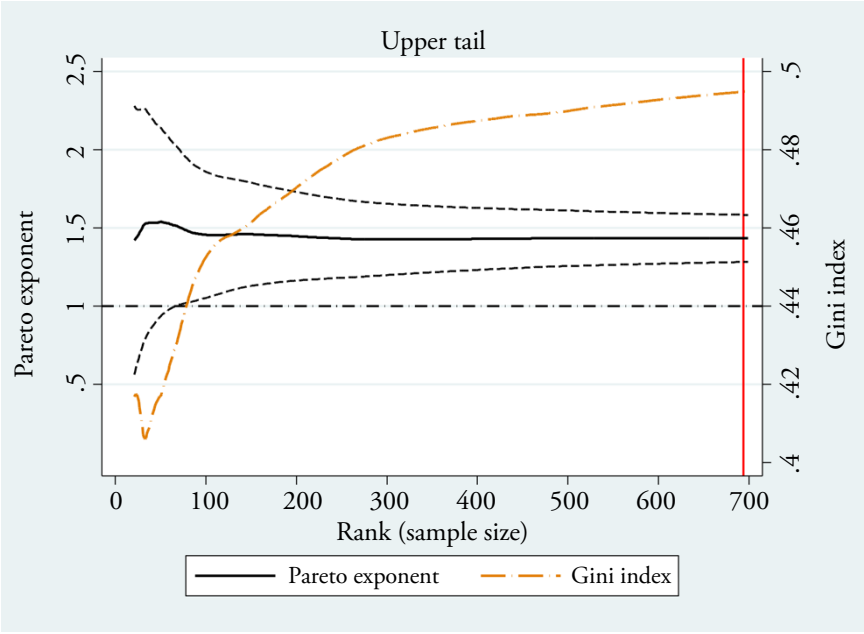
[3] These results are available from the authors upon request.
[4] The Gini coefficient is bounded by 0 (indicating perfect equality of city sizes) and 1 (indicating complete inequality).
[5] Results of the Clauset et al.'s (2009) test for a power law (not shown) reveal that the Pareto distribution provides a plausible fit to the data for the upper-tail city size distribution in all cases.

FIGURE 1. CONT.
Pareto exponent by sample size, US places in 2000 and 2010

c) US places in 2010 (whole sample)



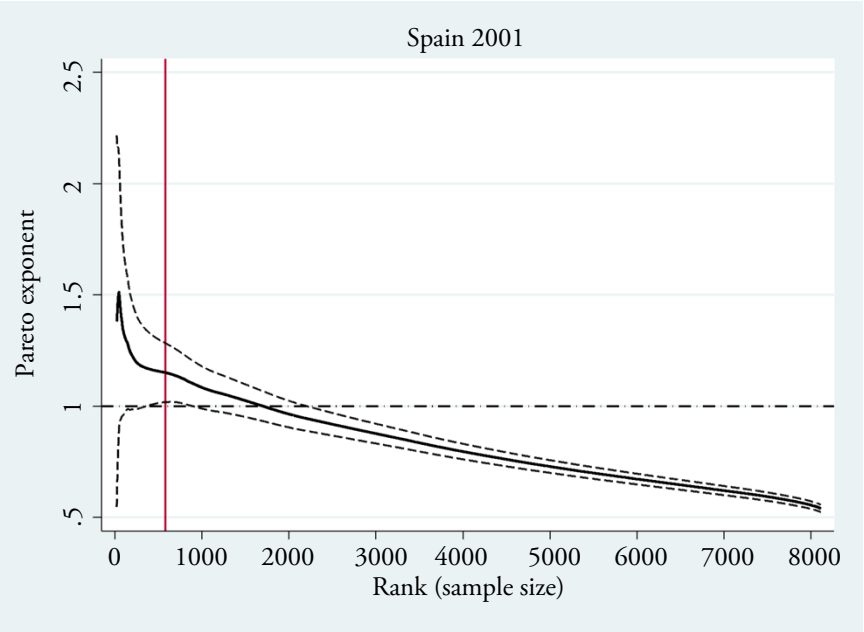
d) US places in 2010 (zoom)



Notes: The Pareto exponent is estimated utilising Gabaix and Ibragimov's Rank-1/2 estimator. Dashed lines represent the standard errors calculated by applying Gabaix and Ioannides's (2004) corrected standard errors: $\mathrm{GI\ s.e.} = \hat{a}.(2/\mathrm{N})^{1/2}$, where *N* is the sample size. The vertical red lines indicate the threshold of the Pareto upper-tail, determined using Clauset et al. (2009)'s methodology. There are 537 places in the upper-tail in 2000 (the population threshold is 57,746) and 694 in 2010 (the population threshold is 55,156).

