A new framework for US city size distribution: Empirical evidence and theory

Ramos, Arturo and Sanz-Gracia, Fernando and González-Val, Rafael

Departamento de Análisis Económico. Universidad de Zaragoza, Institut d'Economia de Barcelona (IEB), Universitat de Barcelona

13 December 2013

A new framework for the US city size distribution: Empirical evidence and theory

ARTURO RAMOS* FERNANDO SANZ-GRACIA†

RAFAEL GONZÁLEZ-VAL‡

December 13, 2013

Abstract

We study the US city size distribution using the Census places data, without size restriction, for the period (1900-2010). Also, we use the recently introduced US City Clustering Algorithm (CCA) data for 1991 and 2000.

We compare the lognormal, two distributions named after Ioannides and Skouras (2013) and the double Pareto lognormal with two newly introduced distributions. The empirical results are overwhelming: One of the new distributions widely outperform any of the previously used density functions for each type of data.

We also develop a theory which generates the new distributions based on the standard geometric Brownian motion for the population in the short term. We propose some extensions of the theory in order to deal with the long term empirical features.

JEL:C13, C16, R00.

Keywords: US city size distribution, population thresholds, lower and upper tail, new statistical distributions

*Department of Economic Analysis, Universidad de Zaragoza (SPAIN) aramos@unizar.es

†Department of Economic Analysis, Universidad de Zaragoza (SPAIN) fsanz@unizar.es

‡Department of Economic Analysis, Universidad de Zaragoza and Institut d'Economia de Barcelona (IEB), Universitat de Barcelona (SPAIN) rafaelg@unizar.es

1 Introduction

The study of city size distribution has a long tradition in urban economics. To cite just a few examples, see Black and Henderson (2003), Ioannides and Overman (2003), Soo (2005), Anderson and Ge (2005), Bosker et al. (2008) and the more recent ones of Giesen et al. (2010) and Ioannides and Skouras (2013).

Along the years, the Pareto distribution (Pareto, 1896) (for the upper tail, subindex "ut") has generated a huge amount of research and great acceptance. The normalized density function for such a distribution reads

$$f_{\rm ut}(x, x_m, \zeta) = \frac{\zeta}{x} \left(\frac{x_m}{x}\right)^{\zeta}, \quad x > x_m \,,$$

where $x > x_m$ is the population of the urban centers, x_m is the minimum threshold size and $\zeta > 0$ is the *Pareto exponent*. [1]

In an influential paper regarding city size distribution, Eeckhout (2004) essentially proposes the lognormal to describe it, using US Census data for the year 2000 of all unincorporated and incorporated places in his analysis. Lognormal distributions had been previously proposed by Parr and Suzuki (1973), but one of the main points in the first reference of this paragraph is that one should take into account the whole set of cities when studying their distribution. Later, a polemic arises by Levy (2009), who argued that the upper tail of the city size distribution, and thus most of the population (for the US places), rather followed a Pareto distribution instead of a lognormal.

On this line of research, the important contribution of Ioannides and Skouras (2013) has appeared, aiming to reconcile both views by means of the proposal of two distributions (IS1 and IS2 hereafter) which have a lognormal body and above an explicit threshold, a Pareto power law (IS1) or a linear combination of Pareto and lognormal (IS2) in the upper tail.

In parallel to the appearance of these works, it has been proposed a distribution which has a lognormal body and power laws in the tails, but without clearly delineating between the three behaviors, called the double Pareto lognormal (dPln); see, e.g., Reed (2002, 2003), Reed and Jorgensen (2004). The fit of such distribution is remarkably good for a number of countries (see Giesen et al. (2010), for eight countries and the recent contribution González-Val et al. (2013b) for a more comprehensive data set).

Let us try to motivate in what follows the appropriateness of our approach (see Section 3 for details). In this context, and summarizing several of the previous con-

[1] The cumulative distribution function is

$$\mathrm{cdf}_{\rm ut}(x, x_m, \zeta) = 1 - \left(\frac{x_m}{x}\right)^{\zeta}, \quad x > x_m$$

so that

$$1 - \mathrm{cdf}_{\rm ut}(x, x_m, \zeta) = \left(\frac{x_m}{x}\right)^{\zeta}$$

and

$$\ln(1 - \mathrm{cdf}_{\rm ut}(x, x_m, \zeta)) = \zeta \ln x_m - \zeta \ln x$$

Thus, for a Pareto distribution, the quantity $\ln(1-\mathrm{cdf})$ is linear in $\ln x$ with negative slope of absolute value ζ. The case of $\zeta = 1$ corresponds to the well-known *Zipf's law* (Zipf, 1949); see the surveys on this subject by Cheshire (1999), and Gabaix and Ioannides (2004). This is the foundation for the well-known *Zipf plots*.

tributions, there is nowadays a certain consensus regarding the study of the city size distribution in the sense that a combination of Pareto and lognormal provides the best fit, IS1 and IS2 having a component of Pareto only in the upper tail and dPln having components of Pareto in the upper and lower tails. We build on this relevant strand of the literature and go further in two ways. First, proposing two new distributions that systematically outperform the lognormal, dPln, IS1 and IS2. Second, offering a theoretical basis for the newly introduced distributions based on the standard geometric Brownian motion process for population and the associated forward Kolmogorov or Fokker–Planck differential equation (Gabaix, 2009, 1999).

For the lower tail (subindex "lt") of city size distributions it has been observed by Reed (2001, 2002, 2003) that they indeed follow a power law, plotting the natural logarithm of cumulative frequencies against that of population.[2] Such a fact seems to be overlooked in the literature, and as we will see below, is one of the important points one should take into account in order to obtain an excellent overall fit.

Against this background, we have decided to compare in detail the distributions IS1 and IS2 proposed by Ioannides and Skouras (2013) with the dPln, and, in order to reconcile both tendencies, we propose two new distributions which take the essence of both views and go a long way ahead. They are:

- The "threshold double Pareto Singh–Maddala" (tdPSM), which is a distribution with a Singh–Maddala one (Singh and Maddala, 1976) in the body and with both tails that follow a power law, but with two thresholds which exactly delineate the switch between the different behaviors. It is like the IS1 of Ioannides and Skouras (2013) but with the lower tail modeled as a pure power law and the body being Singh–Maddala instead of lognormal. As far as we know, the tdPSM is a completely new distribution.

- The "double mixture Pareto Champernowne Pareto" (dm PChP), which is a distribution with a Champernowne distribution (Champernowne, 1952) body and with a linear combination of Champernowne and Pareto in both tails, also with two population thresholds which exactly delineate the switch between the different behaviors. It is like the IS2 of Ioannides and Skouras (2013) but with the lower tail modeled as a mixture of Champernowne and power law, and the lognormal substituted by a Champernowne in general. This is, to the best of our knowledge, also a new distribution.[3]

[2] For the lower tail, we can define the Pareto density function

$$f_{\mathrm{lt}}(x, x_M, \rho) = \frac{\rho}{x}\left(\frac{x}{x_M}\right)^{\rho}, \quad 0 < x < x_M\,,$$

where now x_M is the maximum size threshold and $\rho > 1$ is the Pareto exponent. The cumulative distribution function is then

$$\mathrm{cdf}_{\mathrm{lt}}(x, x_M, \rho) = \left(\frac{x}{x_M}\right)^{\rho}, \quad 0 < x < x_M\,,$$

and therefore $\ln(\mathrm{cdf}_{\mathrm{lt}}(x, x_M, \rho)) = \rho \ln x - \rho \ln x_M$. So, we have that for a lower tail Pareto distribution, the natural logarithm of cdf gives a straight line in $\ln x$ with positive slope ρ. We will plot the previous quantities in the left panels of Figures 1 and 2.

[3] The arrival to the previous two distributions is the outcome of a research process in which we have tried

These distributions yield extremely good, strong and encouraging results, and they rely on the following important improvements:

- The extremely important need to specifically model the lower tail as a power law in order to get an overall good fit, as mentioned above.
- The mixtures in the tails become very important when considering some of our data; this is due to the fact that the tails of such samples are slightly curved on a log-log plot and so the Pareto needs to be combined with another distribution in order to improve the fit notably.
- The use of the Singh–Maddala and Champernowne distributions instead of the lognormal all lead to a very important improvement. This means that the standard theory (Eeckhout, 2004) generating the lognormal can be enhanced notably.

The article is organised as follows. Section 2 describes the databases used. Section 3 motivates the need for the search of new and better distributions. Section 4 shows the definitions and main properties of the distributions studied. Section 5 shows the detailed results. In Section 6 we develop a theory that accommodates the newly preferred distributions and in Section 7 we offer a discussion. Finally, Section 8 concludes and A contains the proofs of statements in Section 6.

2 The databases

We use in this article data of US urban centers from three sources. The first is the decennial data of the US Census Bureau of "incorporated places" without any size restriction, in the period (1900-2000). They include governmental units classified under state laws as cities, towns, boroughs, or villages. Alaska, Hawaii and Puerto Rico have not been considered due to data limitations. The data have been collected from the original documents of the annual census published by the US Census Bureau[4]. This data has been first introduced in González-Val (2010), see therein for details, and later used in other works like González-Val et al. (2013a).

The second source consists of all US urban places, unincorporated and incorporated, and without size restrictions, as provided as well by the US Census Bureau for the years 2000 and 2010. The data for the year 2000 has been first used in Eeckhout (2004) and later in Levy (2009), Eeckhout (2009), Giesen et al. (2010), Ioannides and Skouras (2013) and Giesen and Suedekum (2013). The two samples have been used as well in González-Val et al. (2013a).

different ones. We started with the lognormal for the body as it is used in IS1, IS2. But we realized that a much better performance could be obtained with the Fisk ("Fi") distribution (Fisk, 1961) for the body and (the mixtures at) the Pareto tails. Both of the Singh–Maddala and Champernowne distributions generalize that of Fisk (and have one parameter more) so we tried them as well. For the sake of brevity, we present only the best results obtained, corresponding to the mentioned new distributions. We have also worked with (with obvious notation) tdPln, tdPFi, dm PlnP, dm PFiP, dm PSMP that, although all provide better results than the lognormal, dPln, IS1 and IS2, perform worse than the ones finally presented here.

[4] `http://www.census.gov/prod/www/decennial/` Last accessed: November 1st, 2013.

Table 1: Descriptive statistics of the US data samples used.

US Sample	Obs.	% of US pop.	Mean	SD	Min.	Max.
Inc. Places 1900	10,596	46.99	3,376	42,324	7	3,437,202
Inc. Places 1910	14,135	54.90	3,561	49,351	4	4,766,883
Inc. Places 1920	15,481	58.62	4,015	56,782	3	5,620,048
Inc. Places 1930	16,475	62.69	4,642	67,854	1	6,930,446
Inc. Places 1940	16,729	63.75	4,976	71,299	1	7,454,995
Inc. Places 1950	17,113	63.48	5,613	76,064	1	7,891,957
Inc. Places 1960	18,051	64.51	6,409	74,738	1	7,781,984
Inc. Places 1970	18,488	64.51	7,094	75,320	3	7,894,862
Inc. Places 1980	18,923	61.78	7,396	69,170	2	7,071,639
Inc. Places 1990	19,120	61.33	7,978	71,874	2	7,322,564
Inc. Places 2000	19,296	61.49	8,968	78,015	1	8,008,278
All places 2000	25,359	73.98	8,232	68,390	1	8,008,278
All places 2010	28,664	72.73	7,872	61,632	1	8,175,133
CCA 1991 (2000m)	30,201	97.46	8,180	104,954	1	12,511,237
CCA 1991 (3000m)	23,499	97.46	10,513	147,360	1	15,191,634
CCA 1991 (4000m)	19,912	97.46	12,407	180,751	2	17,064,816
CCA 1991 (5000m)	17,569	97.46	14,062	212,084	2	19,439,862
CCA 2000 (2000m)	30,201	96.08	8,977	108,342	1	12,734,150
CCA 2000 (3000m)	23,499	96.08	11,537	154,157	1	15,594,627
CCA 2000 (4000m)	19,912	96.08	13,615	190,528	1	17,567,010
CCA 2000 (5000m)	17,569	96.08	15,431	223,825	1	19,952,762

The third comes from a different and recent approach to defining city centers, described in detail in Rozenfeld et al. (2008, 2011). They use a so called "City Clustering Algorithm" (CCA) to get "an automated and systematic way of building population clusters based on the geographical location of people." (*loc. cit.*) We use their US clusters data based on the radii of 2, 3, 4, 5 km. and for the years 1991 and 2000. Such data has been used in Ioannides and Skouras (2013) and Giesen and Suedekum (2013).

