US City Size Distribution: Robustly Pareto, but Only in the Tail

Yannis Ioannides[1] **Spyros Skouras**[2]

First version: September 2009

This version:[3] June 18, 2012

[1]Email: yannis.ioannides@tufts.edu; address: Tufts University, Department of Economics, Medford, Massachusetts 02155, USA

[2]Email: skouras@aueb.gr; address: Athens University of Economics and Business; Department of International and European Economic Studies; 76 Patision st, 104 23, Athens, Greece.

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

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 the distribution at smaller sizes is sensitive to the definition used.

Keywords: Gibrat's law, Zipf's law, Pareto law, upper tail, mixture of distributions, switching regressions, urban evolution, urban hierarchy.

JEL codes: D30, D51, J61, R12, C24.

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 cover 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's (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.

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 van Dyck (1999), Handcock and Jones (2004), Clauset, Shalizi and Newman (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 economics has 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, we 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)]. 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 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, and Section 6 discusses the merits of our switching model and compares it to contemporaneous related work. We conclude in section 7.

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 (CDP), consolidated cities, and incorporated places.[1] Starting with the 2000 Census, for the first time, CDPs did not need

[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 authorty but otherwise resemble incorporated places.

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 inhabitants, whereas incorporated places data in Maryland need only have 300 inhabitants. That said, the US Places data by virtue of having been used extensively has become an interesting benchmark from which to start any empirical analysis.

2.2 Area Clusters

Rozenfeld, Rybski, Gabaix, and Makse (2011) introduce, most recently, a City Clustering Algorithm (CCA) to build cities "from the bottom up," in their words. The algorithm defines a 'city' as a maximally connected cluster of populated sites defined at high spatial resolution. Namely, a population cluster is made of contiguous populated sites within a prescribed distance that cannot be expanded: all sites immediately outside the cluster have a population density below a cutoff threshold. Their method defines an agglomeration as a maximally connected cluster of potentially many cells. Cities in this view are clusters of population, i.e., adjacent populated geographical spaces, defined in terms of a single parameter (the cell size). The original US data consist of 61,224 sites with populations ranging from 1500 to 8000 and defined by Federal Information Processing Standards (FIPS), which are in turn associated with corresponding populations provided by the U.S. Census Bureau. Our analysis is based on a cell size of 3 *km* by 3 *km* also used as a benchmark by Rozenfeld *et al.* (2011). We refer to the data thus constructed by Rozenfeld *et al.* as *Area Clusters*. We view this definition of city as focusing on spatially units.

[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`.

2.3 Metropolitan and Micropolitan Areas

A variety of administrative divisions are used by the US Bureau of the Census to approximate the notion of a 'city' as an economic unit. According to the latest update and revision of these definitions in the *Federal Register*, Vol. 75, No. 123, Monday, June 28, 2010, Notices, pp. 37251–37252,[4] the general concept of a metropolitan area (MA), as maintained by the US Office of Management and Budget (OMB), aims at a concept of a contiguous area of relatively high population density and rests on the concept of a "Core Based Statistical Area" (CBSA). This is 'a statistical geographic entity consisting of the county or counties associated with at least one *core* (*urbanized area* or *urban cluster*[5]) of at least 10,000 population, plus adjacent counties having a high degree of social and economic integration with the core as measured through commuting ties with the counties containing the core.' Metropolitan and Micropolitan Statistical Areas are the two categories of Core Based Statistical Areas, with metropolitan areas being associated with at least one urbanized area that has a population of at least 50,000, and micropolitan areas being associated with at least one urban cluster that has a population of at least 10,000, but less than 50,000. Both metropolitan and micropolitan areas are defined in terms of groups of counties that have a high degree of social and economic integration with the central county or counties as defined in terms of commuting patterns. The usual interpretation of micropolitan areas is that they are "small metropolitan areas".[6]

Although judgement calls are involved in this definition by the OMB, the US metropolitan statistical area has been the dominant concept of city used by empirical researchers. However, recently and especially since Duranton (2007), the micropolitan area concept has gained popularity in empirical work, bringing the population threshold down to a theoretical minimum of 10,000 (in the data we use the smallest city in fact has a population of 12,400).

[4] `http://www.whitehouse.gov/sites/ default/files/omb/assets/fedreg_2010/06282010_metro_standards-Complete.pdf`

[5] *Urbanized area* is a geographic entity delineated by the Census Bureau, consisting of densely settled census tracts and blocks and adjacent densely settled territory that together contain at least 50,000 people. *Urban cluster* is a statistical geographic entity delineated by the Census Bureau, consisting of densely settled census tracts and blocks and adjacent densely settled territory that together contain at least 2,500 people.

[6] The counties containing the *core urbanized* area are known as the *central counties* of the metropolitan statistical area. Additional surrounding counties (known as *outlying* counties) can be included in the metropolitan statistical area, provided these counties have strong social and economic ties to the central counties as measured by commuting and employment.

Metropolitan and micropolitan areas are *not* subsets of the Places data and the relation between the two definitions is not obvious. The Metropolitan and Micropolitan definition of city, although clearly an imperfect proxy for urban economic entities, is increasingly being used. We refer to these data as Metro & Micro areas, or Metro/Micro for short.

3 Proposed density models for distinguishing between tail and body

3.1 The Baseline Switching Model

We specify our switching model so that when the switching population level parameter τ tends to infinity, our model converges to the lognormal, while for $\tau = 1$, the lowest urban place population in our data, our density becomes the Pareto. However, our switching approach could easily be modified to accommodate distributions different from the lognormal for the body (this might be appropriate for the Metro/Micro areas and Area Clusters data, as our empirical analysis does in fact suggest) or a different tail distribution (which we will use below). In particular, the density function we adopt is defined in terms of a lognormal density function,

$$f\left(x;\mu,\sigma\right)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(\ln x-\mu\right)^{2}}{2\sigma^{2}}\right),\tau>x>0, \tag{1}$$

where (μ,σ) are the mean and standard deviation of $\ln x$, and of a Pareto density function kernel,

$$g\left(x;\zeta,\tau\right)=\frac{1}{x^{1+\zeta}},x\geq\tau,\zeta>0, \tag{2}$$

where (ζ,τ) are the exponent and cutoff point, respectively. The density function we posit is as follows:

$$h\left(x;\mu,\sigma,\tau,\zeta\right)=\left\{\begin{array}{c} b(\tau,\mu,\sigma,\zeta)\ f\left(x;\mu,\sigma\right),\ \tau>x>0 \\ a(\tau,\mu,\sigma,\zeta)\ b(\tau,\mu,\sigma,\zeta)\ g(x;\zeta,\tau),\ x\geq\tau \end{array}\right\}, \tag{3}$$

where the term $a(\tau, \mu, \sigma, \zeta)$, a normalization constant, is included to scale $g(x; \zeta, \tau)$ so that at the switching point τ from lognormality to Pareto, the density function $h(\cdot)$ is continuous. This requires:

$$\begin{aligned} a(\tau, \mu, \sigma, \zeta) &= \frac{f(\tau; \mu, \sigma)}{g(\tau; \zeta, \tau)} \\ &= f(\tau; \mu, \sigma, \tau)\, \tau^{1+\zeta}. \end{aligned} \tag{4}$$

The term $b(\tau, \mu, \sigma, \zeta)$, a suitably defined function of parameters that serves as a normalization constant, is required to ensure that $h(\cdot)$ is a density, i.e. integrates to one.[7] Substituting (4) into (3) and setting the integral of h across its range to be equal to one, we have:

$$b \int_0^{\tau} f(x; \mu, \sigma)\, dx + b f(\tau; \mu, \sigma, \tau)\, \tau^{1+\zeta} \int_{\tau}^{\infty} g(x; \mu, \sigma)\, dx = 1.$$

Using (1) and (2) and solving for b yields:

$$b(\tau, \mu, \sigma, \zeta) \equiv \frac{1}{\Phi(\tau; \mu, \sigma) + f(\tau; \mu, \sigma, \tau)\, \frac{\tau}{\zeta}}; \tag{5}$$

where $\Phi(x; \mu, \sigma)$ is the lognormal cumulative distribution function with parameters (μ, σ) corresponding to (1). The distribution satisfies the usual regularity conditions required for maximum likelihood estimation and inference. Parameter τ specifies where, if anywhere, a switch occurs from the lognormal to a Pareto. The other parameters μ, σ define the lognormal f. Our model nests Zipf's law, which corresponds to fixing the Pareto exponent ζ to be equal to one.