FIGURE 2.
Pareto exponent by sample size, Spanish municipalities in 2001 and 2011

a) Spanish municipalities in 2001 (whole sample)



b) Spanish municipalities in 2001 (zoom)

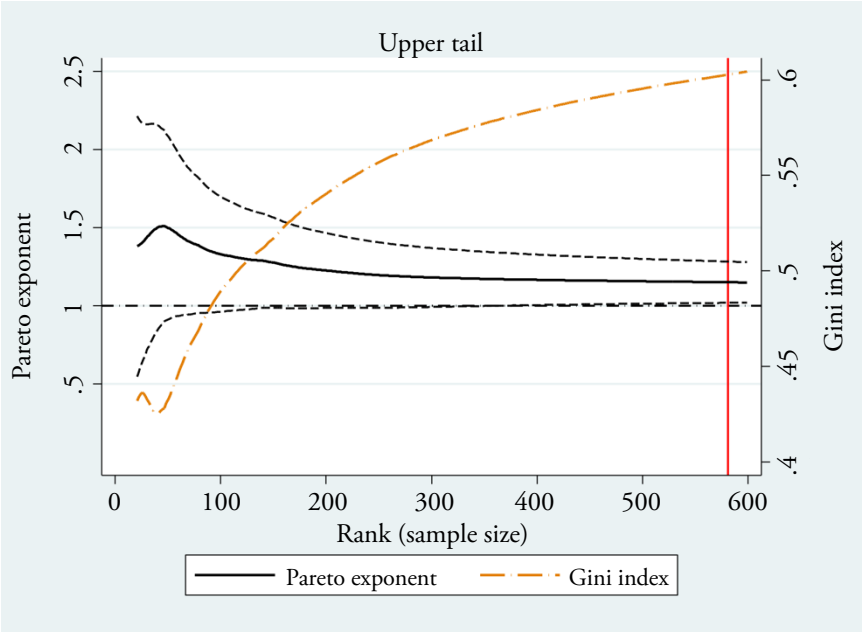
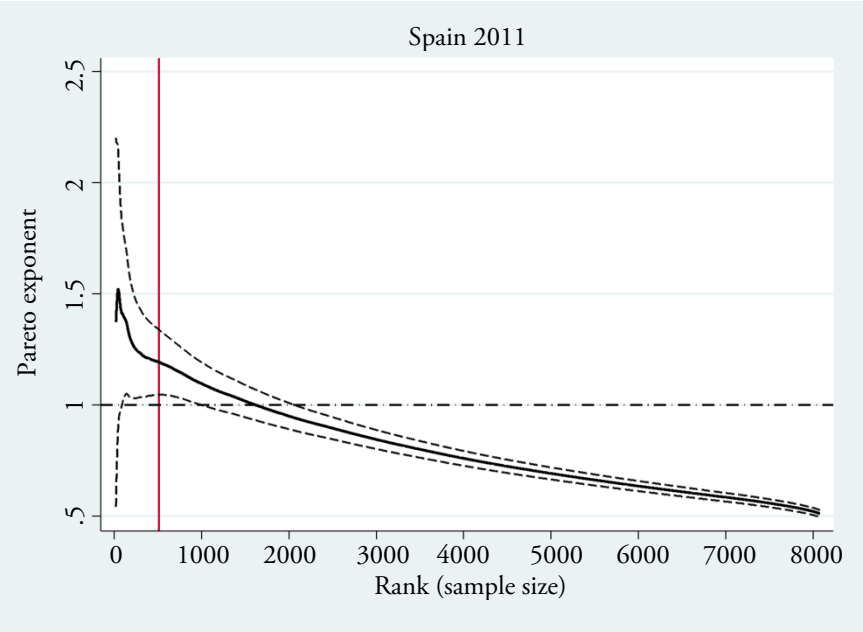
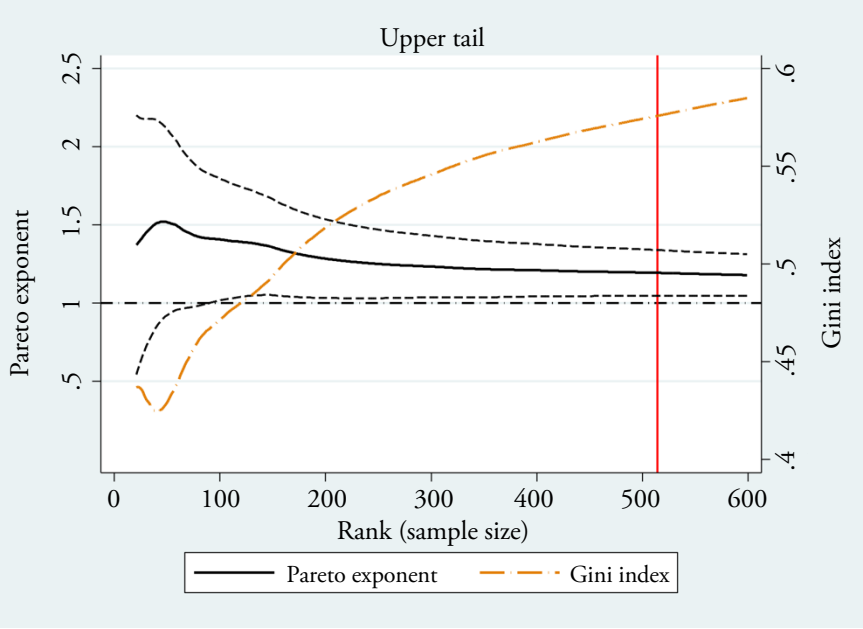