The descriptive statistics of the data can be seen in Table 1. As Giesen and Suedekum (2013) indicate, the CCA data comprises a higher percentage of the whole population than the Census data.

3 Motivation of our approach

As a preliminary analysis, we take the sample of all US places in 2010, in order to see whether the previous dPln, IS1 and IS2 provide a good fit. For the last two, we use in advance some of the estimation results in Tables 2 and 3. In Figure 1 we show, in the left panel, the empirical and estimated (by maximum likelihood, ML) $\ln(\mathrm{cdf})$ against $\ln x$ for the lower tail. In the right panel, the analogous quantities $\ln(1-\mathrm{cdf})$ against $\ln x$ for the upper tail.[5] In the center panel, we show the usual empirical density functions (obtained through an adaptive Gaussian kernel) compared to the estimated density functions, all three for the case of the dPln (estimated previously for all US

[5]The difference of empirical and estimated quantities are amplified by the fact of taking the natural logarithms of cdf or $(1-\mathrm{cdf})$ for the lower and upper tails, respectively (González-Val et al., 2013a).

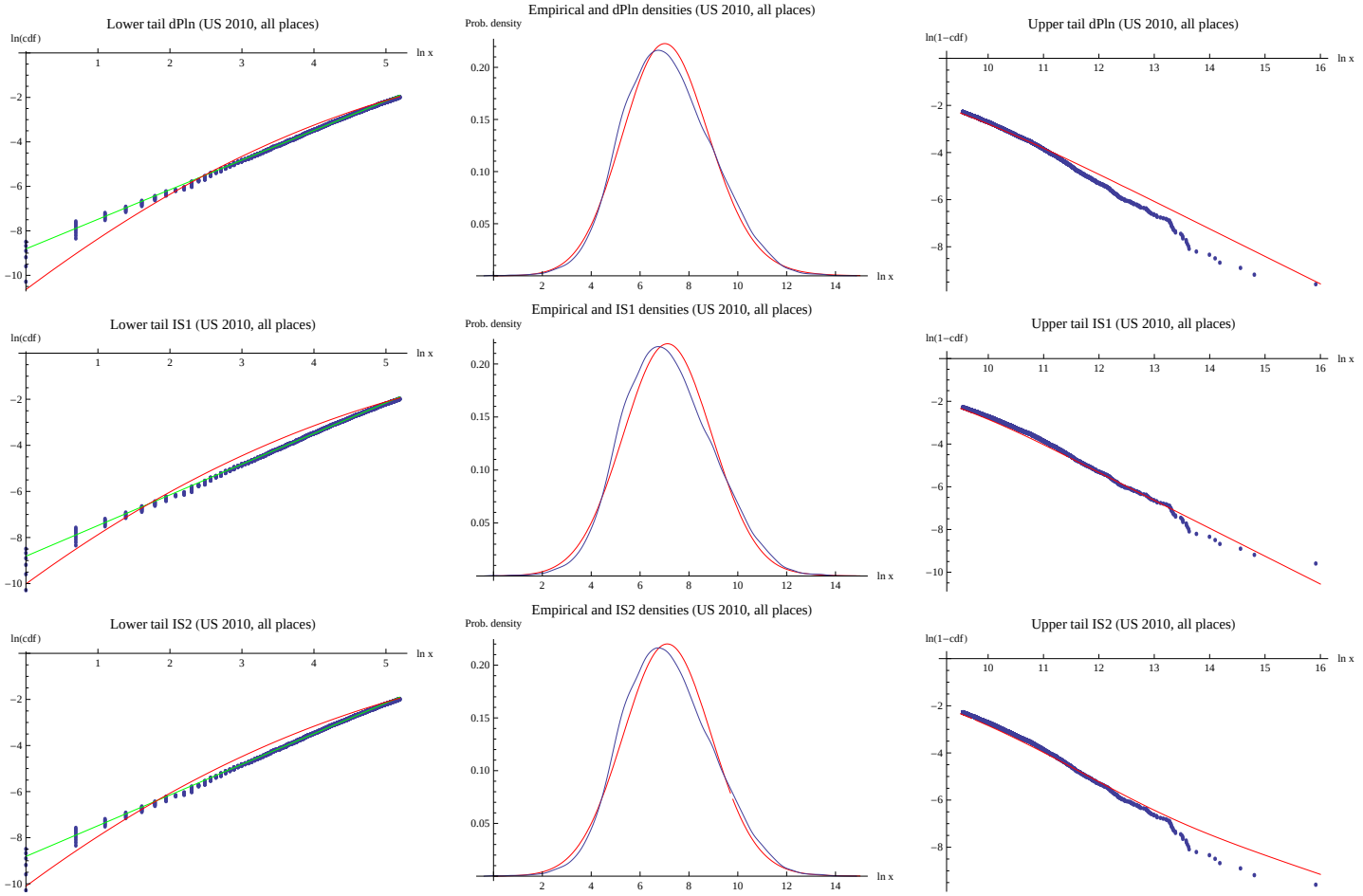

Figure 1: Left column: Empirical and estimated dPln, IS1 and IS2 $\ln(\mathrm{cdf})$ for the lower tail (linear OLS fit in green, empirical in blue, estimated in red). Center column: Empirical (Gaussian adaptive kernel density) and estimated dPln, IS1 and IS2 density functions (empirical in blue, estimated in red). Right column: Empirical and estimated dPln, IS1 and IS2 $\ln(1-\mathrm{cdf})$ for the upper tail (empirical in blue, estimated in red).

places (2010) in González-Val et al. (2013b)) and the IS1, IS2 (firstly estimated in this work for the same sample).

We see, in the left panel of Figure 1, that all of the dPln, IS1 and IS2 (in red) are not so linear as the empirical $\ln(\mathrm{cdf})$.[6] In the middle panel, we observe that the empirical and estimated densities differ clearly in the body and also in the tails. In the right panel, corresponding to the upper tails, we see that the fit is also not so good for the dPln (serious discrepancies starting at $\ln x > 11$, i.e., $x > 59874$ inhabitants), and IS1, IS2 performs better than the dPln to this respect.[7] Advancing some results of Table 8, we will see that both of two standard but demanding tests, given the high sample size, (Kolmogorov–Smirnov (KS) and Cramér-Von Mises (CM)) clearly reject the cited models.[8] Formally, the dPln is slightly more preferred than the IS1 and IS2, as the Akaike Information Criterium (AIC) and Bayesian Information Criterium (BIC) values obtained for the latter two are greater (and therefore unfavored) than those for the former, as Giesen and Suedekum (2013) indicate. This is because IS1 and IS2 fail to take into account the empirical power law behavior of the *lower tail*.

Therefore it makes sense to look for one or some distributions that cannot be rejected in a majority of cases and that offer a better fit to the data. We will see that this can be achieved by introducing some simple but significant changes in IS1 and IS2, which act as our baseline distributions.

4 Description of the distributions used

In this section we will introduce the distributions used along the paper. Firstly, we define some basic functions which will be employed by the distributions of Ioannides and Skouras (2013) and our new ones.

We thus set

$$f_{\ln}(x,\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

$$f_{\mathrm{SM}}(x,\mu,\sigma,\alpha) = \frac{\alpha\,(\mathrm{e}^{-\mu}x)^{1/\sigma}}{x\sigma(1+(\mathrm{e}^{-\mu}x)^{1/\sigma})^{1+\alpha}} \tag{2}$$

$$f_{\mathrm{Ch}}(x,\mu,\sigma,\beta) = \frac{\sin\beta}{x\beta\sigma((\mathrm{e}^{-\mu}x)^{-1/\sigma}+(\mathrm{e}^{-\mu}x)^{1/\sigma}+2\cos\beta)} \tag{3}$$

$$g(x,\zeta) = \frac{1}{x^{1+\zeta}} \tag{4}$$

$$h(x,\rho) = x^{\rho-1} \tag{5}$$

where $\mu,\sigma > 0$ are respectively the mean and the standard deviation of $\ln x$ for the

[6] A linear OLS estimation has been calculated and shown in green, only for reference purposes. Such estimation might be biased if one wants to obtain an accurate result, and an analysis similar to that of Gabaix and Ibragimov (2011) for the inverse rank might be necessary. However, our formal estimations will be performed by the standard maximum likelihood (ML).

[7] If one wants to quickly compare with our new results, see Figure 2.

[8] When performing the tests, we take the whole studied sample, and not subsamples, in order to achieve the maximum power of the KS and CM tests, compare with Giesen and Suedekum (2013).

lognormal density $f_{\ln}$. For the $f_{\rm SM}$, $f_{\rm Ch}$ distributions the corresponding $\mu, \sigma > 0$ are also related to the mean and standard deviation of $\ln x$ (Singh and Maddala, 1976; Champernowne, 1952).[9] The function $g(x, \zeta)$ will model the Pareto part of the upper tail of our distributions and $\zeta > 0$ is the Pareto exponent, and $h(x, \rho)$ corresponds to the Pareto lower tail, being $\rho > 1$ the power law exponent. The functions g, h are not normalized at this stage according to the practice of Ioannides and Skouras (2013).

4.1 The first distribution of Ioannides and Skouras (IS1)

The first distribution studied in Ioannides and Skouras (2013) is a lognormal with a Pareto upper tail, the transition between the two taking place at an exact threshold $\tau > 0$. The requirement is that the composite density function be continuous at $x = \tau$ and normalized to unity[10]. The resulting density function is

$$f_1(x, \mu, \sigma, \tau, \zeta) = \begin{cases} b_1 \, f_{\ln}(x, \mu, \sigma) & 0 < x \leq \tau \\ b_1 \, a_1 \, g(x, \zeta) & \tau < x \end{cases} \tag{6}$$

where a_1, b_1 are constants (depending on the parameters of the distribution) given by the following expressions:

$$a_1 = \frac{f_{\ln}(\tau, \mu, \sigma)}{g(\tau, \zeta)} \tag{7}$$

$$b_1^{-1} = \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{\mu - \ln \tau}{\sqrt{2}\sigma}\right)\right) + \frac{f_{\ln}(\tau, \mu, \sigma)}{\zeta \, \tau^{\zeta} \, g(\tau, \zeta)} \tag{8}$$

where erf denotes the error function associated with the normal distribution. This distribution depends on four parameters $(\mu, \sigma, \tau, \zeta)$ to be estimated. It is easy to see[11] that $f_1 \to f_{\ln}$ when $\tau \to \infty$, using the expressions of a_1 and b_1 given by (7) and (8), respectively.

4.2 The second distribution of Ioannides and Skouras (IS2)

The second distribution studied in Ioannides and Skouras (2013) is a variant of IS1 in which the upper tail is a linear combination of lognormal and Pareto distributions, the parameter θ being the combining coefficient[12]. The requirement of continuity of the density function at the threshold point is analogous to that of IS1 as well as that of the normalization. It is also imposed the condition

$$a_2 \int_{\tau}^{\infty} g(x, \zeta) \, dx = c_2 \int_{\tau}^{\infty} f_{\ln}(x, \mu, \sigma) \, dx$$

[9] We have taken the Champernowne density (2.4) in Champernowne (1952) with $\lambda = \cos \beta$ since this particular specification covers all the estimated cases in this paper. Also, the $f_{\rm SM}$ is directly related to the Burr Type XII distribution (Burr, 1942). See also Kleiber and Kotz (2003).

[10] Composite lognormal-Pareto models have been introduced previously by Cooray and Ananda (2005), Scollnik (2007), Malevergne et al. (2011) and Bee (2012).

[11] Details available from the authors upon request.

[12] The IS2 is referred to as CDGPR in Ioannides and Skouras (2013) because these authors were inspired by a similar combination used in Combes et al. (2012).

in order that the parameter θ controls the proportion of the density in the combination in the upper tail (Ioannides and Skouras, 2013). The resulting composite density is given by:

$$f_2(x,\mu,\sigma,\tau,\zeta,\theta) = \begin{cases} b_2\, f_{\ln}(x,\mu,\sigma) & 0 < x \leq \tau \\ b_2\, \left[(1-\theta)\, c_2\, f_{\ln}(x,\mu,\sigma) + \theta\, a_2\, g(x,\zeta)\right] & \tau < x \end{cases} \tag{9}$$

where now the constants are given as follows:

$$c_2^{-1} = 1 - \theta + \frac{\zeta\, \tau^{\zeta}\, \theta\, g(\tau,\zeta)}{2 f_{\ln}(\tau,\mu,\sigma)} \left(1 + \operatorname{erf}\left(\frac{\mu - \ln \tau}{\sqrt{2}\sigma}\right)\right) \tag{10}$$

$$a_2^{-1} = \frac{2(1-\theta)}{\zeta\, \tau^{\zeta} \left(1 + \operatorname{erf}\left(\frac{\mu - \ln \tau}{\sqrt{2}\sigma}\right)\right)} + \frac{\theta\, g(\tau,\zeta)}{f_{\ln}(\tau,\mu,\sigma)} \tag{11}$$

$$b_2^{-1} = \frac{1}{2}\left(1 - \operatorname{erf}\left(\frac{\mu - \ln \tau}{\sqrt{2}\sigma}\right)\right) + \frac{a_2}{\zeta\, \tau^{\zeta}} \tag{12}$$

This distribution depends on five parameters $(\mu, \sigma, \tau, \zeta, \theta)$ to be estimated. We also have the obvious relation $f_2 = f_1$ when $\theta = 1$.