3.2 Switching Model with Truncation

Our specification of the density function (3) may be easily extended to analyze the Metro & Micro data, which are truncated below at an exogenous value λ, an additional (known) parameter of the model. That is, (3)) only holds above some population level $x \geq \lambda$ where

[7] Such constants appear in all density functions (for example in the lognormal it is $\frac{1}{\sigma\sqrt{2\pi}}$). If X is truncated this constant will also depend on its range as we will see in the next subsection.

λ is an exogenous censoring / truncation point such that $\lambda < \tau$.[8] With truncation at λ, the only difference is that the normalization constant needs to be adapted since the density must integrate to one over the range $[\lambda, \infty)$ instead of $[0, \infty)$. It is straightforward to show using the approach of the previous subsection that instead of (5), b must now satisfy:

$$b(\tau, \mu, \sigma, \zeta, \lambda) = \frac{1}{\Phi(\tau; \mu, \sigma) - \Phi(\lambda; \mu, \sigma) + f(\tau; \mu, \sigma, \tau)\frac{\tau}{\zeta}}. \tag{6}$$

3.3 A Variation of the Baseline Switching Model

As an alternative formulation of a switching model, we develop a variant of the mixture model inspired by Combes, Duranton, Gobillon, Puga, and Roux (2012),[9] which generalizes our baseline switching model. In the notation we employ above:

$$h(x; \mu, \sigma, \tau, \zeta) = \left\{ \begin{array}{c} b(\tau, \mu, \sigma, \zeta) f(x; \mu, \sigma), \ \tau > x > 0 \\ b(\tau, \mu, \sigma, \zeta) g^*(x; \zeta, \tau), x \geq \tau \end{array} \right\}, \tag{7}$$

where f is the lognormal density function, as in (1), truncated from above at $\tau$and g^* is a *mixture* of a lognormal and a Pareto density function with lower cut-off at τ:

$$g^*(x; \zeta, \tau, \mu, \sigma, \theta) = \left[\theta a(\tau, \mu, \sigma, \zeta) g(x; \zeta, \tau) + (1 - \theta) c(\tau, \mu, \sigma, \zeta) f(x; \mu, \sigma, \tau)\right], \tag{8}$$

where g stands for the Pareto kernel defined in (2) above.

We now have *three* normalization terms, a, b, c in our model (7-8), i.e. one more than in the definition of the baseline switching model as described by (3). As before, the parameter b is a multiplicative constant to ensure that h integrates to one as required. This means:

$$b\left[\int_{\tau}^{\infty} g^*(x)\, dx + \int_{0}^{\tau} f(x)\, dx\right] = 1. \tag{9}$$

The terms a and c jointly perform two functions: First, as in the baseline switching model

[8] If the truncation point is above τ, no lognormal component is involved.

[9] We thank the editor, Gilles Duranton, for suggesting this extension, which also serves as a robustness check on our approach.

they ensure that $h(x;\mu,\sigma,\tau,\zeta)$ is continuous at τ (a property that the Combes *et al.* model do not impose). This is achieved by imposing the condition:

$$\theta a g(\tau) + (1-\theta) c f(\tau) = f(\tau). \tag{10}$$

Second, a new requirement is that they make θ a readily interpretable parameter. In particular, we impose the condition that:

$$a \int_{\tau}^{\infty} g(x)\, dx = c \int_{\tau}^{\infty} f(x)\, dx. \tag{11}$$

This is useful as it means that the parameter θ controls the proportion of the density in the mixture $g^*(x;\zeta,\tau,\mu,\sigma,\theta)$, defined in (8), which is determined by the Pareto distribution. This normalization is essential as the difference in scale between the Lognormal and the Pareto in the range of the mixture can be very large. In such cases, θ will be misleading if interpreted as a "weight". Furthermore, such differences in scale may cause numerical instability issues during estimation.

The three normalization constants $a,\ b,\ c$ need to be determined simultaneously. Solving (10) and (11) as a system of two linear equations with two unknowns we obtain the values of a and c. Plugging these values into equation (9) gives the value of b. The solutions are:

$$a(\mu,\sigma,\tau,\zeta,\theta) = \zeta\tau^{\zeta}\left(1-\Phi(\tau)\right) f_X(\tau) \left[\left(1-\Phi(\tau)\right)\frac{\theta\zeta}{\tau} + (1-\theta) f_X(\tau)\right]^{-1};$$

$$b(\mu,\sigma,\tau,\zeta,\theta) = \frac{\left(1-\Phi(\tau)\right)\frac{\theta\zeta}{\tau} + (1-\theta) f_X(\tau)}{\theta\left(1-\Phi(\tau)\right) f(\tau) + (1-\theta) f_X(\tau)\left(1-\Phi(\tau)\right) + \Phi(\tau)\left[\left(1-\Phi(\tau)\right)\frac{\theta\zeta}{\tau} + (1-\theta) f_X(\tau)\right]};$$

$$c(\mu,\sigma,\tau,\zeta,\theta) = f(\tau)\left[\left(1-\Phi(\tau)\right)\frac{\theta\zeta}{\tau} + (1-\theta) f_X(\tau)\right]^{-1}.$$

We refer to this model, our variant of the CDGPR-switching model, as the mixture model.

It is straightforward to verify that for $\theta = 1$ we get the baseline switching model introduced in section 3 above. Similarly it is easy to verify that for $\theta = 0$ we get the pure lognormal. For all values of θ off the boundary, our density h has a lognormal body and a tail above a cutoff τ that is a mixture of a Pareto and a lognormal. Moreover, the lognormal in this mixture is the same as the lognormal that describes the body. Therefore we can think of this density as a lognormal, the tail of which is 'contaminated' by a Pareto at the transition τ, above which it has weight θ.

As discussed earlier, the parameter θ is *not* a good description of the importance of the Pareto versus the lognormal for the tail of large cities when the transition occurs at small city sizes (i.e. τ is smaller than the level at which we would reasonably think of 'large' cities) because the relative magnitudes of the lognormal and the Pareto densities change appreciably in the range (τ, ∞). We therefore also report in Table 1 the relative weight of the Pareto versus the Lognormal at a city size of 500,000, that is:

$$\frac{\theta a g\left(500,000\right)}{\left(1-\theta\right) c f\left(500,000\right)}.$$

The original Combes *et al.* (2012) model[10] corresponds to setting $b = 1$ and $a = \zeta\tau^{\zeta}$ in our notation. In other words, the main difference is that we scale g in the CDGPR model

[10] In CDGPR's notation, the model is:

$$f_M\left(x\right) = \mu f_N\left(x\right) + \left(1-\mu\right) f_P\left(x\right)$$

where

$$\begin{aligned} f_N\left(x\right) &= \frac{1}{x\sqrt{2\pi v}} e^{-\frac{(\ln(x)-m)^2}{2v}}; \\ f_P\left(x\right) &= z b^z x^{-z-1} \text{ for } x \geqslant b, = 0, \text{otherwise.} \end{aligned}$$

To the left of $x = b$ the CDGPR mixture model takes the value:

$$\mu f_N\left(b\right) = \mu \frac{1}{b\sqrt{2\pi v}} e^{-\frac{(\ln(b)-m)^2}{2v}}.$$

To the right of $x = b$ it takes the value:

$$\mu f_N\left(b\right) + \left(1-\mu\right)\frac{z}{b} = \mu \frac{1}{b\sqrt{2\pi v}} e^{-\frac{(\ln(b)-m)^2}{2v}} + \left(1-\mu\right)\frac{z}{b}.$$

Since the data do not exhibit a significant discontinuity, empirical estimates are forced towards parameters for which the model also displays limited discontinuity.

This is quantitatively relevant since the first term above tends to be very small for reasonable values of b.

by a factor $f(\tau;\mu,\sigma,\tau)\frac{\tau}{\zeta}$. As discussed in the previous footnote, we do this to eliminate a discontinuity the CDGPR model has at the threshold point τ, to ensure the distribution used is unimodal and make parameter identification easier. These modifications seem justified for our data though in other empirical contexts (such as that of CDGPR) they may not be necessary.