FIGURE 2. CONT.
Pareto exponent by sample size, Spanish municipalities in 2001 and 2011

c) Spanish municipalities in 2011 (whole sample)



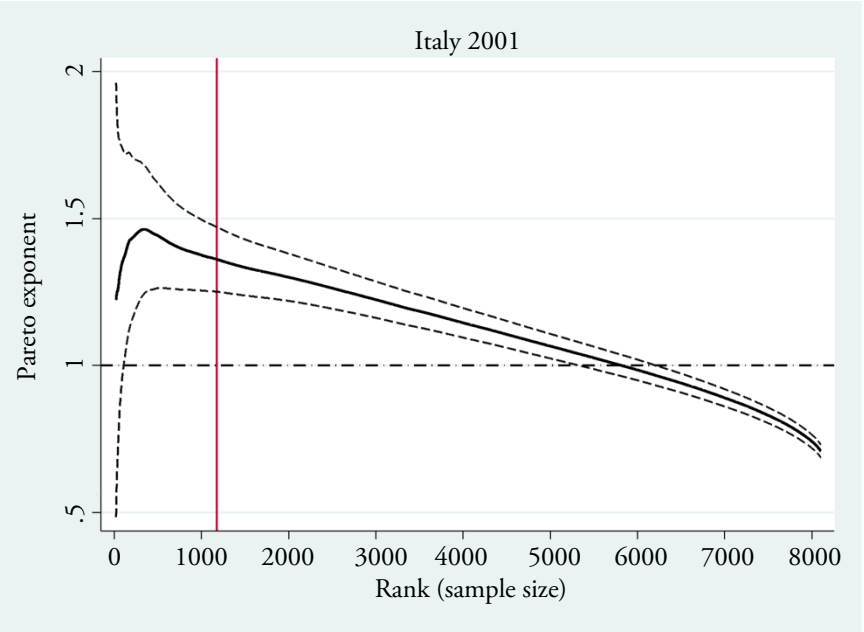
d) Spanish municipalities in 2011 (zoom)



Notes: The Pareto exponent is estimated utilising Gabaix and Ibragimov's Rank-1/2 estimator. Dashed lines represent the standard errors calculated applying by Gabaix and Ioannides's (2004) corrected standard errors: $\text{GI s.e.} = \hat{a}.(2/N)^{1/2}$, where N is the sample size. The vertical red lines indicate the threshold of the Pareto upper-tail, determined using Clauset et al. (2009)'s methodology. There are 581 municipalities in the upper-tail in 2001 (the population threshold is 11,331) and 514 in 2011 (the population threshold is 15,583).