4.3 The double Pareto lognormal distribution (dPln)

The probability density function of the double Pareto lognormal distribution is (Reed, 2002, 2003; Reed and Jorgensen, 2004):

$$\begin{aligned} f_3(x,\alpha,\beta,\mu,\sigma) = {} & \frac{\alpha\beta}{2x(\alpha+\beta)} \exp\left(\alpha\mu + \frac{\alpha^2\sigma^2}{2}\right) x^{-\alpha} \left(1 + \operatorname{erf}\left(\frac{\ln x - \mu - \alpha\sigma^2}{\sqrt{2}\sigma}\right)\right) \\ & - \frac{\alpha\beta}{2x(\alpha+\beta)} \exp\left(-\beta\mu + \frac{\beta^2\sigma^2}{2}\right) x^{\beta} \left(\operatorname{erf}\left(\frac{\ln x - \mu + \beta\sigma^2}{\sqrt{2}\sigma}\right) - 1\right) \end{aligned} \tag{13}$$

where $\alpha, \beta, \mu, \sigma > 0$ are the four distribution parameters to be estimated. The dPln distribution has the property that it approximates different power laws at its two tails, namely $f_3(x) \approx x^{-\alpha-1}$ when $x \to \infty$ and $f_3(x) \approx x^{\beta-1}$ when $x \to 0$, hence the name of double Pareto. The central part of the distribution is approximately lognormal, although it is not possible to exactly delineate the lognormal body part and the Pareto tails (Giesen et al., 2010).

The dPln distribution arises as the steady-state distribution of an evolutionary process of a simple stochastic model of settlement formation and growth based on Gibrat's law and a Yule process. Mathematically, the dPln is the log version of the convolution of the normal distribution and the (asymmetric) double Laplace distribution, see Reed (2002, 2003); Reed and Jorgensen (2004) and references therein for details.

For more recent work on an economic model which incorporates the stochastic derivation of Reed (2002, 2003), see Giesen and Suedekum (2012, 2013). The key in this latest model is the endogenous city creation and the resulting age heterogeneity in cities within the distribution. Giesen and Suedekum (2012, 2013) argue that Eeckhout

(2004) theoretical framework and the lognormal distribution represent a particular scenario of their model, the case when there is no city creation and all cities have the same age.

4.4 The threshold double Pareto Singh–Maddala (tdPSM)

We introduce here the first of our distributions. It is a variant of the IS1 in which we model the lower tail as a Pareto power law and the body as Singh–Maddala instead of lognormal. Thus, The tdPSM has a Singh–Maddala body and Pareto tails, the three regions exactly delineated by two thresholds: $\epsilon > 0$ separates the Pareto power law in the lower tail from the Singh–Maddala body, and $\tau > \epsilon$ separates the body from the Pareto power law in the upper tail. We impose continuity of the density function on the two threshold points and normalization of the former to unity. The resulting density reads

$$f_4(x, \rho, \epsilon, \mu, \sigma, \alpha, \tau, \zeta) = \begin{cases} b_4\, e_4\, h(x, \rho) & 0 < x < \epsilon \\ b_4\, f_{\mathrm{SM}}(x, \mu, \sigma, \alpha) & \epsilon \le x \le \tau \\ b_4\, a_4\, g(x, \zeta) & \tau < x \end{cases} \tag{14}$$

where now

$$e_4 = \frac{f_{\mathrm{SM}}(\epsilon, \mu, \sigma, \alpha)}{h(\epsilon, \rho)} \tag{15}$$

$$a_4 = \frac{f_{\mathrm{SM}}(\tau, \mu, \sigma, \alpha)}{g(\tau, \zeta)} \tag{16}$$

$$b_4^{-1} = e_4 \frac{\epsilon^\rho}{\rho} + \mathrm{e}^{\mu\alpha/\sigma}((\mathrm{e}^{\mu/\sigma} + \epsilon^{1/\sigma})^{-\alpha} - (\mathrm{e}^{\mu/\sigma} + \tau^{1/\sigma})^{-\alpha}) + \frac{a_4}{\zeta\, \tau^\zeta} \tag{17}$$

This distribution depends on seven parameters $(\rho, \epsilon, \mu, \sigma, \alpha, \tau, \zeta)$ to be estimated.

4.5 The double mixture Pareto Champernowne Pareto (dm PChP)

The second distribution we introduce is a variant of the IS2 in the sense that now we consider linear combinations of the Champernowne and respective Pareto distributions in the two tails, while maintaining a Champernowne body. The tails and the body are separated by two exact thresholds ϵ and τ with similar meaning to those of the tdPSM. For the lower tail, the combining coefficient will be denoted by ν, and θ for the upper tail as before. We require as usual continuity of the density function at the threshold points and overall normalization to one. The following conditions are also imposed:

$$a_5 \int_\tau^\infty g(x, \zeta)\, dx = c_5 \int_\tau^\infty f_{\mathrm{Ch}}(x, \mu, \sigma, \beta)\, dx$$
$$e_5 \int_0^\epsilon h(x, \rho)\, dx = d_5 \int_0^\epsilon f_{\mathrm{Ch}}(x, \mu, \sigma, \beta)\, dx$$

in order that the parameters θ, ν control the proportion of the density in the combination in the upper (resp. lower) tail, analogously to the θ of the IS2. The resulting composite

density is given by:

$$f_5(x,\rho,\epsilon,\nu,\mu,\sigma,\beta,\tau,\zeta,\theta) = \begin{cases} b_5\ [(1-\nu)\, d_5\, f_{\mathrm{Ch}}(x,\mu,\sigma,\beta) + \nu\, e_5\, h(x,\rho)] & 0 < x < \epsilon \\ b_5\, f_{\mathrm{Ch}}(x,\mu,\sigma,\beta) & \epsilon \leq x \leq \tau \\ b_5\ [(1-\theta)\, c_5\, f_{\mathrm{Ch}}(x,\mu,\sigma,\beta) + \theta\, a_5\, g(x,\zeta)] & \tau < x \end{cases} \tag{18}$$

where now the constants are given as follows:

$$d_5^{-1} = 1-\nu+\frac{\nu\rho(\beta-\operatorname{arccot}[\cot\beta+(\mathrm{e}^{-\mu}\epsilon)^{1/\sigma}\csc\beta])h(\epsilon,\rho)}{\epsilon^{\rho}\beta f_{\mathrm{Ch}}(\epsilon,\mu,\sigma,\beta)} \tag{19}$$

$$e_5^{-1} = \frac{\beta\epsilon^{\rho}(1-\nu)}{\rho(\beta-\operatorname{arccot}[\cot\beta+(\mathrm{e}^{-\mu}\epsilon)^{1/\sigma}\csc\beta])}+\frac{\nu\, h(\epsilon,\rho)}{f_{\mathrm{Ch}}(\epsilon,\mu,\sigma,\beta)} \tag{20}$$

$$c_5^{-1} = 1-\theta+\frac{\theta\zeta\tau^{\zeta}\operatorname{arccot}[\cot\beta+(\mathrm{e}^{-\mu}\tau)^{1/\sigma}\csc\beta]g(\tau,\zeta)}{\beta f_{\mathrm{Ch}}(\tau,\mu,\sigma,\beta)} \tag{21}$$

$$a_5^{-1} = \frac{\beta(1-\theta)}{\zeta\tau^{\zeta}\operatorname{arccot}[\cot\beta+(\mathrm{e}^{-\mu}\tau)^{1/\sigma}\csc\beta]}+\frac{\theta\, g(\tau,\zeta)}{f_{\mathrm{Ch}}(\tau,\mu,\sigma,\beta)} \tag{22}$$

$$b_5^{-1} = e_5\frac{\epsilon^{\rho}}{\rho}+\frac{1}{\beta}\arctan\left(\frac{\sin\beta}{(\mathrm{e}^{-\mu}\epsilon)^{1/\sigma}+\cos\beta}\right) - \frac{1}{\beta}\arctan\left(\frac{\sin\beta}{(\mathrm{e}^{-\mu}\tau)^{1/\sigma}+\cos\beta}\right)+\frac{a_5}{\zeta\,\tau^{\zeta}} \tag{23}$$

This distribution depends on nine parameters $(\rho,\epsilon,\nu,\mu,\sigma,\beta,\tau,\zeta,\theta)$ to be estimated.

5 Results

5.1 Estimation of the distributions

Maximum likelihood (ML) is a standard technique which allows the estimation of the parameters of a distribution given a sample of data. For the case of the lognormal density function, the corresponding ML estimators can be found easily in an exact closed form (the μ and σ are then the mean and the standard deviation (SD) of the natural logarithm of the data). However, for the other distributions $f_1, \ldots, f_5$ used in this article one must resort to numerical optimization methods in order to find the ML estimators[13]. It is worth noting that the threshold population parameters ϵ and τ present in the cited density functions are to be estimated endogenously by ML, letting the data "decide" what are the optimum threshold values which maximize the log-likelihood.

Previous work on similar matters include that of Bee (2012), which deals with a distribution similar to the IS1 with ML. Also, the log-likelihood function of the dPln is found in Reed and Jorgensen (2004). Of course, Ioannides and Skouras (2013) estimate their IS1 and IS2 by ML. The other cases of this paper can be dealt with in a similar fashion.[14]

[13] We have used MATLAB in order to perform the ML estimations as Ioannides and Skouras (2013) did.

[14] More details are available from the authors upon request.

Table 2: Estimators and 95% confidence intervals of the parameters of the IS1 for the US (places) samples. The estimators for the lognormal are the mean and the standard deviation of $\ln(pop)$

US Sample	ln		IS1			
	μ	σ	μ	σ	τ	ζ
Inc. Places 1900	6.65	1.26	6.31±0.03	0.89±0.03	1,131±196	0.91±0.03
Inc. Places 1910	6.65	1.29	6.26±0.03	0.88±0.02	1,025±148	0.87±0.02
Inc. Places 1920	6.67	1.32	6.29±0.03	0.90±0.02	1,074±157	0.86±0.02
Inc. Places 1930	6.69	1.40	6.30±0.03	0.98±0.02	1,184±203	0.81±0.02
Inc. Places 1940	6.78	1.43	6.38±0.03	1.01±0.02	1,324±215	0.79±0.02
Inc. Places 1950	6.84	1.50	6.51±0.03	1.15±0.02	1,896±321	0.79±0.02
Inc. Places 1960	6.92	1.61	6.61±0.04	1.28±0.03	2,566±445	0.76±0.02
Inc. Places 1970	7.00	1.67	6.74±0.04	1.38±0.03	3,599±680	0.76±0.03
Inc. Places 1980	7.11	1.66	6.86±0.04	1.40±0.03	4,343±832	0.77±0.03
Inc. Places 1990	7.10	1.74	6.90±0.04	1.53±0.03	6,153±1,381	0.78±0.03
Inc. Places 2000	7.18	1.78	7.01±0.04	1.59±0.03	8,063±1,989	0.79±0.04
All places 2000	7.28	1.75	7.26±0.02	1.73±0.02	60,326±35,844	1.25±0.11
All places 2010	7.13	1.83	7.12±0.02	1.82±0.02	93,350±66,640	1.31±0.15

When performing the estimations, not all density functions can be treated always by our numerical procedure, because it seems that in the corresponding cases the estimators simply do not exist. This may happen when dealing with composite densities, see, e.g., Bee (2012) for a theoretical discussion in a related sample situation. Specifically, for the US places data typically the dm PChP cannot be estimated, so for the sake of comparison and brevity we include only the results of the new distributions which can be all estimated for each type of data (US places and CCA clusters) and for all periods, and provide the best performance.

We present such results of the estimation procedure for the US places data in Tables 2, 3, 4 and 5. For the sample of the US (2000, all places) we essentially replicate the results of Ioannides and Skouras (2013),[15] Giesen et al. (2010) and Giesen and Suedekum (2013). We have found that the log-likelihood function is smooth near its maximum in all of the estimated cases, see also Bee (2012).