3.4 Variant of the CDGPR-Switching Model with Truncation

As with our baseline model, we also work with a truncated version of the CDGPR-switching model, where truncation is assumed to occur at a known point $\lambda < \tau$. The only change that is necessary in this case is to adapt b to ensure that the density integrates to one. This means substituting (9) with:

$$b\left[\int_{\tau}^{\infty} g^{*}(x)\,dx + \int_{\lambda}^{\tau} f(x)\,dx\right] = 1$$

The values of a and c remain unchanged. Plugging them into the above expression and solving gives:

$$b(\tau,\mu,\sigma,\zeta,\lambda) = \frac{(1-\Phi(\tau))\frac{\theta\zeta}{\tau} + (1-\theta)f(\tau)}{\theta(1-\Phi(\tau))f(\tau) + (1-\theta)f(\tau)(1-\Phi(\tau)) + [\Phi(\tau)-\Phi(\lambda)]\left[(1-\Phi(\tau))\frac{\theta\zeta}{\tau} + (1-\theta)f(\tau)\right]}.$$

Typically, z will be close to one, so the proportionate jump in density around b will be:

$$\frac{(1-\mu)}{\mu}\left(\frac{1}{\sqrt{2\pi v}}e^{-\frac{(\ln(b)-m)^2}{2v}}\right)^{-1}.$$

Since typically $b >> m$, $\frac{1}{\sqrt{2\pi v}}e^{-\frac{(\ln(b)-m)^2}{2v}}$ will typically be very small so the only way to avoid an implausible jump in the density is to have μ close to 1. But when $\mu = 1$, b and z cannot be identified. This is probably the source of the computational issues reported in CDGPR, which ultimately have to do with a very seriously misspecified model which only nests a plausible model (lognormality) as a limiting case. But as is well known when the true parameters are at the model's boundary, estimation can be extremely difficult (e.g. Andrews, 2001). A further issue is that discontinuous density functions do not satisfy the usual regularity conditions required by classical estimation theory.

Another issue is that this distribution will be bimodal if $\ln(b) < \mu$. Finally, because of the different behaviour in the tail, even a very large value of β is consistent with the lognormal playing a negligible role in the tail.

It is straightforward to verify that for $\theta = 1$ we get the same expression as the baseline model.

3.5 Relation to recent empirical analyses of US city size distributions

There exist two other broadly related papers that we became aware after the first version of this paper was distributed online in September 2009. Malevergne, Pisarenko and Sornette (2011) report results on a uniformly most powerful unbiased (UMPU) test for the null of the Pareto distribution against a truncated lognormal, using the 2000 US Places data. Fig. 3, p. 036111-7, *ibid.*, plots the maximum likelihood value against the threshold value beyond which a Pareto law holds. It shows the maximum occurring at 37235, making it the most appropriate estimate of the threshold. Fig. 4, p. 036111-7, *ibid.*, left panel, plots the $p-$ value of the test of the null hypothesis that the size distribution is Pareto against the alternative that it is a truncated lognormal as a function of city rank of the US Places data for 2000. The $p-$value becomes very low for ranks above 1000, thus supporting the lognormal for cities below those ranks. Similar support in favor of the Pareto for large cities is provided by plotting Hill's estimates for the inverse Pareto exponent [Fig. 4, p. 036111-7, *ibid.*, right panel]. These estimates also reject a strict Zipf's law, $\zeta = 1$. We regard these results as complementary to ours and indeed their estimate of the threshold is within the confidence interval of our estimate.

Giesen, Zimmermann and Suedekum (2010) use data from several countries[11] including notably US Places data for 2000, to estimate the so-called "double Pareto lognormal" (DPLN) model. This model, reported by Reed (2002) [see also Reed and Jorgensen (2005)], is obtained from a "pure" Gibrat's law with constant mean and instantaneous variance by sampling it at exponentially distributed times. The resulting law for the values of the process when it is sampled randomly exists in closed form and is a weighted sum of two Pareto distributions, where the weights are the cumulative and the countercumulative functions of a

[11] Germany, 2006; US Places, 2000; France, 2006; Brazil, 2007; Czech Republic, 2009; Hungary, 2009, Italy, 2009; Switzerland, 2008.

certain normal distribution. Specifically, by adapting our own notation, the DPLN density function is written as:

$$h(x) = \frac{\zeta_1\zeta_2}{\zeta_1+\zeta_2}\left[x^{-\zeta_1-1}e^{\zeta_1\mu+\frac{\zeta_1^2\sigma^2}{2}}\Phi\left(\frac{\ln x-\mu-\zeta_1\sigma^2}{\sigma}\right)+x^{\zeta_2-1}e^{-\zeta_2\mu+\frac{\zeta_2^2\sigma^2}{2}}\left[1-\Phi\left(\frac{\ln x-\mu+\zeta_2\sigma^2}{\sigma}\right)\right]\right],$$

where ζ_1, ζ_2 are the Pareto exponents for the upper and the lower tails, respectively. One of the weights scales up the Pareto tail while the other scales it down. It has been estimated by Reed (2002) as well as by Giesen *et al.* in terms of the parameters $(\zeta_1, \zeta_2, \mu, \sigma^2)$. As Giesen *et al.*, p. 131, note, "[i]t is not possible to exactly delineate the lognormal body and and the Pareto distributed tails. That is, we cannot pin down parametrically at which city size the upper tail of the DPNL ...". The solid statistical foundations of the DPLN in a growth process model [Reed, *op. cit.*] is clearly an asset of the Giesen *et al.* approach. Reed does not provide microfoundations based on economic theory to support his growth based explanation. Therefore, his growth model, while attractive, does not necessarily confer a theoretical pedigree upon the DPLN. Thus, we respectfully view the DPLN as practical specification in empirical applications.

In contrast, it is we think a strength of our approach that we estimate the threshold τ in a transparent way. The switching point in our model is neither exogenous nor an arbitrary function of the Pareto exponent so the two are jointly estimated by maximum likelihood, whereas in the DPLN the transition to the Pareto law will occur at a population level that depends (in an intractable way) on the exponent. Notably, our approach eschews focusing on the lower tail, which does deviate from the lognormal but these deviations are less "economically significant" as they involve small cities (and a small fraction of the population), while the definition of such cities can be wildly different. In any case, the different data used, the statistics reported, and the results of our work should be viewed as complimentary to that of Giesen, Zimmermann and Suedekum (2010), yielding several orthogonal insights. Beyond direct estimates of the transition point, we most notably find that lognormality of the body may in fact be an artifact of measuring cities in terms of

administrative units, such as Places, rather than economic units (Metro & Micro Areas) or spatial units (Area Clusters).

4 Empirical Analysis

Recall Eeckhout (2004) uses data on US Places and finds that a lognormal fits the size distribution better than a Pareto. For those observations for which 1990 data also exist, Eeckhout claims that the growth rate of cities is independent of city size, making the growth process proportionate which asymptotically over time leads to a lognormal size distribution. Eeckhout (2009) interprets non-rejection of lognormality by a Lilliefors test as evidence that the size distribution of cities is best modeled as lognormal everywhere, notably including the upper tail.

However, the reason lognormality cannot be rejected by a Lilliefors test is that this test has very little power to detect deviations from a hypothesized distribution when these deviations occur in the tail. This is immediately apparent from the confidence intervals in Eeckhout (2009), Figure 2, in the tail. These confidence intervals are consistent with almost any imaginable distribution for cities larger than $\exp(12) \simeq 160,000$, including Pareto laws with a very broad range of exponent, e.g. 0.1 to 10. In contrast, several previous estimates of distributions of cities in this range have delivered fairly narrow confidence intervals[12] suggesting that the Lilliefors test has little power for distinguishing between tail behaviors. Indeed, it would be very puzzling if city sizes were actually lognormal and this was the narrowest possible confidence interval. If this were the case, we would expect significant statistical fluctuations of the behavior of distribution tails across different samples usually viewed as independent, such as samples from distant time periods and from different countries. However, they are at least somewhat clustered around power laws with exponents close to 1 (though often larger).

More fundamentally, a Lilliefors approach to testing the adequacy of the lognormal speci-

[12] See Gabaix and Ioannides (2004), p. 2345, where they discuss Krugman's estimate of 1.005, with a standard error of 0.010, obtained for the top 135 US metropolitan areas in 1990, and *ibid.,* 2351–2354, for a discussion of estimates obtained with similar precision by many others.

fication is not appropriate. These omnibus tests are designed to detect an arbitrary deviation from a hypothesized distribution rather than a specific deviation of interest such as the behavior of the tail. Since we are interested in the fit of the lognormal with the distribution of the size of large places we can consider this much more directly by estimating confidence intervals for the tail by drawing repeated samples from the lognormal equal in size to the observed sample (25358 observations) and examining the distribution function across simulations.

Indeed, we drew 1000 such samples, calculated the simulated distribution function for each sample and at each population value determined the $\alpha\%$ highest and lowest frequencies observed across distribution functions. These $\alpha\%$ confidence intervals are reported in Figures 1a and 1b for $\alpha = 5, 1$ and 0.1. Formally these are 'local' confidence intervals for the null that the empirical distribution is drawn from a lognormal when we consider the empirical distribution at a single *particular* population size. As such they can also be used to detect at which population values, if any, this null hypothesis can be rejected and therefore where the fit is good and where it is not. It is clear from Figure 1 that for any population size above 100,000 the empirical distribution function behaves differently to what we expect to occur with as little as 0.1% probability under a lognormal. Of course the same is true *a fortiori* with probabilities 1% or 5% in which case the confidence intervals are narrower.