FIGURE 3.
Pareto exponent by sample size, Italian municipalities in 2001 and 2011

a) Italian municipalities in 2001 (whole sample)



b) Italian municipalities in 2001 (zoom)

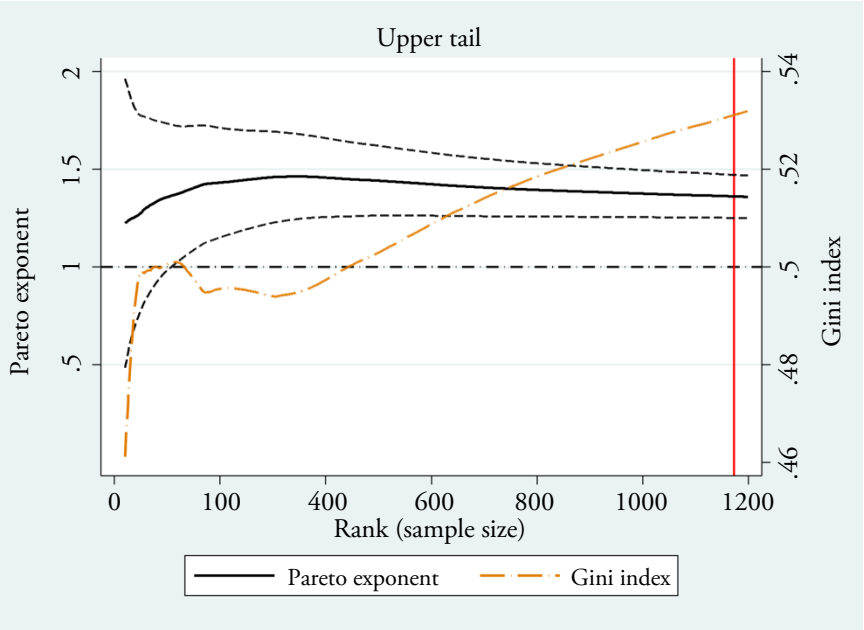
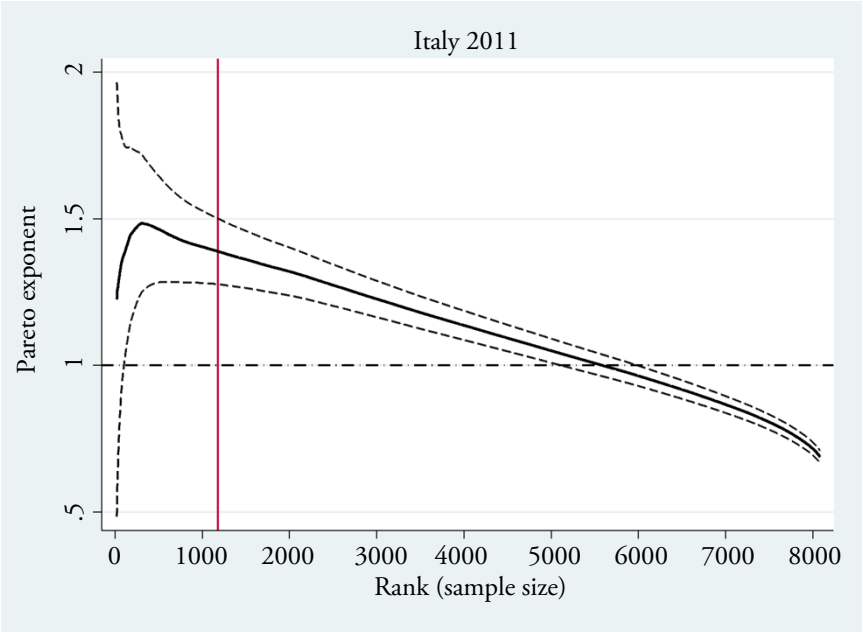
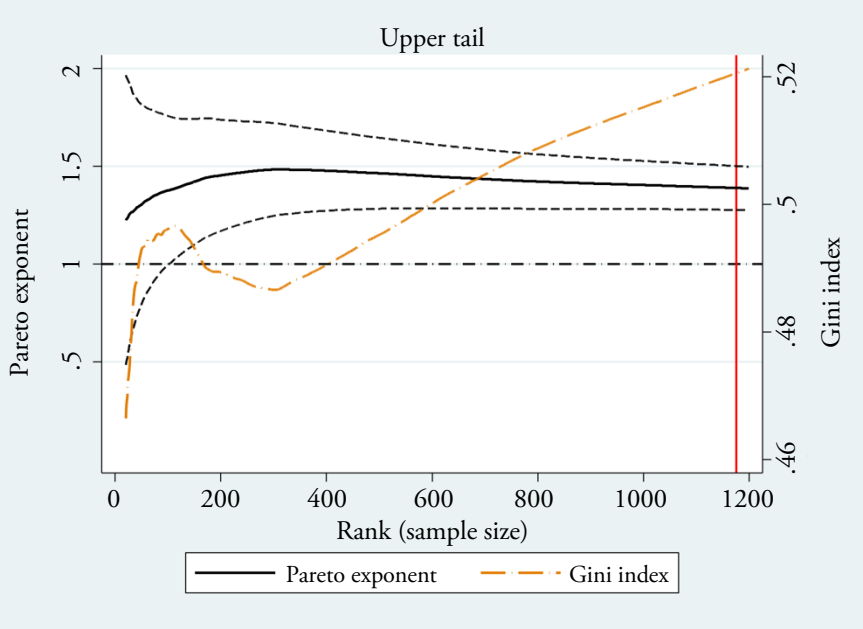


