Journal of Urban Economics 73 (2013) 18–29

Contents lists available at SciVerse ScienceDirect

Journal of Urban Economics

www.elsevier.com/locate/jue

US city size distribution: Robustly Pareto, but only in the tail ☆

Yannis Ioannides [a,*], Spyros Skouras [b]

[a] Tufts University, Department of Economics, Medford, MA 02155, USA
[b] Athens University of Economics and Business, Department of International and European Economic Studies, 76 Patision St., 104 34 Athens, Greece

ARTICLE INFO

Article history:
Received 8 March 2010
Revised 22 June 2012
Available online 7 July 2012

JEL classification:
D30
D51
J61
R12
C24

Keywords:
Gibrat's law
Zipf's law
Pareto law
Upper tail
Mixture of distributions
Switching regressions
Urban evolution
Urban hierarchy

ABSTRACT

We establish empirically using three different definitions of US cities that the upper tail obeys a Pareto law and not a lognormal distribution. We emphasize estimation of a switching point between the body of the city size distribution (which includes most cities) and its upper tail (which includes most of the population). For the 2000 Census Places data, in particular, our preferred model suggests that switching from a lognormal to a Pareto law occurs within a narrow confidence interval around population 60,290, with a corresponding Pareto exponent of 1.25. *Most cities* obey a lognormal; but the upper tail and therefore *most of the population* obeys a Pareto law. We obtain qualitatively similar results for the upper tail with the Area Clusters data of Rozenfeld et al. (2011), and the US Census combined Metropolitan and Micropolitan Areas data, though the shape of that distribution at smaller sizes is sensitive to the definition used.

1. Introduction

The study of city size distributions continues to attract attention. Numerous statistical and econometric investigations point to an important similarity across very different economies regarding the upper tail, thus suggesting that knowledge of the underlying probability laws may improve our understanding of the urban structure worldwide. Understanding the upper tail is particularly important because that is where most of the population lives. For example, in the data on US Census Places used by Eeckhout (2004) and Levy (2009), only 15% of all places have population above 10,000 people, but they accommodate 80% of the population. More dramatically, 1% of all places are larger than 100,000 people but accommodate 63% of the population.

Eeckhout (2004) made the notable observation that the population distribution of US Places data, a proposed definition of US cities that extends the entire range of city sizes, is best empirically analyzed with a lognormal distribution. However, this conclusion generated controversy: Levy (2009) uses a mostly graphical analysis to counter Eeckhout (2004) claim that populations of places are lognormally distributed throughout the range of observed size distributions. Levy's evidence suggests that the tail is in fact Pareto law distributed, contrary to Eeckhout (2004). Eeckhout (2009) in response points to several drawbacks in Levy's critique and concludes reaffirming his original claim that "the tail of the distribution is indeed lognormal" (*ibid*., p. 1676).

The main purpose of this paper is to model econometrically the behavior of the upper tail of city size distributions when data for the entire size range are available and to definitively characterize the behavior of the tail of US city sizes. To do this, we introduce a new statistical model and related tests and statistical procedures tailored for this purpose. Our second goal is to examine the robustness of our finding by going beyond the Places data by means of all other city size data sets that we could avail ourselves of that extend across a broad size range and might reflect reasonable measurements of what economists understand as "cities." Indeed, it is not clear that Places, in spite of their recent popularity, reflect the most appropriate definition of city.

☆ We would like to thank Gilles Duranton (The Editor) and several anonymous referees for detailed suggestions that have greatly improved this paper.

* Corresponding author.
E-mail addresses: yannis.ioannides@tufts.edu (Y. Ioannides), skouras@aueb.gr (S. Skouras).

0094-1190/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jue.2012.06.005

The main difficulty in analyzing the tail behavior of cities using data that extends across all sizes is that this requires an assumption about where the body ends and the tail begins. If this is *ad hoc,* the conclusions may be seriously biased. We avoid this problem because we use a distribution that models separately and parametrically the 'body' of the distribution as lognormal and the 'tail' of the distribution as Pareto. We thus allow the data to determine where, if anywhere, a switch occurs from one behavior to the other. This provides a simple solution to the otherwise complex statistical problem of jointly estimating and conducting inference about Pareto tail exponents, cut-offs and related parameters [see Caers and Dyck (1999), Handcock and Jones (2004), Clauset et al. (2009), Arnold and Press (1989)]. Our approach relies on maximum likelihood methods which make inference and specification tests straightforward, while the parametric nature of our approach is tailored around economically meaningful parameters. For example, we report point estimates and standard errors for the threshold population level at which Pareto behavior begins, which to the best of our knowledge we are the first to do.

Our benchmark data is the US Census Places data which have been the focus of most recent research in this area. By using a more powerful approach than Eeckhout (2009), we find economically and statistically very significant deviations from lognormality in the upper tail. We apply this approach to two other data sets and obtain the striking result that: the tail of large cities is very robust across definitions and obeys consistently a Pareto distribution with exponent not far from one (Zipf's law), broadly speaking, and yet the same is not true for the body of the size distribution. According to our preferred benchmark model, the switch to the Pareto tail behavior occurs in the population range of 30 to 60 thousand for all definitions of cities, while the shape of the body is extremely sensitive to the city definition. Indeed, researchers who do not consider Places to be an attractive US city definition might view our results as evidence that lognormality is altogether an unsatisfactory description of the size distribution at *any* size.