The confidence intervals of Figure 1 are local in the sense that they describe the interval in which the observed population should be (with some probability under the null), at the specific location of the simulated cumulative distribution function. We also used the methodology of Duranton and Overman (2005) who in a different context construct global confidence intervals which determine the uniform interval around the model in which the entire range of the observed population should be. This approach applied without modification in a context like ours (where the *local* confidence intervals strictly increase with population) is equivalent to imposing the local confidence interval for the largest population size as a *global* confidence interval. This extremely conservative approach still leads to rejections at the 1% level, if only because as is evident from Figure 1b the very largest city (New York) is inconsistent with its local confidence interval which is also the global confidence interval

for all cities. A related type of global confidence interval would be to consider the size of the uniform confidence interval within which lie $a\%$ of all 25358*1000 cities across all 1000 simulations with 25,358 observations each across the entire size range. This number turns out very small, around ± 0.03 for $\alpha = 0.1\%$ (in the units of Figure 1) which implies a $\pm 7\%$ population band around the population levels predicted by the lognormal. We do not overlay this confidence interval on to Figure 1 as it would be visually indistinguishable from the lognormal model itself.

Our starting point is therefore that even in the Places data for which Eeckhout (2004, 2009) claims lognormality, the fit of the lognormal in the tail is highly questionable, which motivates our switching model. The model is analyzed in detail below across the Places and alternative data sets.

4.1 Estimation of switching model

Next we report max likelihood estimates for the parameters of several variants of our switching model $h(x;\mu,\sigma,\tau,\zeta)$, defined in section 3 above. The respective truncated models are used, presented in sections 3.2 and 3.4 respectively, with the Metro & Micro areas data which are truncated by design. The maximum likelihood estimator is quite simply:

$$\max_{\mu,\sigma,\tau,\zeta} : \sum_{i=1}^{N} \ell n\left[h(x;\mu,\sigma,\tau,\zeta)\right], \tag{12}$$

given data $\{x_i\}_{i=1}^N$. When working with the Metro & Micro data, which are by construction left censored at populations of 12,400, we use the truncated version of our baseline model as described in Section 3.2. In all estimates we constrain $\tau > \exp(\mu)$ in order to ensure the density is unimodal. The contrary is empirically implausible, may lead to numerical issues with multiple optima in the likelihood function, and is in conflict with the interpretation that we are seeking to fit the *tail* of the distribution.

Turning to our parameter estimates[13] in Table 1, we report that using our baseline

[13]Maximization was performed using Matlab's fmincon function from various starting values, which performs optimization based on numerical gradients. Starting from any 'reasonable' values such as the max

switching model, the switching to a Pareto tail occurs at populations τ of 60,290 in the Places data. The switching is at somewhat lower populations in the Metro & Micro Areas (34,853) and Area Clusters data (30,635). Interestingly, this means that 2% of Places (501 in number) are in the tail (>60,290) but contain 46% of the Places population. Similarly, 2% of the Area Clusters (632 in number) are in what we estimate as the tail (>30,635) and contain 40% of the Area Clusters population. Because Metro & Micro Areas are left censored, the corresponding numbers are harder to interpret but for the sake of completeness, we report that 77% (746 in number) of all Metro & Micro Areas are in the Pareto tail (>34,853) and contain 98% of the Metro & Micro Area population. In comparison, the threshold estimate implied by Malevergne *et al.* (2011) is 37235.

Turning now to the mixture model presented in section 3.3 above, we note that an additional parameter θ is involved and estimated. From (8), this parameter defines the mixture of Pareto and lognormal describing the upper tail. The larger is θ, the more important is the Pareto in the mixture. The estimates of θ for our three data sets are 0.25, 1.00, 0.88, respectively. It is thus not surprising that the Pareto is least important for the Places data and most important for the Micro & Metro areas, which are by construction data on the upper tail. The estimate of the cut-off point τ is not affected for the Area Clusters and essentially for the Micro & Metro areas, as well, but is very significantly affected for the Places data, for which turns out to be much smaller.

The estimates of the Pareto exponent, ζ, are 1.25, 0.92, and 0.74, for the respective data sets and the baseline model, and 0.82, 0.92, and 0.86, for the CDGPR-switching model. They are all estimated with very high precision and therefore in most cases are statistically different from Zipf's law. Many researchers, and most forcefully Simon (1968), have emphasized that given the extremely sharp prediction of Zipf's law it should be tested leniently, with evidence of an approximate Pareto with a parameter not too far from 1 being supportive even if the

likelihood parameters of a lognormal model for μ and σ, Zipf for ζ and τ around 500,000 we always got the values reported to several digits of accuracy. Starting far from these parameters occasionally produced other optima which were however always inferior. We also tried optimization using pattern search, simulated annealing and genetic algorithms which produced the same or inferior values depending on starting values and optimization algorithm parameter settings. We are therefore confident that our estimates represent global maxima to the level of accuracy reported in the tables.

law can be rejected in the strict statistical sense. After all any null hypothesis can be rejected with enough data and Zipf's is a very sharp hypothesis. Note also that the size of the exponent ζ is larger in the Places data where the switching population is also larger, implying perhaps that when a large range of populations is described by a lognormal body, the tail is also thicker. In comparison, the estimate by Giesen *et al.* (2010) of the Pareto exponent ζ_1 that corresponds to ours varies from a maximum of 1.442 for Germany to a minimum of 1.028 for France; the one for US Places is 1.225.

It is worth noting that in almost all cases, the uncensored lognormal parameter estimates of Eeckhout (2004), $\hat{\mu} = 7.28$, $\hat{\sigma} = 1.75$, reproduced on Column 1, Table 1 below, are not very different from the corresponding μ, σ parameter estimates in our switching models. Nevertheless, the estimate of τ clearly suggests a Pareto tail missed by Eeckhout's model. If the Pareto behavior were not empirically relevant, τ would be very large (or be associated with a very large standard error) in comparison to the range of the data, which would have indicated that all observed cities are well described by a lognormal. The fact that this is not the case provides evidence further refined below by means of a battery of tests, according to which the switching model is preferable to a pure lognormal. In other words, our estimates confirm previous research as far as the body of the distribution is concerned, but nevertheless lead to very different conclusions regarding the tail.

Note also the very small standard errors of most estimated parameters. This suggests that our proposed model is not unduly flexible for the data analyzed, since all parameters, including the switching population parameter τ are all estimated with good precision. This is one important way to evaluate whether a model is overfit, but we also confirm this conclusion with standard model complexity criteria below.

4.2 Qualitative specification analysis

We have already discussed the evidence in Figure 1 according to which the lognormal model is a poor fit for the upper tail of the US Census places data. By contrast, Figure 2 illustrates the good fit of all variants of our switching model to the Places data. The $q - q$ plots in

Figure 3 further reinforce the conclusion that our model fits the data reasonably well, with deviations being small in terms of economic importance. The mixture model seems to add little to justify the more complex interpretation and computation it requires.

Figure 2 also makes it evident that there are deviations from lognormality in the data which are extremely economically significant. For the largest cities, the lognormal density is *one fifth to one hundredth* the magnitude of the actual observed empirical distribution! In other words, it grossly understates the frequency of very large cities. For example, according to the fitted lognormal, there is an approximately one-in-25,000 chance of observing a city larger than one million yet we observe around ten cities larger than that, with New York eight times larger in a sample of just 25,358 observations. In contrast, the Pareto law describes the size of large cities very well. Large cities are an extremely important part in the US urban structure, are home to a large fraction of the population and play a dominant role in the US economy so this deviation is very important. Furthermore, our benchmark lognormal-Pareto model has good predictions for the frequency of all city sizes, large and small, without errors of much economic significance. Similar conclusions hold if we constrain our model to fit Zipf's law exactly, i.e. impose $\zeta = 1$ in which case it has only one additional parameter (the switching population) relative to the simple lognormal model. Similarly, the CDGPR variant of our model has a much better fit than the lognormal; its fit is visually comparable to our other switching model variants despite the fact that it nests them.

