Economics Letters 120 (2013) 232–237

Contents lists available at SciVerse ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

The size distribution of US cities: Not Pareto, even in the tail

Marco Bee [a], Massimo Riccaboni [b,c,*], Stefano Schiavo [d,a,e]

[a] Department of Economics and Management, University of Trento, via Inama 5, 38122 Trento, Italy
[b] LIME, IMT Institute for Advanced Studies, Piazza San Ponziano 6, 55100 Lucca, Italy
[c] Department of Managerial Economics, Strategy and Innovation (MSI), K.U. Leuven, Naamsestraat 69, 3000 Leuven, Belgium
[d] School of International Studies, University of Trento, Italy
[e] OFCE-DRIC, France

HIGHLIGHTS

- The entire distribution of US city size is neither a Pareto one nor a lognormal one.
- Based on multiple tests, we find that the largest US cities are not Pareto distributed.
- Tests on real data and samples draws from a lognormal distribution yield similar Pareto tails.
- Bootstrap exercises show that the length of the Pareto tail shrinks by increasing sample size.

ARTICLE INFO

Article history:
Received 3 March 2013
Received in revised form
17 April 2013
Accepted 19 April 2013
Available online 24 April 2013

JEL classification:
C14
D30
R12
R23

Keywords:
City size distribution
Zipf distribution
Power-law
Lognormal distribution
Maximum entropy

ABSTRACT

We question the claim that the largest US cities are Pareto distributed. We show that results of multiple tests on real data are similar to those obtained when the true distribution is lognormal, and largely depend on sample sizes.

1. Introduction

Recently, a lively debate has emerged on whether city size data are better approximated by a Pareto distribution or by a lognormal one (Eeckhout, 2004; Levy, 2009; Eeckhout, 2009; Malevergne et al., 2011; Rozenfeld et al., 2011; Ioannides and Skouras, 2013).

Beside the specific intellectual curiosity the issue may raise, there are broader theoretical reasons for investigating the matter, as competing models yield different implications. Indeed, while the seminal paper by Gabaix (1999) predicts a Zipf's law, Eeckhout (2004) proposes an equilibrium theory to explain the lognormal distribution of cities. This debate is hampered by the difficulty to distinguish lognormal versus Pareto tails (Embrechts et al., 1997; Bee et al., 2011). Moreover, the contention is partly based on the difficulty of properly defining what a city is and, empirically, what is the correct measure to use.[1]

* Corresponding author at: Department of Managerial Economics, Strategy and Innovation (MSI), K.U. Leuven, Naamsestraat 69, 3000 Leuven, Belgium.
E-mail addresses: marco.bee@unitn.it (M. Bee), massimo.riccaboni@kuleuven.be (M. Riccaboni), stefano.schiavo@unitn.it (S. Schiavo).

0165-1765/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.econlet.2013.04.035

[1] This point is made in Rozenfeld et al. (2011), who propose a new methodology to define cities based on microdata and a clustering algorithm that identifies a city as the maximal connected cluster of populated sites. By applying this methodology to both US and UK data, the authors find that a Zipf's law approximates well the

While early studies focus on the largest US Metropolitan Statistical Areas (MSAs) only, recent contributions use data for all the populated places of the US and other countries. By so doing, Eeckhout (2004) shows that the size distribution of US cities is a lognormal one, not a power-law one as previously thought (at least since Zipf, 1949). A few years later, Levy (2009) acknowledged that the body of the city size distribution is well approximated by a lognormal distribution, but claimed that there are significant departures in the upper tail. In particular, the top 0.6% of the distribution, i.e., the MSAs, appear to fit better a power-law distribution. Eeckhout (2009) replied to these new findings by highlighting potential problems associated with the procedures used by Levy (2009) to identify the power-law tail.[2] Recently Malevergne et al. (2011) have suggested that the debate rests on the small power of the tests employed by both Eeckhout (2004) and Levy (2009). They claim the issue can be definitely settled by adopting a better testing procedure, namely the uniformly most powerful unbiased test of the exponential versus truncated normal distribution in log-scale developed by del Castillo and Puig (1999). Last, Ioannides and Skouras (2013) applied a switching model and found that the distribution is lognormal in the body, but robustly Pareto in the upper tail (top 5%).

We contribute to this debate by providing new evidence based on a through analysis of the tail behavior of the distribution and a number of counterfactual exercises. We conclude that the power-law behavior of the upper tail is less robust than previously claimed, due to the limited power of the available statistical tests (Perline, 2005).