We see that there are two distributions, apart from the lognormal, for which the estimates are rather stable or present a soft trend, without, first, "sudden jumps" (for sudden jumps see, e.g., the estimates of τ for the IS1 and IS2 when passing to all places), and without, secondly, surprisingly low estimates for the upper threshold τ (see, e.g., the estimates of τ of IS2 for the samples of US incorporated places in the whole period 1900-2000), namely the dPln and tdPSM. Of these two, only the last offer an estimate of the lower (ϵ) and upper (τ) thresholds ($\epsilon \in (99, 178)$ and $\tau \in (3405, 55274)$). This is an observed first good feature of the tdPSM.

We show next the estimation results for the US CCA samples in Tables 6 and 7. For these data, we also replicate essentially the results of Ioannides and Skouras (2013) and Giesen and Suedekum (2013). Moreover, the estimation results yield in general more

[15] We provide 95% confidence intervals meanwhile Ioannides and Skouras (2013) provide standard errors. Both quantities are related and give essentially the same information. Also, there are slight differences in the values of τ but within the confidence intervals.

Table 3: Estimators and 95% confidence intervals of the parameters of the IS2 for the US (places) samples

US Sample	IS2 μ	σ	τ	ζ	θ
Inc. Places 1900	6.90±0.11	1.05±0.03	395±2	0.81±0.03	0.66±0.05
Inc. Places 1910	7.06±0.11	1.10±0.03	364±2	0.80±0.03	0.67±0.04
Inc. Places 1920	7.01±0.10	1.08±0.03	361±1	0.77±0.03	0.67±0.03
Inc. Places 1930	7.34±0.12	1.25±0.03	384±2	0.78±0.03	0.71±0.03
Inc. Places 1940	7.51±0.11	1.31±0.03	405±2	0.79±0.03	0.67±0.03
Inc. Places 1950	7.44±0.09	1.38±0.02	408±2	0.75±0.03	0.58±0.03
Inc. Places 1960	7.59±0.08	1.50±0.02	399±2	0.74±0.03	0.50±0.03
Inc. Places 1970	7.66±0.08	1.59±0.02	437±2	0.75±0.04	0.44±0.03
Inc. Places 1980	7.76±0.07	1.59±0.02	487±3	0.77±0.04	0.42±0.03
Inc. Places 1990	7.72±0.07	1.69±0.02	481±3	0.78±0.05	0.37±0.04
Inc. Places 2000	7.82±0.07	1.74±0.02	518±4	0.80±0.06	0.34±0.04
All places 2000	7.25±0.02	1.72±0.02	16,111±10,888	0.82±0.16	0.25±0.17
All places 2010	7.11±0.02	1.81±0.02	16,397±11,108	0.80±0.16	0.20±0.15

Table 4: Estimators and 95% confidence intervals of the parameters of the dPln for the US (places) samples

US Sample	dPln α	β	μ	σ
Inc. Places 1900	0.92±0.03	2.64±0.27	5.95±0.04	0.58±0.04
Inc. Places 1910	0.89±0.03	2.96±0.35	5.86±0.04	0.61±0.04
Inc. Places 1920	0.87±0.03	2.78±0.27	5.88±0.04	0.60±0.04
Inc. Places 1930	0.80±0.02	2.21±0.14	5.89±0.04	0.57±0.04
Inc. Places 1940	0.79±0.02	2.20±0.15	5.96±0.04	0.61±0.04
Inc. Places 1950	0.80±0.03	2.15±0.17	6.06±0.05	0.78±0.04
Inc. Places 1960	0.80±0.03	2.24±0.26	6.11±0.06	0.96±0.05
Inc. Places 1970	0.83±0.03	2.62±0.22	6.18±0.05	1.13±0.04
Inc. Places 1980	0.86±0.02	3.65±0.02	6.23±0.02	1.19±0.01
Inc. Places 1990	0.87±0.02	3.59±0.01	6.23±0.01	1.31±0.003
Inc. Places 2000	0.87±0.02	3.55±0.01	6.32±0.02	1.36±0.003
All places 2000	1.23±0.03	3.16±0.003	6.78±0.01	1.52±0.002
All places 2010	1.17±0.03	2.97±0.004	6.61±0.01	1.59±0.008

Table 5: Estimators and 95% confidence intervals of the parameters of the tdPSM for the US (places) samples

US Sample	tdPSM ρ	ϵ	μ	σ	α	τ	ζ
Inc. Places 1900	2.32±0.14	172±1	5.64±0.09	0.42±0.06	0.32±0.07	3,405±97	1.02±0.05
Inc. Places 1910	2.48±0.15	147±1	5.62±0.06	0.44±0.04	0.34±0.05	8,190±308	1.09±0.06
Inc. Places 1920	2.36±0.12	167±1	5.60±0.08	0.45±0.05	0.33±0.06	4,310±127	0.98±0.04
Inc. Places 1930	2.06±0.09	178±1	5.52±0.06	0.45±0.05	0.31±0.05	8,465±222	1.00±0.05
Inc. Places 1940	2.01±0.09	177±1	5.53±0.06	0.44±0.05	0.28±0.05	10,359±229	1.06±0.05
Inc. Places 1950	1.89±0.09	150±1	5.62±0.08	0.54±0.06	0.34±0.05	11,741±382	1.06±0.05
Inc. Places 1960	1.72±0.07	148±1	5.55±0.09	0.61±0.07	0.32±0.06	13,917±405	1.07±0.05
Inc. Places 1970	1.60±0.07	141±1	5.71±0.10	0.69±0.07	0.38±0.06	25,937±682	1.18±0.07
Inc. Places 1980	1.69±0.08	129±1	5.84±0.10	0.69±0.06	0.38±0.05	34,196±571	1.30±0.08
Inc. Places 1990	1.51±0.06	140±1	5.91±0.14	0.85±0.08	0.48±0.08	41,945±1,003	1.31±0.08
Inc. Places 2000	1.60±0.08	99±1	5.88±0.11	0.79±0.06	0.40±0.05	47,386±851	1.35±0.08
All places 2000	1.46±0.06	127±1	6.80±0.24	1.14±0.08	0.82±0.14	36,081±746	1.33±0.07
All places 2010	1.31±0.04	133±1	6.84±0.28	1.31±0.11	1.00±0.18	55,274±1,063	1.45±0.09

stable and precise values. The estimation process is smoother than for the places data, and the distribution dm PChP can be estimated for all of these samples. This is a remarkable feature of the cluster data: The City Clustering Algorithm considers as an urban center an actual agglomeration of people within a prescribed radius, irrespectively of legally established borders, giving an economic and physical entity to the considered clusters. This fact seems to reflect in the obtained data, which allows the estimation of more density functions and, in general, with narrower confidence intervals. For the dm PChP ϵ varies between 1118 and 2671 and τ between 14253 and 20381.

We have used the graphical tools in Section 3 to introduce the need of going further in the search of distributions with better fit. But when performing a high precision exercise, such graphical tools can be misleading in assessing the quality of fit, see González-Val et al. (2013a). So we resort to standard statistical tests and information criteria to see when the hypothesized distributions offer a good fit and what model is the selected one amongst the studied ones. This is done in the following subsections.

5.2 Standard statistical tests

In this subsection we provide independent tests in order to verify the goodness of fit in all of the studied cases. As in González-Val et al. (2013b) we have chosen the Kolmogorov–Smirnov (KS) test, which is also mentioned in Giesen et al. (2010), Giesen and Suedekum (2012, 2013) and is standard in the literature; also the Cramér-von Mises (CM) test, cited in turn in Ioannides and Skouras (2013).

Moreover, the KS and CM tests have similar power: It is quite low for small sample sizes but very high for large sample sizes (Razali and Wah, 2011). Both tests are extremely precise for large and very large sample sizes as the ones used in this paper, for which the non rejections only occur if the deviations (statistics) are extremely small. Significance level is chosen to be always 5%. Non rejections are indicated in boldface.

Table 6: Estimators and 95% confidence intervals of the parameters of the IS1, IS2 and dPln for the US CCA clusters samples. The estimators for the lognormal are the mean and the standard deviation of $\ln(pop)$

US Sample	ln		IS1			
	μ	σ	μ	σ	τ	ζ
CCA 1991 (2000m)	8.33	0.85	8.29±0.01	0.77±0.01	29,944±1,223	1.02±0.08
CCA 1991 (3000m)	8.32	0.89	8.26±0.01	0.75±0.01	25,709±990	0.88±0.06
CCA 1991 (4000m)	8.32	0.92	8.24±0.01	0.75±0.01	23,207±886	0.85±0.06
CCA 1991 (5000m)	8.33	0.95	8.23±0.01	0.75±0.01	21,891±856	0.85±0.06
CCA 2000 (2000m)	8.44	0.87	8.40±0.01	0.80±0.01	37,224±1,667	1.03±0.09
CCA 2000 (3000m)	8.43	0.91	8.37±0.01	0.79±0.01	30,635±1,262	0.92±0.07
CCA 2000 (4000m)	8.42	0.94	8.34±0.01	0.78±0.01	27,571±1,125	0.87±0.06
CCA 2000 (5000m)	8.42	0.97	8.33±0.01	0.79±0.01	26,679±1,125	0.85±0.06

US Sample	IS2				
	μ	σ	τ	ζ	θ
CCA 1991 (2000m)	8.29±0.01	0.77±0.01	28,121±1,481	0.98±0.11	0.93±0.07
CCA 1991 (3000m)	8.26±0.01	0.75±0.01	27,191±1,229	0.93±0.09	1.06±0.05
CCA 1991 (4000m)	8.24±0.11	0.75±0.01	23,880±1,107	0.86±0.08	1.02±0.05
CCA 1991 (5000m)	8.23±0.01	0.76±0.01	21,202±1,039	0.83±0.08	0.97±0.06
CCA 2000 (2000m)	8.40±0.01	0.80±0.01	34,321±1,978	0.98±0.12	0.92±0.08
CCA 2000 (3000m)	8.37±0.01	0.79±0.01	30,906±1,550	0.92±0.10	1.01±0.06
CCA 2000 (4000m)	8.34±0.01	0.78±0.01	27,433±1,371	0.87±0.09	1.00±0.06
CCA 2000 (5000m)	8.33±0.01	0.79±0.01	26,608±1,362	0.85±0.08	1.00±0.06

US Sample	dPln			
	α	β	μ	σ
CCA 1991 (2000m)	1.95±0.04	1.85±0.03	8.36±0.01	0.14±0.02
CCA 1991 (3000m)	1.76±0.04	1.86±0.04	8.29±0.01	0.11±0.02
CCA 1991 (4000m)	1.64±0.03	1.88±0.04	8.25±0.01	0.10±0.02
CCA 1991 (5000m)	1.54±0.03	1.87±0.05	8.22±0.01	0.10±0.03
CCA 2000 (2000m)	1.86±0.04	1.82±0.03	8.45±0.01	0.18±0.02
CCA 2000 (3000m)	1.66±0.03	1.83±0.04	8.37±0.01	0.16±0.02
CCA 2000 (4000m)	1.55±0.03	1.84±0.05	8.32±0.02	0.15±0.03
CCA 2000 (5000m)	1.46±0.03	1.83±0.05	8.29±0.02	0.14±0.03

Table 7: Estimators and 95% confidence intervals of the parameters of the dm PChP for the US CCA clusters samples

US Sample	dm PChP ρ	ϵ	ν	μ	σ	β	τ	ζ	θ
CCA 1991 (2000m)	0.59±0.07	2,091±136	0.22±0.04	8.35±0.01	0.37±0.02	1.29±0.22	17,171±898	0.96±0.11	0.78±0.10
CCA 1991 (3000m)	0.63±0.09	2,134±161	0.19±0.05	8.31±0.01	0.37±0.02	1.31±0.24	16,903±853	0.87±0.08	0.90±0.08
CCA 1991 (4000m)	0.63±0.11	1,963±173	0.18±0.06	8.29±0.01	0.39±0.03	1.45±0.24	16,495±864	0.83±0.08	0.92±0.08
CCA 1991 (5000m)	0.57±0.12	2,671±314	0.09±0.03	8.27±0.01	0.42±0.03	1.62±0.21	15,773±852	0.83±0.08	0.92±0.09
CCA 2000 (2000m)	0.54±0.07	1,371±114	0.36±0.07	8.44±0.01	0.39±0.02	1.13±0.24	20,381±1,231	0.95±0.12	0.69±0.11
CCA 2000 (3000m)	0.56±0.09	1,323±134	0.32±0.08	8.40±0.01	0.40±0.02	1.21±0.25	19,912±1,122	0.87±0.09	0.84±0.10
CCA 2000 (4000m)	0.57±0.11	1,118±140	0.33±0.09	8.38±0.01	0.42±0.02	1.36±0.24	20,083±1,173	0.84±0.09	0.89±0.10
CCA 2000 (5000m)	0.58±0.12	1,279±166	0.26±0.09	8.35±0.01	0.42±0.03	1.26±0.30	14,253±797	0.71±0.08	0.71±0.08