Figure 4 makes it apparent that while the top 100 cities have a very good fit to a Pareto across all data sets, the shape of their body varies considerably across definitions. Indeed, Figure 5 shows that the good fit of the lognormal for the body of the size distribution is specific to the Places data and quite poor for Area Clusters and Metro & Micro definitions. The robust feature across all data definitions is, instead, that there is a tail Pareto behavior that starts at a population level somewhere below a population level of 87,000 (whose $\log_{10} 87,000 \approx 4.93$). Lognormality of the body thus seems to be a property of the definition; it is there for the Places data, but is sensitive to the city definition used.

We also note that separate analysis reveals that at very small sizes of less than a few hundred inhabitants, the Places and Areas data have somewhat anomalous and unexpected

behaviour: the empirical distribution of the Places data has a clear second mode at around 150 inhabitants while that of the Area Clusters data is distinctly U-shaped in the range of 1-1000 inhabitants. This reinforces our evidence that the behaviour of cities that are not "large' is sensitive to the definition used. Cities as empirical and theoretical entities may only be meaningful above a certain size and if this view is adopted, this size can be easily estimated using a switching model.

4.3 Formal specification tests

One might object that the qualitative superiority of our model over the lognormal is of limited interest because a nested model always perform worse than the nesting model. Below we show that the switching model is superior to the lognormal in terms of formal statistical specification tests that penalize our models for their greater generality. We also compare the models according to economic metrics. These are important in order to gauge the extent to which each model captures or misses economically meaningful aspects of the data.

Statistical metrics: The poor fit of the lognormal is easily confirmed with an Anderson normality test on the log of sizes (see Table 1, top). The same test easily rejects lognormality for the Area Clusters data. We cannot apply this test to the Metro & Micro Areas data because they are left censored by construction.

We have already shown that our model parameter estimates in Table 1 have small standard errors based on which the lognormal special case (where τ tends to infinity) is extremely unlikely. In addition, the confidence intervals of the parameters are such that a null hypothesis that the additional parameters beyond those of a pure lognormal are statistically significant is not rejected. The fact that the confidence intervals for the parameters are clearly very far from the limiting lognormal case is strong evidence that the deviations from it are very statistically significant.

Table 1 also reports AIC, BIC and Bayes' factors against the lognormal, all of which suggest that all three variants of our model are preferred to the lognormal for all three datasets. Note that likelihood ratio tests and some other standard statistical specification

tests cannot be easily applied here because the lognormal special case is nested but only as a limiting case [Andrews (2001)]. Comparing the three variants of the switching model to each other gives a somewhat mixed picture depending on the data used; in any case, the differences in performance are small. The most striking result is that the mixture model for the Area Clusters data gives identical estimates to the lognormal-Pareto model which is nested in it.

We note that all these statistics have a serious drawback which is that the fit at small cities is weighted equally to the fit at large cities. But the fit at small cities is a very marginal issue from the point of view of the entire population since it is only relevant for a small fraction. For example, Places smaller than population 500 account for less than 1% of the population across all Places; but there are over 7265 such observations and the model fit at those observations receives equal weight to its fit at the 7265 largest observations which account for over 91% of the population. For this reason, economic metrics are essential when evaluating city size distributions over large ranges.

Economic metrics: The specification statistics we discuss next are designed to have an economic interpretation. One is the model population error, i.e. the number of people that would need to be shuffled across cities to get the distribution to fit exactly the estimated model; another is the pseudo-R^2, defined by Duranton (2007) as the variance of the model's population errors divided by the variance of population across cities. Both of these are reported on Table 1. By both criteria all variants of our model dominate the lognormal with one exception (population error for Metro & Micro areas data but this criterion can be very statistically noisy). Among our three variants, the lognormal-Pareto is preferred for the Places data though the mixture model provides an excellent fit to the Metro & Micro Areas.

It is also appropriate to note that our model has statistically significant *deviations* from the data. This is confirmed by the Cramer-von-Mises test, reported in Table 1, which easily rejects all model specifications for all three data sets. With large data sets like the ones we employ here, even more refined models can be proposed. With data sets of this size, parsimony is a secondary issue from an empirical modelling perspective since there is little

estimation uncertainty relative to model uncertainty.

However, it is not clear to us that further statistical accuracy would be very *economically* meaningful: our model provides a reasonable fit for large cities and these are economically the most important cities. We cannot think of an economic context where more accurate estimation of the distribution of large cities would be necessary. Regarding smaller cities, our data analysis suggests that definitional and measurement problems may plague any effort to develop a detailed understanding of their size distribution. If a more accurate description of the body is nevertheless required, it is straightforward to accommodate it using our switching approach that differentiates between body and tail by substituting the lognormal with another density.

In sum, the proposed model provides a statistically and economically more attractive description of the data than the lognormal alone, even when parsimony is accounted for. It also allows us to easily estimate and report key parameters of interest, such as tail exponents and switching points that define the tail, even when lognormality is an imperfect description of the body of the size distribution as is the case with the Area Clusters and Metro & Micro Areas.

5 Implications for the Economic Theory of Urban Structure

Several researchers, notably including Gabaix (1999) have proposed simple economic models for the city size distribution that predict a Pareto distribution. Our empirical estimates reported above indicate that while these models can explain the size behavior of all metropolitan areas and a significant number of large micropolitan areas, they can explain only a tiny fraction of cities (around 2%) and less than half the population living in all cities of the Places and Area Clusters data. On the other hand, economic models that predict lognormality (such as Eeckhout, 2004) can only explain the body of the Places data (though not even the body of the other city definitions).

In hindsight, this failure of extant models to explain the entire city size distribution is not surprising. These models do not allow for structural differences between larger and smaller cities and so they cannot accommodate changing qualitative behavior of the size distribution across different ranges. But the system of cities literature has argued theoretically and empirically [Black and Henderson (2003); Duranton (2007); Henderson (1974; 1997); Rossi-Hansberg and Wright (2007)] that cities specialize. Ioannides (2012), Ch. 9, shows that even though under special assumptions the dynamics of growth of specialized and diversified cities coincide, in general they would not. Growth of specialized cities is subject to economy wide shocks that are transmitted to each city through the terms of trade. Diversified cities, on the other hand, host many types of industries and their dynamics combine characteristics of many industries. Therefore specialized cities' own growth processes may be subject to different forces once their characteristics become set. Indeed, it is natural that size might be correlated with economic characteristics and as Henderson (1997) argues it is medium size cities that specialize. Similarly, these arguments may also be cast in terms of functional rather than sectoral specialization [Duranton and Puga (2005)]. Larger cities host a diverse set of industries, a notion bolstered by the urban hierarchy literature as well. That is, according to the urban hierarchy principle, industries found in a city of a given size will also be found in cities of larger size [Hsu (2009); Mori and Smith (2009)]. Much of the theory of systems of cities has implicitly assumed agglomerations of sizable populations, and simply does not make sense for small populations, like of size, 1, 2 or even one thousand. Such considerations also provide a strong theoretical rationale for allowing the data to determine whether different statistical laws rule in different parts of the city size distribution. This is particularly important, since e.g. models predicting lognormality also imply very economically important deviations from reality - in particular far fewer large cities than we actually observe.

Recall our discussion of how introduction of heterogeneity in each city's growth within a country whose size is non-stochastic implies that each city has its own steady-state size distribution. As a result, the tail of the overall cross-sectional distribution is dominated by those cities whose dynamics lead them to distributions that have the fattest tails among all

those in the cross-section. But each steady-state cannot be heavier-tailed than Zipf's law (which is the fattest-tailed distribution consistent with a finite mean), otherwise the size of the entire country would be stochastic. Therefore the fattest-tailed cities must obey Zipf's law, and since these dominate the cross-section, the cross-section has a Zipf distribution as well.

This explanation is consistent with a broad range of behaviors for the body of the size distribution, which of course depends on the respective definition of city, which is in turn associated with different growth dynamics. Such heterogeneity may mask changes in their economic roles over time. The details of these dynamics determine the body of the distribution while the Zipf tail appears across *all* definitions. This perspective allows a single approach to explaining the distributions of city sizes under all the various definitions of cities and plausibly suggests the growth processes of moderately sized cities might be more sensitive to the city definition used than the processes of large cities.

6 Conclusions

Our results confirm rigorously and in a statistically robust manner that the upper tail of the US city size distribution fits a Pareto distribution. This result is robust across city definitions, whereas Eeckhout's (2004) claim of city size lognormality in fact only applies to the body of the size distribution of only one somewhat arbitrary definition of city. Nevertheless, Eeckhout's (2004) pioneering step to use the US Census Places data that extend across the full size range to estimate the size distribution should be credited for leading to the remarkable finding that the tail of the US city size distribution obeys a *different* law than its body. Many international data sets use different lower cut-off populations, typically set arbitrarily. Such data sets do not allow easy detection of changing laws describing different parts of the distribution. Naturally, each city definition is associated with different growth dynamics. The details of these dynamics determine the body of the distribution. The persistence of the Pareto tail across *all* definitions suggests that the upper tail is shaped by different forces. This perspective would allow a single approach to explaining all the various

distributions of city sizes under alternative definitions of cities. It plausibly suggests the growth processes of moderately sized cities might be more sensitive to the city definition used than the processes of large cities.