2. Data and methodology

2.1. Data

We analyze the distribution of US city size: information is derived from the 2010 Census Data collected by the US Census Bureau. The elementary unit of analysis, corresponding to disaggregate data, is the population of 6 127 259 census blocks. These figures are then aggregated into administrative units that represent populated places. As in Eeckhout (2004), we take populated places as the unit of analysis at the aggregate level.[3] Since it has been argued that the way cities are defined (i.e., the way elementary units are aggregated) is not neutral with respect to the shape of the resulting city size distribution, we perform our analysis using both the administrative definition of cities and the clusters identified by Rozenfeld et al. (2011).[4]

2.2. Testing for a power-law tail

Discriminating between power-law (Pareto) and lognormal tail behavior is a difficult task. Although asymptotically the two distributions are mathematically different, the convergence of the lognormal to the asymptotic distribution is extremely slow (Perline, 2005), so the difference may be very small, to the extent that they are often practically indistinguishable for any finite sample size.

Given these difficulties, several tests have been proposed: the uniformly most powerful unbiased (UMPU) test developed by del Castillo and Puig (1999) and used by Malevergne et al. (2011); the maximum entropy (ME) test by Bee et al. (2011); and the test proposed by Gabaix and Ibragimov (GI henceforth; see Gabaix and Ibragimov, 2011).

The UMPU test is based on the fact that the logarithm of a truncated lognormal distribution is truncated normal, and the logarithm of a Pareto distribution is exponential. del Castillo and Puig (1999) have shown that the likelihood ratio test for the null hypothesis of exponentiality against the alternative of truncated normality is given by the clipped sample coefficient of variation $\bar{c} = \min\{1, \hat{\sigma}/\hat{\mu}\}$ of the logarithms of the observations, where μ and σ are the parameters of the truncated normal. The UMPU test only compares the null of a power-law distribution against the alternative of a lognormal distribution, and rejects the null hypothesis for small values of the coefficient of variation c. However, the coefficient of variation does not uniquely identify distributions with power-law tails. This implies that the UMPU test works well (i.e., its power is high) in cases such as the lognormal–Pareto mixture, namely when the data-generating process is such that $c \geq 1$ above the threshold that separates the lognormal and the Pareto distributions and $c < 1$ below the threshold (Bee et al., 2011). On the other hand, if the distribution below the threshold is not a power-law one but nonetheless has $c \geq 1$, as happens, for example, for the Weibull distribution with shape parameter equal to 1, the UMPU test is completely unreliable. A case that illustrates this point is the aggregate city size distribution studied below (see Section 3).

The ME approach entails maximizing the Shannon information entropy under k moment constraints $\mu^i = \hat{\mu}^i$ $(i = 1, \ldots, k)$, where $\mu^i = \mathrm{E}[T(x)^i]$ and $\hat{\mu}^i = \frac{1}{n}\sum_j T(x_j)^i$ are the ith theoretical and sample moments, n is the number of observations, and T is the function defining the characterizing moment.[5] The solution (that is, the ME density) takes the form $f(x) = e^{-\sum_{i=0}^{k} \lambda_i T(x)^i}$. If $T(x) = x$, the logarithm of the Pareto (i.e., the exponential) distribution is an ME density with $k = 1$, whereas the logarithm of the lognormal (i.e., the normal) distribution is an ME with $k = 2$. A log-likelihood ratio (llr) test of the null hypothesis $k = k^*$ against $k = k^* + 1$ is given by

$$\mathrm{llr} = -2n\left(\sum_{i=0}^{k^*+1} \hat{\lambda}_i \hat{\mu}^i - \sum_{i=0}^{k^*} \hat{\lambda}_i \hat{\mu}^i\right).$$

From standard limiting theory, the llr test is asymptotically χ^2_1 and is optimal (Cox and Hinkley, 1974; Wu, 2003).[6]

The ME test is a by-product of a more general non-parametric approach to density estimation. It can indeed be shown that, when the whole distribution is of interest, the method can be used for fitting the best approximating density, with the optimal k found by the llr test (Bee, 2013). Referring the interested reader to Wu (2003) for details, the main advantages of the technique are that (i) it delivers the best (according to the ME criterion) approximating density, and allows one to assess whether it belongs to certain parametric families; (ii) if the true distribution is a Pareto one, it provides an estimate of the shape parameter; (iii) it does not consider a single alternative model, so pitfalls such as the one discussed for the UMPU test in Section 3 below are avoided.

distribution of 1947 US cities with more than 12,000 inhabitants (about 1000 cities with more than 5000 inhabitants for the UK).

[2] Specifically, Eeckhout (2009) suggests that the graphical procedure based on visual inspection of a log–log plot introduces significant biases in the right tail of the distribution.

[3] In the rest of the paper, the terms city and populated place are used interchangeably.

[4] Data on clusters are available at http://lev.ccny.cuny.edu/~hmakse/soft_data.html.

[5] The two most common cases are $T(x) = x$ and $T(x) = \log(x)$, corresponding respectively to arithmetic and logarithmic moments.

[6] The routines for implementing the UMPU and ME tests are available at https://sites.google.com/site/sschiavo7788/home/software.