FIGURE 3. CONT.
Pareto exponent by sample size, Italian municipalities in 2001 and 2011

c) Italian municipalities in 2011 (whole sample)



d) Italian municipalities in 2011 (zoom)



Notes: The Pareto exponent is estimated utilising Gabaix and Ibragimov's Rank-1/2 estimator. Dashed lines represent the standard errors calculated by applying Gabaix and Ioannides's (2004) corrected standard errors: $\text{GI s. e.} = \hat{a}.(2/\text{N})^{1/2}$, where N is the sample size. The vertical red lines indicate the threshold of the Pareto upper-tail, determined using Clauset et al. (2009)'s methodology. There are 1,173 municipalities in the upper-tail in 2001 (the population threshold is 9,358) and 1,176 in 2011 (the population threshold is 10,282).

We can extract some general conclusions for the three countries. First, Zipf's law holds only for small sample sizes (approximately up to 100 cities) in all cases except for Spain in 2001, in which holds for all the upper-tail (Figure 2(b)), because the value of one only falls within the confidence bands for those small

sample sizes. Second, the estimation of the Pareto exponent for the upper-tail is independent of the sample size; in fact, it is not significantly different from a horizontal line. This finding implies that Proposition 1 in Eeckhout (2004) does not hold in the upper-tail in any case. Furthermore, this result also implies that the Pareto exponent does not provide any information about the degree of inequality in the upper-tail; the Pareto exponent is a flat line, which indicates that adding more cities does not increase or decrease inequality in city sizes. On the other hand, the Gini index clearly changes with the sample size. In most cases, the Gini index increases with increasing sample size (Figures 1(b), 1(d), 2(b) and 2(d)). That result points to increasing inequality with increasing sample size. However, the case of Italy is special: for sample sizes between 100 and 400 the Gini index decreases (Figures 3(b) and 3(d)). That finding implies growing homogeneity among city sizes; however, for sample sizes above the 400 largest cities, inequality increases like in the other countries.

Third, for sample sizes that extend beyond the upper-tail distribution (delineated by the vertical red line in the graphs) we observe exactly the same pattern in all cases: the estimate of the Pareto exponent decreases with sample size (Figures 1(a), 1(c), 2(a), 2(c), 3(a) and 3(c)). That is, once the sample size enters the lognormal body of the distribution and the lognormality assumption is fulfilled, Proposition 1 of Eeckhout (2004) is valid.

5. Conclusions

The current paradigm in the city size distribution literature states that, although most of the city size distribution is nonlinear, the Pareto distribution (and Zipf's law) holds for the largest cities (Giesen et al., 2010; Ioannides and Skouras, 2013). However, we found that Zipf's law holds only for small samples of the largest cities, not for the entire upper-tail.

Furthermore, we found that the Pareto exponent decreases with the sample size, but only for the lognormal body of the distribution. That finding partially supports Proposition 1 of Eeckhout (2004). For the upper-tail distribution, the exponent does not vary with the sample size. This result supports the current approach in the literature to estimating Zipf's law by considering only the upper-tail distribution, because if the upper-tail is Pareto-distributed the estimated exponent will be quite stable for all sample sizes within the upper-tail. However, a potential issue that remains is that, as the method for defining the Pareto upper-tail is still an open debate (Fazio and Modica, 2015; Schluter, 2021), different population thresholds (and sample sizes) across methods can lead to biased estimates of the Pareto exponent.

Acknowledgments

This research was funded by the Spanish Ministerio de Ciencia e Innovación and Agencia Estatal de Investigación, MCIN/AEI/10.13039/501100011033 (projects PID2020-114354RA-I00 and PID2020-112773GB-I00), DGA (project S39_20R, ADETRE research group), and ERDF. Comments received from one anonymous referee have improved the version originally submitted. All remaining errors are ours.

References

Arshad, S., Hu, S., and Ashraf, B. N. (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*, 75–92.