The switching model we introduce may be useful in modeling many variables beyond population such as wealth, income, internet traffic, ecology, and other instances where there have been suggestions of similar behavior. This heterogeneity of the behavior of the size distribution is not surprising from a theoretical perspective. Size is likely to play a role in determining cities' economic properties, growth, degree of specialization etc. Yet, this is something the city size distribution literature has not yet dealt with: Eeckhout's theoretical model is a poor fit in the range of the largest and most important cities traditionally of concern to urban economics. It is also a poor fit for data based on definitions of US cities other than Places. On the other hand, theoretical models predicting Pareto laws for all cities [Gabaix (1999)] get it right for the top 500 or so most important cities but wrong for the overwhelming majority, that is 98%, of smaller cities. The estimated Pareto exponent is close but not equal to 1, whereas the implied tail behavior is dramatically different from that of the lognormal. To return to Krugman's (1996) notable adjective, the Pareto law of city sizes and its exponent remain *spooky*!

7 References

Ades, Alberto F., and Edward L. Glaeser. 1995. "Trade and Circuses: Explaining Urban Giants." *Quarterly Journal of Economics* 110:195–227.

Andrews, Donald W.K. 2001. "Testing when a Parameter is on the Boundary of the Maintained Hypothesis", *Econometrica* 69(3): 683–734.

Arnold, B.C. and S.J. Press, 1989, "Bayesian Estimation and Prediction for Pareto Data", *Journal of the American Statistical Association*, 408: 1079-1084.

Black, Duncan, and J. Vernon Henderson. 2003. "Urban Evolution in the USA." *Journal of Economic Geography* 3:343–372.

Bogue, Donald, 1953. *Population Growth in Standard Metropolitan Areas 1900 – 1950.* Scripps Foundation in Research in Population Problems, Oxford, Ohio.

Caers, Hef and Jozef Van Dyck, 1999, "Nonparametric Tail Estimation using a Double Bootstrap Method," *Computational Statistics and Data Analysis,* 29:191-211.

Combes, Pierre-Philippe, Gilles Duranton, Laurent Gobillon, Diego Puga, and Sébastien Roux. 2012. "The Productivity Advantages of Large Cities: Distinguishing Agglomeration from Firm Selection." working paper, February; *Econometrica*, forthcoming.

Clauset, Aaron, Cosma R. Shalizi, and Mark E.J. Newman. "Power-law Distributions in Empirical Data." *SIAM Review* 51(4):661-703.

Dobkins, Linda H., Ioannides, Yannis .M. 2000. "Dynamic Evolution of the U.S. City Size Distribution." In: Huriot, J.-M., Thisse, J.-F. (Eds.), *The Economics of Cities.* Cambridge University Press, Cambridge, pp. 217–260.

Dobkins, Linda H., and Yannis M. Ioannides. 2001. "Spatial interactions among U.S. cities: 1900–1990." *Regional Science and Urban Economics* 31:701–731.

Duranton, Gilles. 2007. "Urban Evolutions: The Fast, the Slow, and the Still." *American Economic Review,* 97(1):197—221.

Duranton, Gilles, and Diego Puga. 2005. "From Sectoral to Functional Urban Specialization." *Journal of Urban Economics* 57(2):343–370.

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

Eeckhout, Jan. 2009. "Gibrat's Law for (All) Cities: A Reply." *American Economic Review*, 99(4)): 1676–1683.

Gabaix, Xavier. 1999. "Zipf's Law for Cities: An Explanation." *Quarterly Journal of Economics* 114(3):739–67.

Giesen, Kristian, Arndt Zimmermann and Jens Suedekum. 2010. 'The Size Distribution Across All Cities - Double Pareto Lognormal Strikes." *Journal of Urban Economics* 68(2):129–137.

Handcock, M.S. and J.H. Jones, 2004, *Theoretical Population Biology* 65:413.

Henderson, J. Vernon. 1974. "The Size and Types of Cities." *American Economic Review*, 64(4):640–656.

Henderson, J. Vernon. 1997. "Medium Size Cities." *Regional Science and Urban Economics* 27:583-612.

Hsu, Wen-Tai. 2009. "Central Place Theory and City Size Distribution." working paper, Department of Economics, Chinese University of Hong Kong. August.

Ioannides, Yannis M. 2012. *From Neighborhoods to Nations: The Economics of Social Interactions.* Princeton University Press, (forthcoming).

Ioannides, Yannis M., Henry G. Overman, Esteban Rossi-Hansberg, and Kurt Schmidheiny. 2008. "The Effect of Information and Communication Technologies on Urban Structure." *Economic Policy* 203–242.

Krugman, Paul. 1996. "Confronting the Myth of Urban Hierarchy." *Journal of the Japanese and the International Economies* 10: 399-418.

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

Maddala, G.S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press.

Malevergne, Y., V. Pisarenko, and D. Sornette. 2011. Y.Malevergne, V. Pisarenko and D. Sornette. 2011: "Testing the Pareto against the Lognormal Distributions with the Uniformly most Powerful Unbiased Test Applied to the Distribution of Cities", *Physical Review E* 83, 036111.

Mori, Tomoya, and Tony E. Smith. 2009. 'On the New Empirical Regularities among Industrial Location Behavior, Industrial Diversities and Population Sizes of Cities in Japan." *Brookings Wharton Papers on Urban Affairs,* 175–216.

Reed, William J. 2002. 'On the Rank-Size Distribution for Human Settlements." *Journal of Regional Science* 42(1):1–17.

Reed, William J., and Murray Jorgensen. 2005. 'The Double Pareto–Lognormal — A New Parametric Model for Size Distributions." *Communications in Statistics – Theory and Methods* 33(8):1733–1753.

Rozenfeld, Hernán D., Diego Rybski, Xavier Gabaix, Hernán A. Makse. 2011. 'The Area and Population of Cities: New Insights from a Different Perspective on Cities." *American Economic Review* 101(5):2205–2225.

Simon, Herbert. 1968. "On Judging the Plausibility of Theories." In *Logic, Methodology and Philosophy of Sciences, Vol III.* ed. B. van Rootsellar and J. F. Staal, 439–459. Amsterdam: North-Holland.

Skouras, Spyros. 2010. "Explaining Zipf's law for US Cities." Working paper,

`http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1527497`

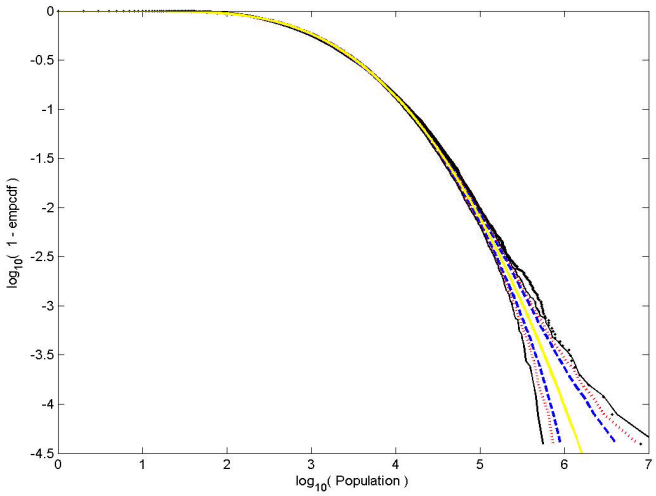

Figure 1a: The countercumulative of the empirical distribution function (circles) is plotted against the countercumulative of Eeckhout's fitted lognormal (central solid line) with confidence intervals for each population level. The confidence intervals are at the 5% (dashed), 1% (dotted) and 0.1% (solid) levels. The countercumulative of the empirical distribution function can also be interpreted as each city's size rank from large to small.