Table 8: Results of the Kolmogorov–Smirnov (KS) and Cramér–Von Mises (CM) tests for the US places samples and the used density functions. Non-rejections are marked in boldface

US Sample	ln		IS1	
	KS	CM	KS	CM
Inc. Places 1900	0 (0.07)	0 (17.22)	0.04 (0.01)	0.02 (0.62)
Inc. Places 1910	0 (0.07)	0 (21.81)	0 (0.02)	0.003 (1.10)
Inc. Places 1920	0 (0.07)	0 (25.87)	0.002 (0.02)	0.003 (1.09)
Inc. Places 1930	0 (0.07)	0 (27.59)	0.001 (0.02)	0 (1.34)
Inc. Places 1940	0 (0.07)	0 (25.59)	0 (0.021)	0 (1.86)
Inc. Places 1950	0 (0.06)	0 (17.55)	0 (0.020)	0 (1.91)
Inc. Places 1960	0 (0.05)	0 (14.26)	0 (0.026)	0 (2.82)
Inc. Places 1970	0 (0.05)	0 (12.88)	0 (0.026)	0 (2.85)
Inc. Places 1980	0 (0.04)	0 (11.36)	0 (0.027)	0 (3.24)
Inc. Places 1990	0 (0.04)	0 (9.10)	0 (0.027)	0 (3.24)
Inc. Places 2000	0 (0.04)	0 (9.35)	0 (0.030)	0 (3.72)
All places 2000	0 (0.02)	0 (2.69)	0 (0.02)	0 (2.31)
All places 2010	0 (0.02)	0 (1.41)	0 (0.03)	0 (4.53)
US Sample	IS2		dPln	
	KS	CM	KS	CM
Inc. Places 1900	**0.28 (0.01)**	**0.29 (0.19)**	0.03 (0.01)	**0.07 (0.42)**
Inc. Places 1910	**0.07 (0.01)**	**0.19 (0.25)**	0.001 (0.02)	0.02 (0.66)
Inc. Places 1920	0.045 (0.012)	**0.10 (0.35)**	0.02 (0.013)	**0.09 (0.37)**
Inc. Places 1930	0.03 (0.012)	0.04 (0.52)	0 (0.017)	0 (1.19)
Inc. Places 1940	0.04 (0.011)	**0.05 (0.45)**	0 (0.021)	0 (1.60)
Inc. Places 1950	**0.068 (0.010)**	**0.06 (0.44)**	0 (0.021)	0 (1.64)
Inc. Places 1960	0.049 (0.011)	**0.054 (0.45)**	0 (0.024)	0 (2.02)
Inc. Places 1970	0.029 (0.011)	0.037 (0.51)	0 (0.021)	0 (1.75)
Inc. Places 1980	**0.10 (0.009)**	**0.071 (0.40)**	0 (0.021)	0 (1.99)
Inc. Places 1990	**0.11 (0.009)**	**0.070 (0.40)**	0 (0.021)	0 (2.03)
Inc. Places 2000	0.02 (0.012)	**0.080 (0.38)**	0 (0.020)	0 (2.28)
All places 2000	0 (0.02)	0 (2.25)	0.005 (0.01)	0.005 (1.00)
All places 2010	0 (0.02)	0 (3.93)	0 (0.02)	0 (1.83)

US Sample	tdPSM	
	KS	CM
Inc. Places 1900	**0.99 (0.005)**	**0.97 (0.03)**
Inc. Places 1910	**0.62 (0.007)**	**0.84 (0.06)**
Inc. Places 1920	**0.50 (0.007)**	**0.65 (0.09)**
Inc. Places 1930	**0.96 (0.004)**	**0.97 (0.03)**
Inc. Places 1940	**0.90 (0.005)**	**0.96 (0.03)**
Inc. Places 1950	**0.87 (0.005)**	**0.78 (0.06)**
Inc. Places 1960	**0.93 (0.004)**	**0.85 (0.05)**
Inc. Places 1970	**0.94 (0.004)**	**0.96 (0.03)**
Inc. Places 1980	**0.54 (0.006)**	**0.48 (0.12)**
Inc. Places 1990	**0.71 (0.006)**	**0.75 (0.07)**
Inc. Places 2000	**0.88 (0.005)**	**0.90 (0.05)**
All places 2000	**0.65 (0.005)**	**0.47 (0.13)**
All places 2010	**0.17 (0.007)**	**0.29 (0.19)**

Table 9: Results of the Kolmogorov–Smirnov (KS) and Cramér–Von Mises (CM) tests for the US CCA clusters samples and the used density functions. Non-rejections are marked in boldface

US Sample	ln		IS1		IS2	
	KS	CM	KS	CM	KS	CM
CCA 1991 (2000m)	0 (0.09)	0 (92.70)	0 (0.09)	0 (66.53)	0 (0.09)	0 (65.57)
CCA 1991 (3000m)	0 (0.10)	0 (86.75)	0 (0.08)	0 (43.14)	0 (0.08)	0 (45.35)
CCA 1991 (4000m)	0 (0.11)	0 (78.08)	0 (0.08)	0 (35.06)	0 (0.08)	0 (33.26)
CCA 1991 (5000m)	0 (0.11)	0 (74.02)	0 (0.08)	0 (28.57)	0 (0.07)	0 (27.85)
CCA 2000 (2000m)	0 (0.09)	0 (73.26)	0 (0.08)	0 (49.12)	0 (0.08)	0 (49.37)
CCA 2000 (3000m)	0 (0.09)	0 (71.00)	0 (0.07)	0 (33.92)	0 (0.07)	0 (33.44)
CCA 2000 (4000m)	0 (0.09)	0 (62.27)	0 (0.07)	0 (23.98)	0 (0.07)	0 (24.97)
CCA 2000 (5000m)	0 (0.10)	0 (58.44)	0 (0.07)	0 (20.50)	0 (0.07)	0 (20.27)

US Sample	dPln		dm PChP	
	KS	CM	KS	CM
CCA 1991 (2000m)	0 (0.02)	0 (1.84)	**0.86 (0.004)**	**0.82 (0.06)**
CCA 1991 (3000m)	0 (0.02)	0 (2.42)	**0.64 (0.005)**	**0.74 (0.07)**
CCA 1991 (4000m)	0 (0.03)	0 (2.46)	**0.86 (0.005)**	**0.69 (0.08)**
CCA 1991 (5000m)	0 (0.03)	0 (2.21)	**0.61 (0.006)**	**0.60 (0.10)**
CCA 2000 (2000m)	0 (0.02)	0.003 (1.09)	**0.58 (0.005)**	**0.73 (0.07)**
CCA 2000 (3000m)	0 (0.02)	0 (1.18)	**0.55 (0.006)**	**0.43 (0.14)**
CCA 2000 (4000m)	0 (0.04)	0 (1.79)	**0.36 (0.007)**	**0.28 (0.19)**
CCA 2000 (5000m)	0 (0.05)	0 (2.22)	**0.46 (0.007)**	**0.51 (0.12)**

We show in Table 8 the results for the samples of US places. We offer the p-values of the tests jointly with the values of the statistics (in parentheses). A first observation is that the lognormal model is very strongly rejected for all samples, and the IS1 is as well rejected always although with a lower value of the tests' statistics than for the lognormal. The dPln is as well rejected in almost all cases (two exception). The IS2 in turn is not rejected the 53.84% of the cases: The mixing lognormal-Pareto in the upper tail means an improvement. And a big jump in performance is obtained with the tdPSM. Indeed, such distribution is not rejected 100% of the cases. Thus, modeling both tails as a pure Pareto and the body as the Singh–Maddala distribution means a striking better improvement. Thus, the tdPSM reveals itself as an excellent and robust specification for the US places size distribution.

We move to the results of the tests for the US CCA clusters in Table 9. Again, we show the p-values and the tests' statistics in parentheses. Here, the lognormal is again strongly rejected always and also the IS1 and IS2. The dPln is rejected always as well (with lower values of the tests' statistics). Again, a wide jump is obtained when considering the dm PChP: It is not rejected 100% of all instances. This means that modeling the two tails as a mixing of Pareto-Champernowne and the body as Champernowne leads to an excellent fit. These final results are robust to the different radii the clusters are constructed with (2, 3, 4 and 5 km.), and to the years studied (1991 and 2000). We obtain in this way an excellent model for the US CCA clusters size distribution: The dm PChP.

In the next subsection we study the distributions with the information criteria.

5.3 Information criteria

In order to select a model between the studied distributions, we follow another approach: To compute two information criteria very well suited to the maximum likelihood method which we have used in order to estimate the parameters of the distributions studied: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (see, e.g., Burnham and Anderson (2002, 2004); Giesen et al. (2010) and references therein). In the first two of these references it is argued, theoretically and by means of simulations, that the AIC is preferable to the BIC, and in case of discrepancy between the two information criteria, we prefer to follow the outcome of the former.

We show in Table 10 the results for the US places samples and the presented distributions. We obtain a similar result to those of the KS and CM tests: Choosing an ordering of ascending values of the AIC for each sample (the results with the BIC are almost exactly the same) we obtain a robust ordering of the distributions (the lower the value of AIC, the more preferred the distribution). For the US incorporated places and all places samples in the period (1900-2010) we have

$$\mathrm{AIC}_{\mathrm{tdPSM}} < \mathrm{AIC}_{\mathrm{dPln}} < \mathrm{AIC}_{\mathrm{IS2}} < \mathrm{AIC}_{\mathrm{IS1}} < \mathrm{AIC}_{\mathrm{ln}}$$

Therefore, the selected model is the tdPSM 100% of the samples. This, jointly with the outcomes of the KS and CM tests, yields a new and strong result: The US city size distribution (incorporated places and all places) can be safely taken as the new tdPSM.

For the US CCA cluster samples, we refer to Table 11. We have again strong regularities. The ordering of the distributions by ascending values of AIC is (the ordering by BIC is practically the same)

$$\mathrm{AIC}_{\mathrm{dmPChP}} < \mathrm{AIC}_{\mathrm{dPln}} < \mathrm{AIC}_{\mathrm{IS1}} < \mathrm{AIC}_{\mathrm{IS2}} < \mathrm{AIC}_{\mathrm{ln}}$$

The difference between IS1 and IS2 is very small (they are tied in two out of the eight samples). And it is striking the result that our new distribution dm PChP is systematically preferred to others known up to now in the literature. In short, we have that the selected model (amongst those studied here and others not shown for the sake of brevity) is the dm PChP the 100% of the cases, with values of the AIC and BIC quite lower than for the other previously know distributions. This, jointly with the results of the KS and CM tests, yields a second strong and new result: The US city size distribution (CCA clusters) can be safely taken as the new dm PChP.

In both of the cases of US places and CCA clusters samples, we have another result: To achieve an exceptional performance, it seems to be essential to model both tails as a Pareto distribution, in a pure form (places), with Singh–Maddala body, or as part of a mixing with the Champernowne distribution (clusters), with body of the same type of the latter.