Arshad, S., Hu, S., and Ashraf, B. N. (2019). Zipf's law, the coherence of the urban system city size distribution: Evidence from Pakistan. *Physica A: Statistical Mechanics and its Applications*, *513*, 87–103.

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

Beirlant, J., P. Vynckier, and J. L. Teugels, (1996). Tail Index estimation, Pareto quantile plots, and regression diagnostics. *Journal of the American Statistical Association*, *91*(436), 1659–1667.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*, 661–703.

Clemente, J., González-Val, R., and Olloqui, I. (2011). Zipf's and Gibrat's laws for migrations. *The Annals of Regional Science*, *471*, 235–248.

Cottineau, C., (2017). MetaZipf. A dynamic meta-analysis of city size distributions. *PLoS ONE*, *8*, 1–22.

Cristelli, M., Batty, M., and Pietronero, L. (2012). There is more than a power law in Zipf. *Scientific Reports*, *2*, 1–7.

Düben, C., and Krause, M. (2021). Population, light, and the size distribution of cities. *Journal of Regional Science*, *61*(1), 189–211.

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

Fazio, G., and Modica, M. (2015). Pareto or log-normal? Best fit and truncation in the distribution of all cities. *Journal of Regional Science*, *55*(5), 736–756.

Fernholz, R. T., and Kramer, R. (2024). Racing to Zipf's law: Race and metropolitan population size 1910-2020. *Journal of Regional Science*, *64*(3), 649–670.

Gabaix, X., (1999). Zipf's law for cities: An explanation. *Quarterly Journal of Economics*, *114*(3), 739–767.

Gabaix, X., and Ibragimov, R. (2011). Rank-1/2: A simple way to improve the OLS estimation of tail exponents. *Journal of Business & Economic Statistics*, *29*(1), 24–39.

Gabaix, X., and Ioannides, Y. M. (2004). The evolution of city size distributions. In J. V. Henderson, and J. F. Thisse, (eds.), *Handbook of Regional and Urban Economics* (vol. 4, 2341-2378). Elsevier Science, North-Holland.

Giesen, K., Zimmermann, A., and Suedekum, J. (2010). The size distribution across all cities – double Pareto lognormal strikes. *Journal of Urban Economics*, *68*, 129–137.

González-Val, R., (2010). The evolution of US city size distribution from a long term perspective (1900-2000). *Journal of Regional Science*, *50*(5), 952–972.

González-Val, R., (2016). War size distribution: Empirical regularities behind conflicts. *Defence and Peace Economics*, 276, 838–853.

González-Val, R., Ximénez-de-Embún, D. P., and Sanz-Gracia, F. (2024). A long-term, regional-level analysis of Zipf's and Gibrat's laws in the United States. *Cities*, *149*, 104946.

Hackmann, A., and Klarl, T. (2020). The evolution of Zipf's Law for U.S. cities. *Papers in Regional Science*, *99*(3), 841–852.

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

Ioannides, Y. M., and Skouras, S. (2013). US city size distribution: Robustly Pareto, but only in the tail. *Journal of Urban Economics*, *73*, 18–29.

Luckstead, J., and Devadoss, S. (2014). Do the world's largest cities follow Zipf's and Gibrat's laws? *Economics Letters*, *125*(2), 182–186.

Nitsch, V., (2005). Zipf zipped. *Journal of Urban Economics*, *57*, 86–100.

Peng, G. (2010). Zipf's law for Chinese cities: Rolling sample regressions. *Physica A*, *389*, 3804–3813.

Peña, G., and Sanz-Gracia, F. (2021). Zipf's exponent and Zipf's law in the BRICS: A rolling sample regressions approach. *Economics Bulletin*, *41*, 2543–2549.

Puente-Ajovín, M., Ramos, A., and Sanz-Gracia, F. (2020). Is there a universal parametric city size distribution? Empirical evidence for 70 countries. *The Annals of Regional Science*, *65*, 727–741.

Schluter, C., (2021). On Zipf's law and the bias of Zipf regressions. *Empirical Economics*, *61*(2), 529–548.

Zanette, D. H., (2006). Zipf's law and the creation of musical context. *Musicae Scientiae*, *101*, 3–18.

Zipf, G. K., (1949). *Human behavior and the principle of least effort.* Addison-Wesley.

ORCID

Rafael González-Val https://orcid.org/0000-0002-2023-5726

Fernando Sanz-Gracia https://orcid.org/0000-0003-3725-0022