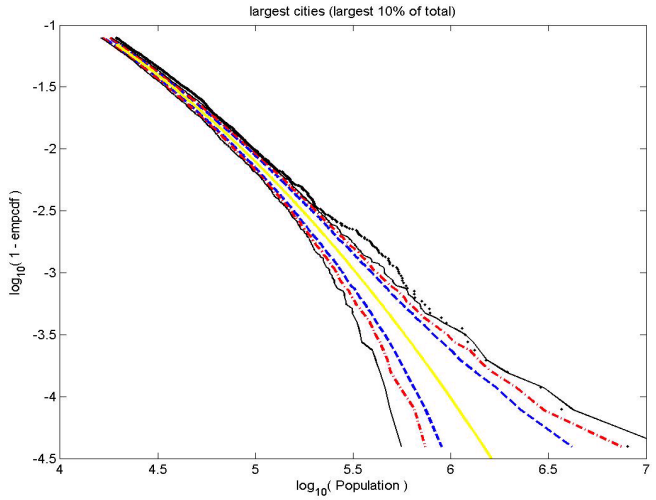

Figure 1b: A magnification into the tail of the countercumulative distributions depicted in Figure 1a. The lognormal provides a poor fit for the 2000 largest cities depicted here (population larger than roughly 20,000).

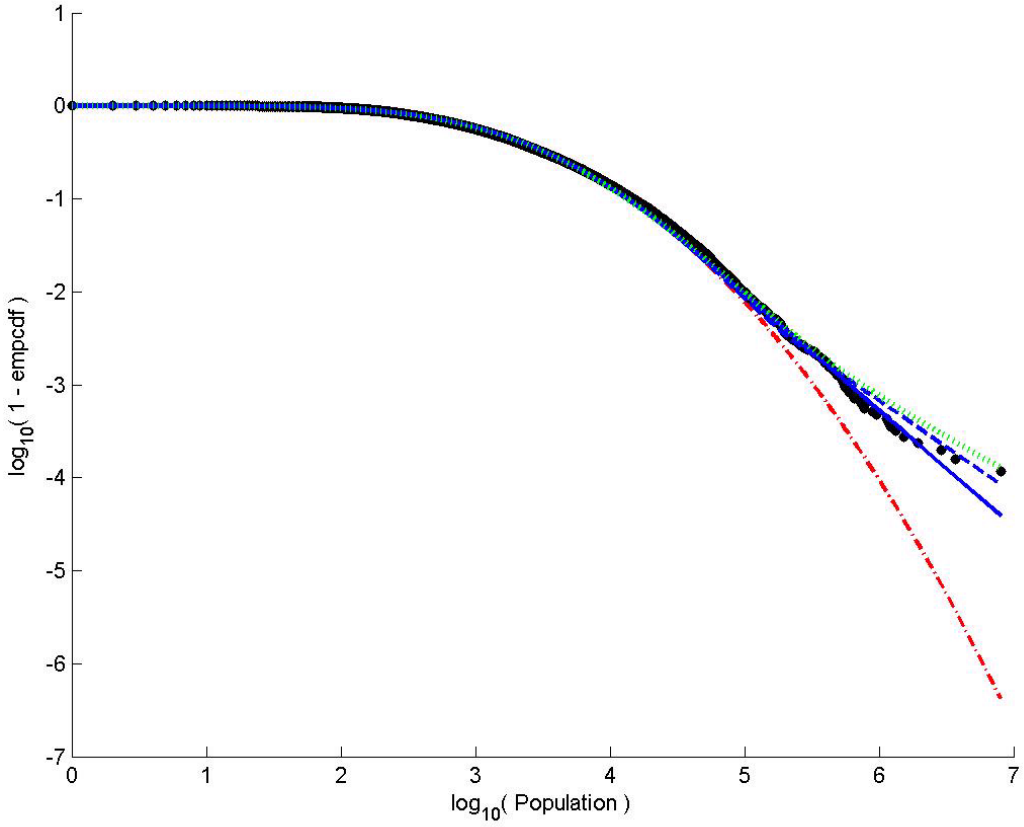

Figure 2: The countercumulative of the empirical distribution function (circles) is plotted against the countercumulative of Eeckhout's fitted lognormal (dash-dot line); the estimated proposed lognormal-power distribution (solid); the same constrained to satisfy Zipf's law (dashed); and the CDGPR mixture model. Clearly all variants of the proposed distribution provide a much better fit even when constrained to fit Zipf's law. The lognormal misses the empirical distribution by *two orders of magnitude* in the tail.

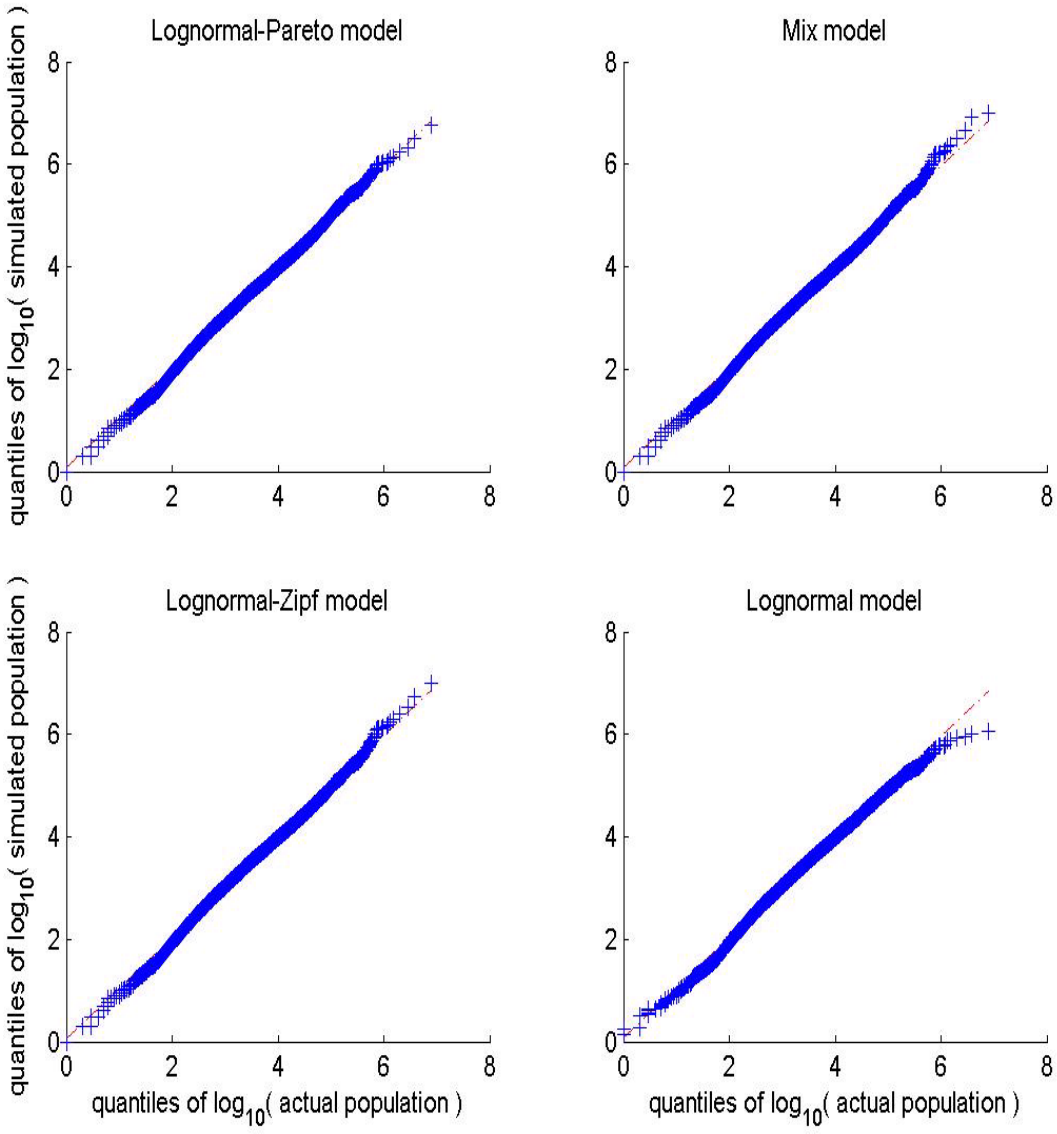

Figure 3: These qqplots for the Places data confirm the decent fit of our switching models. Systematic deviations from lognormality are apparent in the bottom plot. Each plot depicts the deviation of the actual Places populations from a sample simulated from the corresponding model with parameters equal to our max likelihood estimates.