As a complement of the KS, CM, AIC and BIC results, we show in Figure 2 an informal graphical approximation of the obtained fits in two different cases: The first row for the sample of all US places (2010) and the tdPSM, and the second for the

Table 10: Maximum log-likelihoods, AIC and BIC for the used distributions and the US places data. The lowest values of AIC and BIC for each sample are marked in boldface

US Sample	ln log-likelihood	AIC	BIC	IS1 log-likelihood	AIC	BIC
Inc. Places 1900	-87,943	175,891	175,905	-87,290	174,588	174,617
Inc. Places 1910	-117,640	235,284	235,299	-116,769	233,546	233,576
Inc. Places 1920	-129,580	259,164	259,179	-128,576	257,160	257,191
Inc. Places 1930	-139,194	278,392	278,407	-138,254	276,516	276,547
Inc. Places 1940	-143,097	286,198	286,213	-142,289	284,586	284,617
Inc. Places 1950	-148,254	296,512	296,528	-147,679	295,366	295,397
Inc. Places 1960	-159,142	318,288	318,304	-158,758	317,524	317,555
Inc. Places 1970	-165,171	330,346	330,362	-164,907	329,822	329,853
Inc. Places 1980	-171,088	342,180	342,196	-170,864	341,736	341,767
Inc. Places 1990	-173,472	346,948	346,964	-173,333	346,674	346,705
Inc. Places 2000	-177,127	354,258	354,274	-177,031	354,070	354,101
All places 2000	-234,773	469,550	469,566	-234,756	469,519	469,552
All places 2010	-262,440	524,884	524,901	-262,433	524,874	524,907

US Sample	IS2 log-likelihood	AIC	BIC	dPln log-likelihood	AIC	BIC
Inc. Places 1900	-87,273	174,555	174,592	-87,254	174,516	174,545
Inc. Places 1910	-116,732	233,474	233,512	-116,727	233,462	233,492
Inc. Places 1920	-128,539	257,088	257,126	-128,521	257,050	257,081
Inc. Places 1930	-138,164	276,338	276,377	-138,129	276,266	276,297
Inc. Places 1940	-142,174	284,358	284,397	-142,179	284,366	284,397
Inc. Places 1950	-147,574	295,158	295,197	-147,593	295,194	295,225
Inc. Places 1960	-158,605	317,220	317,259	-158,679	317,366	317,397
Inc. Places 1970	-164,741	329,492	329,531	-164,831	329,670	329,701
Inc. Places 1980	-170,682	341,374	341,413	-170,777	341,562	341,593
Inc. Places 1990	-173,152	346,314	346,353	-173,243	346,494	346,525
Inc. Places 2000	-176,827	353,664	353,703	-176,931	353,870	353,901
All places 2000	-234,750	469,510	469,551	-234,710	469,428	469,461
All places 2010	-262,427	524,864	524,905	-262,375	524,758	524,791

US Sample	tdPSM log-likelihood	AIC	BIC
Inc. Places 1900	-87,232	**174,478**	**174,529**
Inc. Places 1910	-116,690	**233,393**	**233,446**
Inc. Places 1920	-128,485	**256,983**	**257,037**
Inc. Places 1930	-138,060	**276,134**	**276,188**
Inc. Places 1940	-142,074	**284,162**	**284,216**
Inc. Places 1950	-147,486	**294,986**	**295,040**
Inc. Places 1960	-158,530	**317,073**	**317,128**
Inc. Places 1970	-164,680	**329,375**	**329,430**
Inc. Places 1980	-170,625	**341,265**	**341,320**
Inc. Places 1990	-173,106	**346,226**	**346,281**
Inc. Places 2000	-176,775	**353,563**	**353,618**
All places 2000	-234,633	**469,280**	**469,337**
All places 2010	-262,252	**524,518**	**524,576**

Table 11: Maximum log-likelihoods, AIC and BIC for the used distributions and the US CCA clusters data. The lowest values of AIC and BIC for each sample are marked in boldface

US Sample	ln log-likelihood	AIC	BIC	IS1 log-likelihood	AIC	BIC
CCA 1991 (2000m)	-289,460	578,923	578,940	-288,236	576,481	576,514
CCA 1991 (3000m)	-226,140	452,284	452,300	-224,434	448,876	448,908
CCA 1991 (4000m)	-192,249	384,502	384,518	-190,431	380,871	380,902
CCA 1991 (5000m)	-170,343	340,690	340,706	-168,608	337,224	337,255
CCA 2000 (2000m)	-293,311	586,627	586,643	-292,300	584,608	584,641
CCA 2000 (3000m)	-229,171	458,347	458,363	-227,733	455,474	455,507
CCA 2000 (4000m)	-194,701	389,406	389,422	-193,134	386,277	386,309
CCA 2000 (5000m)	-172,389	344,783	344,798	-170,864	341,735	341,766

US Sample	IS2 log-likelihood	AIC	BIC	dPln log-likelihood	AIC	BIC
CCA 1991 (2000m)	-288,236	576,482	576,523	-284,288	568,584	568,617
CCA 1991 (3000m)	-224,433	448,876	448,916	-221,851	443,710	443,742
CCA 1991 (4000m)	-190,431	380,872	380,912	-188,584	377,177	377,209
CCA 1991 (5000m)	-168,608	337,225	337,264	-167,096	334,201	334,232
CCA 2000 (2000m)	-292,299	584,608	584,650	-288,879	577,765	577,798
CCA 2000 (3000m)	-227,733	455,476	455,516	-225,494	450,996	451,028
CCA 2000 (4000m)	-193,134	386,279	386,318	-191,552	383,112	383,143
CCA 2000 (5000m)	-170,864	341,737	341,776	-169,586	339,179	339,211

US Sample	dm PChP log-likelihood	AIC	BIC
CCA 1991 (2000m)	-283,584	**567,186**	**567,261**
CCA 1991 (3000m)	-221,218	**442,454**	**442,526**
CCA 1991 (4000m)	-188,065	**376,148**	**376,219**
CCA 1991 (5000m)	-166,669	**333,356**	**333,426**
CCA 2000 (2000m)	-288,309	**576,635**	**576,710**
CCA 2000 (3000m)	-225,020	**450,057**	**450,130**
CCA 2000 (4000m)	-191,176	**382,370**	**382,441**
CCA 2000 (5000m)	-169,277	**338,572**	**338,642**

sample of US CCA clusters (2000, 2km.). We see that the lower tail of the first sample fits nicely (the empirical $\ln(\mathrm{cdf})$ of that of clusters is not so linear), for the upper tails the fit is quite remarkable in the two cases, and for the middle panel it is very hard to see discrepancies between the empirical and estimated density functions, compare with Figure 1.

Table 12: Percentages of population and urban units (places, clusters) in the tails and the body of the tdPSM for places and the dm PChP for clusters. For the definition of tails and body we use in each case the corresponding thresholds ϵ and τ of Table 5 for places and Table 7 for clusters

	Population			Units		
	Lower tail	Body	Upper tail	Lower tail	Body	Upper tail
Inc. Places 1900	0,3%	20,8%	78,9%	7,4%	81%	11,6%
Inc. Places 1910	0,2%	29,5%	70,3%	5,7%	89%	5,3%
Inc. Places 1920	0,2%	19,6%	80,2%	7,5%	82,2%	10,3%
Inc. Places 1930	0,3%	23,5%	76,2%	9,9%	83,7%	6,4%
Inc. Places 1940	0,2%	25,6%	74,2%	9,2%	84,8%	6%
Inc. Places 1950	0,1%	25,4%	74,5%	7,7%	86,1%	6,2%
Inc. Places 1960	0,1%	26%	73,9%	8,5%	84,9%	6,6%
Inc. Places 1970	0,1%	33,8%	66,1%	8,2%	87,5%	4,3%
Inc. Places 1980	0,1%	39,5%	60,4%	6,2%	90,2%	3,6%
Inc. Places 1990	0,1%	41,2%	58,7%	8,6%	88,2%	3,2%
Inc. Places 2000	0%	41,4%	58,6%	5,2%	91,5%	3,3%
All places 2000	0,1%	42,9%	57%	7,1%	89%	3,9%
All places 2010	0,1%	50,9%	49%	9,9%	87,7%	2,4%
CCA 1991 (2000m)	2%	53,2%	44,8%	12,3%	84,5%	3,2%
CCA 1991 (3000m)	1,8%	39,3%	58,9%	13,9%	82,2%	3,9%
CCA 1991 (4000m)	1,3%	32,7%	66%	12,3%	83,2%	4,5%
CCA 1991 (5000m)	3,1%	26,2%	70,7%	24,6%	70%	5,4%
CCA 2000 (2000m)	0,4%	56,7%	42,9%	4,7%	92,2%	3,1%
CCA 2000 (3000m)	0,3%	42,2%	57,5%	4,8%	91,3%	3,9%
CCA 2000 (4000m)	0,2%	35,1%	64,7%	3,7%	92%	4,3%
CCA 2000 (5000m)	0,3%	27,7%	72%	5%	87,6%	7,4%

We also show in Table 12 the percentages of population and urban units in the tails and the body of the selected distributions for each type of data (places and clusters). As an approximation, we classify urban units in the lower tail as those having a population less than the value of the ϵ threshold, those in the upper tail having a population greater than the τ threshold, and the body is formed by urban units with population between ϵ and τ. The values of these thresholds for places are those of Table 5 and for clusters those of Table 7. It is observed that although the percentages of population in the lower tails are generally quite low, the percentages of urban units in the lower tail are comparable to or even higher than those in the upper tail. This fact explains the need of taking into account the appropriate modeling of the lower tail in order to obtain an excellent overall fit.

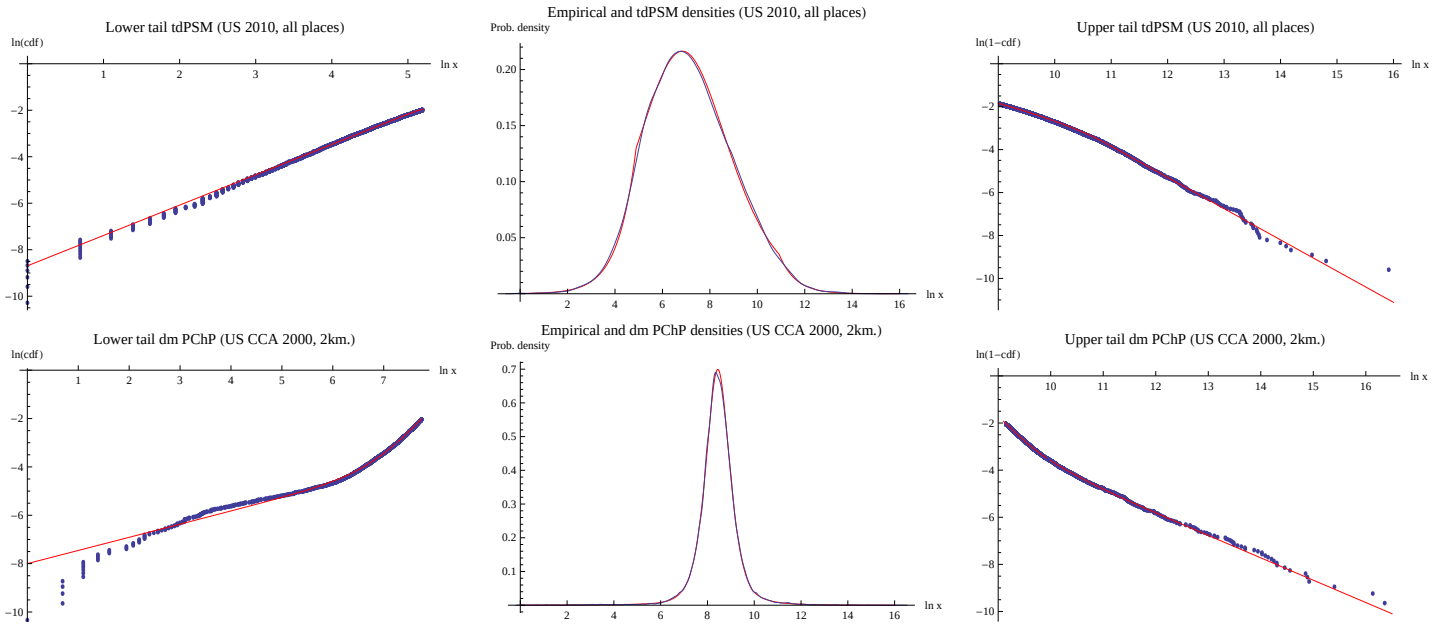

Figure 2: Left column: Empirical and estimated tdPSM and dm PChP $\ln(\text{cdf})$ for the lower tail. Center column: Empirical (Gaussian adaptive kernel density) and estimated tdPSM and dm PChP density functions. Right column: Empirical and estimated tdPSM and dm PChP $\ln(1-\text{cdf})$ for the upper tail. First row: US all places (2010). Second row: US CCA clusters (2000, 2km.). Empirical in blue, estimated in red in all cases.

6 Theoretical underpinnings

We develop in this section a theory yielding the distributions of this paper with best performance, namely the tdPSM for the US incorporated places and all places in the period (1900-2010), and the dm PChP for the US CCA clusters. We build on previous concepts used by many authors, for example Gabaix (1999, 2009) and also Reed (2002, 2003), amongst others.

Namely, consider a continuous time model in which the (natural logarithm of the) sizes X_t obeys the stochastic differential equation

$$dX_t = a(t)X_t\,dt + b(t)X_t\,dB_t \tag{24}$$

where $a(t)$, $b(t)$ are functions of time t and B_t is a Brownian motion. Such an equation is sometimes considered as an implementation of the Gibrat's law, see Gabaix (1999, 2009) and references therein. The probability density function of the variable x (in our paper, population of urban nuclei), depending also on time, namely $f(x,t)$, obeys the *forward Kolmogorov* equation, also known as *Fokker–Planck* equation, which is the partial differential equation:

$$\frac{\partial f(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left(a(t)xf(x,t)\right) + \frac{\partial^2}{\partial x^2}\left(\frac{1}{2}x^2b(t)^2f(x,t)\right) \tag{25}$$

See Payne (1967) for a concise and rather complete exposition[16].