On a more positive note, urban economists have dwelled on the practical difficulty of defining cities, let alone measuring their size, and have therefore held doubts as to whether reported empirical regularities are actually reliable. Our showing that the shape of the distribution's tail is very robustly Pareto across widely differing definitions of cities, will help allay such worries. At the same time, considerable progress in defining and measuring cities must be made before fully credible empirical analyses are feasible for cities of populations less than a few tens of thousands. Indeed, it may be meaningful to distinguish between cities and smaller settlements (e.g. because production technologies, agglomeration economies and congestion externalities are very different in a "city" with a few residents than in a city of size ten million), but this is a distinction that city size distribution literature has not yet fully emphasized. There exist, of course, notable studies that focus on particular ranges of size, like Henderson (1997) who considers medium size cities and Ades and Glaeser (1995) who study the urban economy by distinguishing between primate cities and all other cities.

Our results have bearing on theoretical models that aim to explain city size distributions [Gabaix (1999), Rossi-Hansberg and Wright (2007), Eeckhout (2004), Skouras (2010), Ioannides (2012), Ioannides et al. (2008)]. The finding of a robust Pareto tail but varying estimates about the shape of the body suggests it is very important that a good explanation of the Pareto–Zipf tail should be consistent with a broad range of shapes for the body of the size of the distribution. The empirical evidence we offer suggests we should be somewhat skeptical of explanations of Pareto–Zipf tails that are heavily tied to some specific functional form for the body, such as lognormality. Skouras (2010) proposes an economic explanation for a size distribution that is "flexible" in the body but leads to Zipf in the upper tail. He shows that heterogeneity in urban growth dynamics across cities leads to a steady state size distribution that has a tail close to Zipf's law but an arbitrary body that will depend on details of each city's growth process. It follows that a version of Gabaix's (1999) model with heterogeneity across cities can reproduce the observed transition from lognormality to Pareto behavior making heterogeneity an important neglected feature of the data. Heterogeneity across cities is key to explaining the qualitatively different behavior of the density of size distribution of cities at different size ranges.

The remainder of this paper starts with a discussion of three alternative data sets for the United States, which have been used by several researchers in the recent literature. Section 3 discusses empirical models for city size distributions that have been used recently and presents our new distribution function which allows for switching between a lognormal and a Pareto across its range. Section 4 turns to our empirical analysis. We begin by showing that the Lilliefors test, employed by Eeckhout (2009), is too weak as a method for testing the hypothesis of lognormality in the tail, which is the aim of Eeckhout (2004, 2009). Next we provide several alternative tests that reject lognormality in the tail. Across all data sets that we employ, we estimate a switch from lognormality to Pareto behavior around cities of population in the range of 30 to 60 thousand depending on the definition of city. We discuss the merits of our switching model and compare it to contemporaneous related work. We conclude that our lognormal–Pareto model is preferable to a simple lognormal alone according to several formal tests and across *all* data sets we analyze, thus providing definitive evidence in favor of a Pareto tail. Section 5 discusses the implications of our findings for the theoretical modeling of urban systems. We conclude in Section 6.

2. Data

The heart of our paper is empirical results using three alternative definitions for US cities, and therefore we consider it important to provide details of their definitions. Eeckhout (2004) pioneered use of the US Census *Places* data. Places range from the smallest to the largest city sizes. It is our main benchmark and we take it up first.

2.1. US Census Places

As defined by the US Census Bureau, places are concentrations of population that may or may not have legally prescribed limits, powers or functions. They must have a name and be locally recognized. They include census designated places (CDPs), consolidated cities, and incorporated places.[1] Starting with the 2000 Census, for the first time, CDPs did not need to meet a minimum population threshold to qualify for tabulation of census data.[2] Consequently, observations on 25,358 places in the 2000 US Census range in population from 1 to over 8 million inhabitants, 'including cities, towns, and villages' [Eeckhout (2004), p. 1431]. As the detailed description of the places data in the US Census literature[3] shows, incorporated places must adhere to specific criteria for 'incorporation' provided for by legislation that varies across US states. This creates an obvious source of arbitrariness and bias in the data definition, which has not been fully recognized by previous users of the data. For example, incorporated places in Massachusetts must have more than 12,000

[1] An *incorporated place* is a "governmental unit incorporated under state law as a city, town (except in the New England states, New York, and Wisconsin), borough (except in Alaska and New York), or village and having legally prescribed limits, powers, and functions". The unincorporated counterpart is called a census-designated place; such places lack their own local authority but otherwise resemble incorporated places.

[2] http://www.census.gov/geo/www/tiger/glossry2.pdf, Appendix A, p. A.17.

[3] Geographical Areas Reference Manual, US Census, Chapter 9,http://www.census.gov/geo/www/GARM/Ch9GARM.pdf.