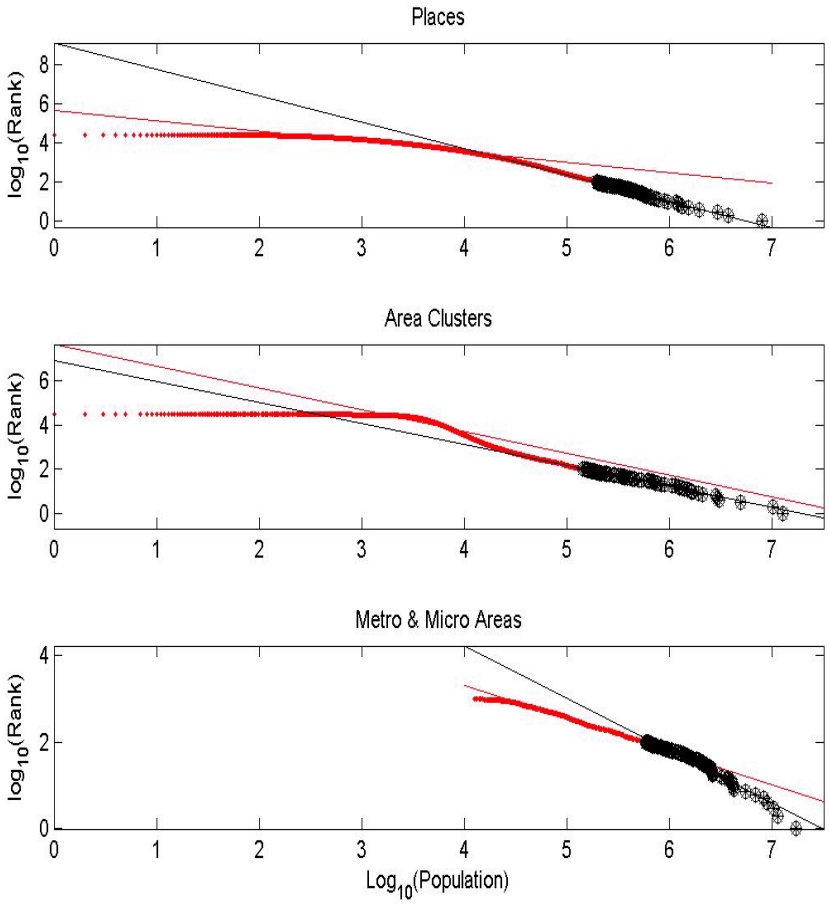

Figure 4: We present rank-size plots by each city definition. We distinguish the top 100 cities (heavy black circles) from the entire distribution (red dots). The least squares line on the tail cities is presented separately to the line for all cities. Evidently, the Pareto law is a good fit across all definitions but only for the tail cities. The shape of the distribution of smaller cities differs drastically with definition.

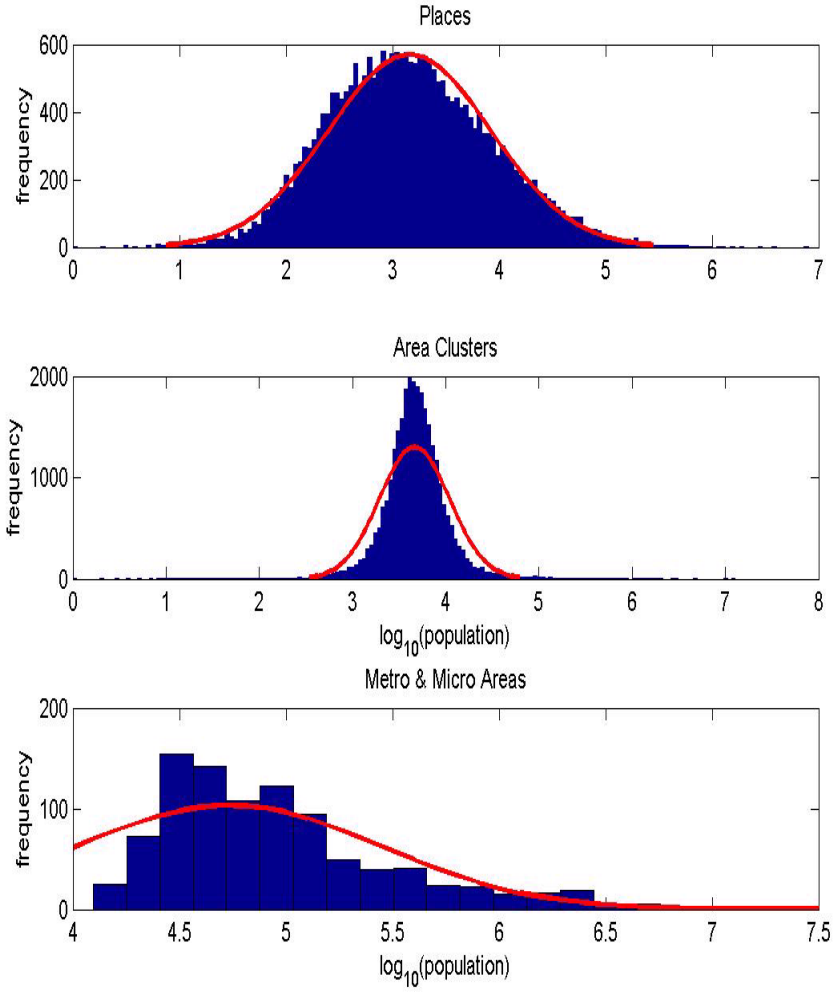

Figure 5: We present histograms for city populations according to each city definition. Consistent also with Figure 4, the shape of the distribution differs drastically with definition. Furthermore, the lognormal is only a good fit for the Places data.

		Places				Area Clusters				Metro & Micro Areas			
# observations		25358				23499				964			
min - max		1 - 8008278				1 - 15594627				12463 - 16846046			
Anderson normality		4.79				22.55				-			
	p-val	1				1				-			
Model		LogN	LogN-Zipf	LogN-Pareto	mixture	LogN	LogN-Zipf	LogN-Pareto	mixture	LogN	LogN-Zipf	LogN-Pareto	mixture
Parameter estimates										nb: Truncated model used			
μ		7.28	7.27	7.26	7.25	8.43	8.37	8.37	identical	10.90	10.96	12.13	13.26
	st. error	(0.01)	(0.01)	(0.01)	(0.01)	(0.005)	(0.005)	(0.005)	estimates	(0.1)	(0.09)	(2.60)	(0.29)
σ		1.75	1.74	1.73	1.72	0.91	0.79	0.79	to logN-	1.64	0.78	0.92	1.15
	st. error	(0.01)	(0.01)	(0.01)	(0.01)	(0.004)	(0.004)	(0.004)	Pareto	(0.07)	(0.11)	(0.59)	(0.05)
τ		-	268169	60290	16312	-	29564	30635		-	105718	34853	34998
	st. error	-	(33868)	(18480)	(5401)	-	(628)	(644)		-	(35105)	(222)	(212)
ζ		-	1*	1.25	0.82	-		0.92		-	1*	0.74	0.86
	st. error	-	-	(0.05)	(0.08)	-		(0.03)		-	-	(0.03)	(0.06)
θ		-	-	-	0.25	-	-	-		-	-	-	0.88
	st. error	-	-	-	(0.09)	-	-	-		-	-	-	(0.04)
Weight of Pareto At Pop 500K			-	-	2:1	-	-	-			-	-	2:1
Log-Likelihood		-234773	-234761	-234756	-234750	-229171	-227736	-227733		-12534	-12494	-12472	12467
Specification fit and comparison													
pseudo-R2		0.997	**0.999**	**0.999**	**0.999**	0.09	**0.98**	**0.98**		0.833	0.961	0.994	**0.998**
Population error (mn)		44.9	25	**21.1**	30.9	148.9	65.7	**35.4**		147.2	107	172.6	**26.7**
AIC		469550	469528	469519	**469510**	458347	455477	**455474**		25072	24994	**24952**	**24952**
BIC		469566	469552	469552	469551	458363	**455502**	455507		28082	25008	**24971**	24975
Bayes' factor		-	0.001	0.001	0	-	0	0		-	0	0	0
Cramer-von Mises		1.89	1.75	1.64	1.61	41.84	23.97	24		3.5	2.53	0.22	0.24
	p-val	1	1	1	1	1	1	1		1	1	0.65	0.7

Table 1: Anderson normality statistic is for log size (not calculated for Metro & Micro data due to censoring). The Pareto law parameter ζ is fixed at one for all lognormal-Zipf models (and marked by a star), while all estimated parameters are estimated by max likelihood. All asymptotic standard errors are derived from numerical hessians. The Pseudo-R2 is calculated as in Duranton (2007) as 1-(MSE of the model cdf)/var(population) in log-log space (as in Figure 2). The Bayes' factors is computed relative to the lognormal as exp(0.5(BIC-BIC of lognormal)) and is decisive against the lognormal in all cases according to Jeffreys' scale. The Cramer stat is the average across 1000 samples simulated from each model. Preferred models according to a particular statistic are indicated in bold.