The equation (25) has several well-known solutions, like the (time-dependent) log-normal, see, e.g., the recent work of Toda (2012) and references therein, or the upper tail Pareto distribution (with a lower threshold), see Gabaix (1999, 2009).

Now, in order to accommodate the preferred models obtained in previous sections, we should first investigate under which conditions the building blocks of such distributions, the (lower and upper tail) Pareto, the Singh–Maddala and Champernowne distributions are itself solutions of (25). We begin with the Pareto distributions.

Proposition 1 *The (time-dependent) lower tail Pareto distribution $A(t)h(x,\rho(t))$ is a solution of the equation (25) if and only if*

$$\rho'(t) = 0 \Rightarrow \rho(t) = \rho$$
$$A(t) = \exp\left(\int_0^t \frac{1}{2}\rho((1+\rho)b(s)^2 - 2a(s))\,ds\right)$$

Likewise, the (time-dependent) upper tail Pareto distribution $C(t)g(x,\zeta(t))$ is a solution of the equation (25) if and only if

$$\zeta'(t) = 0 \Rightarrow \zeta(t) = \zeta$$
$$C(t) = \exp\left(\int_0^t \frac{1}{2}\zeta((\zeta-1)b(s)^2 + 2a(s))\,ds\right)$$

[16]Strictly speaking, the $f(x,t)$ of (25) is a probability density function conditional on the initial data. We will simply take the obtained solutions of (25) evaluated at $t = 0$ as the initial conditions.

Proof. See appendix.

The second part of this last result is related to a derivation of Gabaix (1999, 2009) of the Pareto distribution as a stationary solution of (25). It is remarkable that the Pareto exponents ρ and ζ must be constants in order that the Pareto distributions be solutions of the equation (25). Note as well that these two Pareto distributions satisfy (25) also in the case of having $b(t) = 0$. We continue next with the Singh–Maddala distribution.

Proposition 2 *The (time-dependent) Singh–Maddala distribution* $D(t)f_{\mathrm{SM}}(x, \mu(t), \sigma(t), \alpha(t))$ *is a solution of the equation (25) if*

$$\begin{aligned}
&\mu'(t) = a(t) \Rightarrow \mu(t) = \int_0^t a(s)\, ds \\
&\sigma'(t) = 0 \Rightarrow \sigma(t) = \sigma \\
&\alpha'(t) = 0 \Rightarrow \alpha(t) = \alpha \\
&D'(t) = 0 \Rightarrow D(t) = D \\
&b(t) = 0
\end{aligned}$$

Proof. See appendix.

Also, the result for the Champernowne distribution is similar:

Proposition 3 *The (time-dependent) Champernowne distribution* $E(t)f_{\mathrm{Ch}}(x, \mu(t), \sigma(t), \beta(t))$ *is a solution of the equation (25) if*

$$\begin{aligned}
&\mu'(t) = a(t) \Rightarrow \mu(t) = \int_0^t a(s)\, ds \\
&\sigma'(t) = 0 \Rightarrow \sigma(t) = \sigma \\
&\beta'(t) = 0 \Rightarrow \beta(t) = \beta \\
&E'(t) = 0 \Rightarrow E(t) = E \\
&b(t) = 0
\end{aligned}$$

Proof. See appendix.

The main novelty of these last two results is that a necessary condition for the (time-dependent) Singh–Maddala and Champernowne density functions to be *always* a solution of the equation (25) is that $b(t) = 0$, namely, the *diffusion* term in (25) vanishes and also the stochastic term in (24) vanishes. We will comment on the economic meaning of such a requirement later.

Because of the importance of the selected models obtained in previous sections, it is worth studying the case of $b(t) = 0$ in more detail. We have that in such a case the equation (25) reduces to

$$\frac{\partial f(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left(a(t)xf(x,t)\right) \tag{26}$$

which can be written as

$$\frac{\partial f(x,t)}{\partial t} + a(t)x\frac{\partial f(x,t)}{\partial x} = -a(t)f(x,t) \tag{27}$$

namely, a first-order linear partial differential equation in two variables, tractable with standard methods. We have the following result:

Proposition 4 *The general solution of the equations (26) or (27) can be expressed as*

$$f(x,t) = \frac{1}{x}\, j\left(\ln x - \int_0^t a(s)\, ds\right)$$

where $j(\cdot)$ is an arbitrary function of its argument (positive and differentiable almost everywhere).

Proof. See appendix.

This last result shows that the probability density functions which satisfy the equation (26) are inversely proportional to x, with a multiplying function which depends on x and t only through the combination $\ln x - \int_0^t a(s)\, ds$. Such a simple result is essential in what follows, since our preferred models will fit into such a framework.

Corresponding to the selected distribution for US incorporated places and all places, the tdPSM, we have the following result:

Theorem 1 *The time-dependent function associated to the tdPSM*

$$f_{4t}(x,t) = \begin{cases} b_4(t)\, e_4(t)\, h(x,\rho(t)) & 0 < x < \epsilon(t) \\ b_4(t)\, f_{\mathrm{SM}}(x,\mu(t),\sigma(t),\alpha(t)) & \epsilon(t) \le x \le \tau(t) \\ b_4(t)\, a_4(t)\, g(x,\zeta(t)) & \tau(t) < x \end{cases} \tag{28}$$

is a solution of the equation (26) if and only if the following conditions hold:

$$\begin{aligned} &\mu(t) = \int_0^t a(s)\, ds\,, \quad \sigma(t) = \text{const.} \\ &b_4(t) = \text{const.}\,, \quad \alpha(t) = \text{const.} \\ &\mathrm{e}^{-\mu(t)}\epsilon(t) = \text{const.}\,, \quad \mathrm{e}^{-\mu(t)}\tau(t) = \text{const.} \\ &\rho(t) = \text{const.}\,, \quad e_4(t)\mathrm{e}^{\rho(t)\mu(t)} = \text{const.} \\ &\zeta(t) = \text{const.}\,, \quad a_4(t)\mathrm{e}^{-\zeta(t)\mu(t)} = \text{const.} \end{aligned}$$

Proof. See appendix.

Likewise, correspondingly to the selected model in the case of US CCA clusters, the dm PChP, we have the following result:

Theorem 2 *The time-dependent function associated to the dm PChP*

$$
\begin{aligned}
&f_{5t}(x,t) \\
&\quad = \begin{cases} b_5(t)\ [(1-\nu(t))\, d_5(t)\, f_{\mathrm{Ch}}(x,\mu(t),\sigma(t),\beta(t)) + \nu(t)\, e_5(t)\, h(x,\rho(t))] & 0 < x < \epsilon(t) \\ b_5(t)\, f_{\mathrm{Ch}}(x,\mu(t),\sigma(t),\beta(t)) & \epsilon(t) \le x \le \tau(t) \\ b_5(t)\ [(1-\theta(t))\, c_5(t)\, f_{\mathrm{Ch}}(x,\mu(t),\sigma(t),\beta(t)) + \theta(t)\, a_5(t)\, g(x,\zeta(t))] & \tau(t) < x \end{cases}
\end{aligned}
\tag{29}
$$

is a solution of the equation (26) if and only if the following conditions hold:

$$
\begin{aligned}
&\mu(t) = \int_0^t a(s)\, ds\,, \quad \sigma(t) = \text{const.} \\
&b_5(t) = \text{const.}\,, \quad \beta(t) = \text{const.} \\
&\mathrm{e}^{-\mu(t)}\epsilon(t) = \text{const.}\,, \quad \mathrm{e}^{-\mu(t)}\tau(t) = \text{const.} \\
&(1-\nu(t))d_5(t) = \text{const.}\,, \quad (1-\theta(t))c_5(t) = \text{const.} \\
&\rho(t) = \text{const.}\,, \quad \nu(t)e_5(t)\mathrm{e}^{\rho(t)\mu(t)} = \text{const.} \\
&\zeta(t) = \text{const.}\,, \quad \theta(t)a_5(t)\mathrm{e}^{-\zeta(t)\mu(t)} = \text{const.}
\end{aligned}
$$

Proof. See appendix.

Thus, our preferred models are able to satisfy equation (26) provided the relation $\mu(t) = \int_0^t a(s)\, ds$ holds and some other quantities remain constant. The parameter $\sigma(t)$ is a constant as well as $b_4(t)$ or $b_5(t)$. These results could be anticipated from our preliminary study of the Singh–Maddala or Champernowne distributions as a solution of (25). Also, it is also predicted that the Pareto exponents $\rho(t)$, $\zeta(t)$ remain constant (the individual Pareto distributions yielded the same results). And there are other constants arising from the fact of having a composition/mixing of the distributions. The most remarkable are those relating the threshold parameters $\mathrm{e}^{-\mu(t)}\epsilon(t) = \text{const.}$ and $\mathrm{e}^{-\mu(t)}\tau(t) = \text{const.}$ It is worth noting that this theory does not predict the precise value of the Pareto exponents ρ, ζ, only that they remain constant. To predict the value of ζ, other approaches (using as well a version of equation (25)) exist (Gabaix, 1999, 2009), so our theory can be regarded as complementary to the cited references.

As an informal test on how well our theory works, we have computed the values of the presumed constants for the empirical results corresponding to the samples of US incorporated and all places in the period (1900-2010) and that of US CCA clusters, using the estimated parameters by ML and the expressions (15), (16), (17) of the constants (constants in the sense of Section 4) e_4, a_4, b_4 in the first case and (19), (20), (21), (22), (23), of d_5, e_5, c_5, a_5, b_5 in the second case. The results are shown in Tables 13 and 14.

For the US incorporated and all places, we see that σ increases, even quite slowly, so one of the basic assumptions of our theory, the absence of diffusion, is not exactly satisfied. There exists diffusion, although very small in the short term (say, one decade). The quantity b_4 remains in the interval $(1.04, 1.13)$. The parameter α is in the interval

$(0.28, 1)$. The lower tail Pareto exponent ρ decreases slowly with time from 2.32 in 1900 to 1.31 in 2010. Likewise, the upper tail Pareto exponent ζ increases slowly from 1.02 in 1900 to 1.45 in 2010. Both variations are due to the effective existence of diffusion in practice. The quantity $\mathrm{e}^{-\mu}\epsilon$ varies more, in the interval $(0.14, 0.71)$. The analogous relation for the upper tail threshold τ leads to a strong variation of the presumed "constant". It is to be remarked that the number of places in these samples increases greatly with time, cf. Table 1.

In turn, for the US CCA clusters, the variations are in general smaller in all cases but we have to take into account that only a nine-year period is studied with these data. Also, for these data the number of observations is the same for each pair of samples of 1991 and 2000.

In short, the results suggest that when the short term is considered (say, one decade), the theory works well, and if the number of observations is constant, slightly better. In the long term, and if the number of observations varies along time, the theory shows its limitations.

Table 13: Values of the quantities obtained in Theorem 1 for the US incorporated and all places samples corresponding to the tdPSM

US Sample	σ	b_4	α	$\mathrm{e}^{-\mu}\epsilon$	$\mathrm{e}^{-\mu}\tau$	ρ	$e_4\mathrm{e}^{\rho\mu}$	ζ	$a_4\mathrm{e}^{-\zeta\mu}$
Inc. Places 1900	0.42	1.05	0.32	0.61	12.06	2.32	0.52	1.02	1.44
Inc. Places 1910	0.44	1.04	0.34	0.53	29.68	2.48	0.66	1.09	2.21
Inc. Places 1920	0.45	1.06	0.33	0.62	15.93	2.36	0.53	0.98	1.45
Inc. Places 1930	0.45	1.05	0.31	0.71	33.74	2.05	0.39	1	2.06
Inc. Places 1940	0.44	1.05	0.28	0.7	41.14	2.01	0.35	1.06	3.14
Inc. Places 1950	0.55	1.06	0.34	0.54	42.6	1.89	0.43	1.06	3.22
Inc. Places 1960	0.61	1.09	0.32	0.57	54.01	1.71	0.35	1.07	4.55
Inc. Places 1970	0.69	1.08	0.38	0.47	86.13	1.6	0.42	1.19	9.44
Inc. Places 1980	0.69	1.07	0.38	0.37	99.13	1.69	0.52	1.3	17.04
Inc. Places 1990	0.85	1.09	0.48	0.38	113.39	1.51	0.52	1.31	19.32
Inc. Places 2000	0.79	1.08	0.4	0.28	132.98	1.6	0.61	1.35	30.08
All Places 2000	1.14	1.10	0.82	0.14	40.27	1.46	1.67	1.33	6.48
All Places 2010	1.31	1.13	1	0.14	59.42	1.31	1.45	1.45	11.8

7 Discussion

We have just seen that two newly introduced here density functions perform better than some of the previously known: The lognormal used by Eeckhout (2004) and others, the IS1 and IS2 of Ioannides and Skouras (2013); the dPln of Reed (2002, 2003); Giesen et al. (2010) and others, when fitting US city data. Specifically, the tdPSM is the preferred model for US incorporated and all places data and the dm PChP is the preferred density function for the US CCA clusters of Rozenfeld et al. (2008, 2011). We have developed as well a theory that can generate the cited preferred models, and when comparing to the empirical results, it follows that when the short-term is considered (one decade) and the number of observations (urban centers) is almost constant, the theory

Table 14: Values of the quantities obtained in Theorem 2 for the US CCA clusters samples and the dm PChP

US Sample	σ	b_5	β	$\mathrm{e}^{-\mu}\epsilon$	$\mathrm{e}^{-\mu}\tau$	$(1-\nu)d_5$	$(1-\theta)c_5$	ρ	$\nu e_5 \mathrm{e}^{\rho\mu}$	ζ	$\theta a_5 \mathrm{e}^{-\zeta\mu}$
CCA 1991 (2000m)	0.37	0.96	1.29	0.5	4.07	0.94	0.44	0.59	0.03	0.96	0.1
CCA 2000 (2000m)	0.39	0.97	1.13	0.3	4.39	0.89	0.55	0.54	0.02	0.95	0.08
CCA 1991 (3000m)	0.37	0.96	1.31	0.53	4.16	0.94	0.25	0.63	0.03	0.87	0.11
CCA 2000 (3000m)	0.4	0.97	1.21	0.3	4.47	0.9	0.36	0.56	0.02	0.87	0.11
CCA 1991 (4000m)	0.39	0.96	1.45	0.5	4.16	0.95	0.2	0.63	0.02	0.83	0.12
CCA 2000 (4000m)	0.42	0.97	1.36	0.26	4.63	0.89	0.26	0.57	0.02	0.84	0.12
CCA 1991 (5000m)	0.42	0.95	1.62	0.68	4.03	0.98	0.2	0.57	0.02	0.83	0.14
CCA 2000 (5000m)	0.42	0.96	1.26	0.3	3.35	0.92	0.55	0.58	0.02	0.79	0.11

is rather reasonable. However, in the long term (say, one century) and with varying number of observations, the theory shows its limitations.

The basic assumption of our theory in the previous section is that the stochastic term be zero, or at least negligible. Otherwise, we cannot assure that the Singh–Maddala/Champernowne (part of the) distribution be an exact solution of the Fokker–Planck equation. The economic meaning of this outcome is clear: The population and hierarchical structure of cities is very stable in time, at least in the short term (Black and Henderson, 1999; Kim, 2000; Beeson et al., 2001). And this stability or persistence is even corroborated when the cities suffer strong temporal shocks, like the US Civil War (Sanso-Navarro et al., 2013), the IIWW atomic bombing in Japan (Davis and Weinstein, 2002), the IIWW bombing in Germany (Brakman et al., 2004; Bosker et al., 2008), the US bombing in Vietnam (Miguel and Roland, 2011), or the urban terrorism (Glaeser and Shapiro, 2002). This is the interpretation associated to the theoretical condition that the diffusion term needs to be zero in the Fokker–Planck equation to guarantee that the tdPSM and the dm PChP are exact solutions of that equation.

In the long term it is also shown that things become different, and another (perhaps more general) theory should be adopted, for which we provide some ideas below. In the extreme long-term situation, we have the contribution of Batty (2006), which defends that the changes in the internal hierarchy of cities can be very important, although the aggregate distribution appears to be quite stable. This is not incompatible with the short term persistence literature, because Batty's temporal horizon is very large (world data from 430 BC.).

As mentioned, the population evolves in the long term in such a way that our theory does not work so well (for US places it is observed, which is our long term database; for US CCA clusters we do not have enough data samples). Since the hypothesized model of the city size distribution (for US places) can be taken robustly in the whole period (1900-2010) as the tdPSM, we conjecture that it is the evolution equation (26) the one that should be reconsidered. We think of three main variations:

- It is to be added the term $-k(t)f(x,t)$ (or other terms) to the right hand side of (26) in order to model the entry of new urban centers in the sample (Gabaix, 2009, 1999). The specification of $k(t)$ (or of the alternative terms) seems to be delicate. Perhaps the previous work on the distribution of entrant cities (González-Val, 2010; Giesen and Suedekum, 2013) may help in this task.

- The equation to be used is (25) with $b(t) \neq 0$. Then, we cannot assure that the distribution tdPSM be an exact solution of such an equation. We would enter in the realm of approximate solutions, see, e.g., Grasman and van Herwaarden (1999). Additionally, this could be combined as well with the extension exposed in the first item of this list.

- The equation to be used is not (25), but possibly a non-linear Fokker–Planck equation, see, e.g., Frank (1991). This approach seems to be more difficult as one should find a nonlinear Fokker–Planck equation allowing a composite of two Pareto and Singh–Maddala distributions as a solution, and moreover yielding a better agreement with empirical results than the theory exposed here. It would

be a theoretical treasure if the cited equation does exist.

We leave these topics for future research.

8 Conclusions

Elsewhere, since the work of Eeckhout (2004) the risks have been demonstrated of considering only the largest cities; that is, only the upper tail. One of the main lessons of such work is that, when possible, one should use city data without minimum size restrictions.[17] In turn, if the availability of data allows it, the analysis of city size distribution should be done as a long-term analysis. With both considerations as premises, this article uses US Census data for the period (1900-2010), in decades, and all the incorporated/all places. Also, we use the US City Clustering Algorithm (CCA) clusters data of Rozenfeld et al. (2008, 2011) for the years 1991 and 2000 and radii of the clusters of 2, 3, 4 and 5 km.

This work has minutely examined seven density functions: Lognormal, IS1, IS2 and dPln, known in the field of urban economics, and we have thereby explicitly introduced in Section 4 two new density functions, which we call tdPSM and dm PChP, for which the essential point is the modeling of *both* tails as a Pareto distribution with or without mixing with the Singh–Maddala or Champernowne distributions.

These two new distributions are associated to two "philosophical" principles:

i) For the US it seems to be necessary to pay attention to the lower tail of the distribution, despite of representing a small percentage of the population, in order to obtain an excellent overall fit. In a nutshell, *small nuclei do matter*.

ii) The body of the distribution is better described by a Singh–Maddala or Champernowne distribution rather than a lognormal. This constitutes a relevant difference to the evidence accumulated so far.

After estimating the parameters of all of the distributions by maximum likelihood (ML), we have tested the fit provided by each distribution using the Kolmogorov–Smirnov (KS) and Cramér-von Mises (CM) tests. Afterwards, we have computed the AIC and BIC information criteria.

The results are extremely robust and regular. The two new density functions improve notably the performance of the lognormal, IS1, IS2 and dPln. In particular, the tdPSM is a distribution not rejected 100% of the cases by both of the KS and CM, and is the selected model (out of the six distributions studied) by both AIC and BIC for the whole period (1900-2010) of samples of US incorporated and all places. Likewise, the

[17] In this work we have not shown the results corresponding to the data of the so called Metropolitan and Micropolitan areas (MMA), see, e.g., Ioannides and Skouras (2013) for their definition, because in them it is imposed a not small minimum threshold size (about 13,000 inhabitants). We simply mention that the KS and CM tests for a truncated version of all of the distributions used in this paper yield rejection, even having that the sample sizes of MMA data are much lower than for US places or CCA clusters (less than 1,000 observations). This means that the modeling of the MMA size distribution is much more demanding than for the US places or CCA clusters, possibly due to the cut-off imposed to such data.

dm PChP is a distribution not rejected 100% of the instances of CCA clusters by both KS and CM tests, being the selected model for all these samples by both AIC and BIC.

In short, we find empirically that the US city size distribution for places can be safely taken as a *Singh–Maddala* body with pure Pareto tails, the three regions separated by two exact thresholds. For US CCA clusters, an analogous situation occurs but where the body is Champernowne and in the tails it is advantageous *to mix* the Pareto distributions with the Champernowne one. Moreover, we have given a theoretical support for these distributions, a theory which works reasonably well in the short term and when the number of cities is constant. We have provided some ideas for the search of a theory that would be satisfactory also for the long term and varying number of urban nuclei.

A Proofs of Section 6

Proof of Proposition 1. Inserting $f(x,t) = A(t)h(x,\rho(t))$ into (25) written in the following way

$$\frac{\partial f(x,t)}{\partial t} + \frac{\partial}{\partial x}\left(a(t)xf(x,t)\right) - \frac{\partial^2}{\partial x^2}\left(\frac{1}{2}x^2b(t)^2f(x,t)\right) = 0$$

yields

$$\left(a(t)\rho(t) + \frac{A'(t)}{A(t)} - \frac{1}{2}b(t)^2\rho(t)(\rho(t)+1) + \ln(x)\rho'(t)\right)A(t)h(x,\rho(t)) = 0$$

Thus, the expression in the big left parentheses has to be zero. But firstly, the only dependence on x appears in one term with $\ln x$. In order to the equation to be consistent, it should happen that $\rho'(t) = 0 \Rightarrow \rho(t) = \rho$. Imposing this condition it follows

$$a(t)\rho + \frac{A'(t)}{A(t)} - \frac{1}{2}b(t)^2\rho(\rho+1) = 0$$

which is a simple differential equation for $A(t)$. Integrating, the thesis follows. The analysis for $f(x,t) = C(t)g(x,\zeta(t))$ is analogous and is omitted.

Proof of Proposition 2. It is similar to the proof of Proposition 1, but the expressions that appear are very long, so for the sake of brevity we will omit them. We have performed the calculations with the program MATHEMATICA: A notebook file is avalaible from the authors upon request.

Proof of Proposition 3. Again, the procedure is analogous to that of Propositions 1 and 2. The expressions which appear are very long and for the sake of brevity we will omit them. A MATHEMATICA notebook with the calculations is available from the authors upon request.

Proof of Proposition 4. It is an application of standard results, see Theorem 2.5.1 and Example 2.5.1 of Myint-U and Debnath (2007). According to this reference, the equation

$$\frac{\partial f(x,t)}{\partial t} + a(t)x\frac{\partial f(x,t)}{\partial x} = -a(t)f(x,t)$$

has the associated *characteristic equations* (*loc. cit.*)

$$\frac{dt}{1} = \frac{dx}{a(t)x} = \frac{df}{-a(t)f} \tag{30}$$

Equating the first and second members of (30) we have

$$dt = \frac{dx}{a(t)x} \Leftrightarrow a(t)dt = \frac{dx}{x}$$

and integrating we have that $C_1 = \ln x - \int_0^t a(s)\,ds$ is the first associated family of characteristic curves of the system, where C_1 is a constant. Equating the second and third members of (30), we have

$$\frac{dx}{a(t)x} = \frac{df}{-a(t)f} \Leftrightarrow \frac{dx}{x} = -\frac{df}{f}$$

and therefore the second family of characteristic curves is $C_3 = \mathrm{e}^{C_2} = xf$, where C_2 is a constant and C_3 is its exponential. As $x > 0$ it follows that $f > 0$ as well, something that is necessary for a probability density function. Thus, the general solution of the equation is expressed as an arbitrary function k of the expressions of C_1, C_3 equated to zero:

$$k\left(\ln x - \int_0^t a(s)\,ds, xf\right) = 0$$

and therefore, solving for f (*loc. cit.*),

$$f(x,t) = \frac{1}{x} j\left(\ln x - \int_0^t a(s)\,ds\right)$$

where j is an arbitrary function of its argument (positive and differentiable almost everywhere).

Proof of Theorem 1. The result is achieved writing the function f_{4t} as follows:

$$\begin{aligned} f_{4t}(x,t) = {} & b_4(t)(1 - \mathrm{H}(x - \epsilon(t)))e_4(t)\,h(x,\rho(t)) + \\ & b_4(t)\mathrm{H}(x - \epsilon(t))(1 - \mathrm{H}(x - \tau(t)))f_{\mathrm{SM}}(x,\mu(t),\sigma(t),\alpha(t)) + \\ & b_4(t)\mathrm{H}(x - \tau(t))\,a_4(t)\,g(x,\zeta(t)) \end{aligned}$$

where $\mathrm{H}(y)$ is the Heaviside step function. We apply then the Proposition 4 directly. First, we deal with the arguments of the Heaviside functions. We have

$$x - \epsilon(t) = \mathrm{e}^{\ln x - \mu(t)}\mathrm{e}^{\mu(t)} - \epsilon(t) = \mathrm{e}^{\mu(t)}(\mathrm{e}^{\ln x - \mu(t)} - \mathrm{e}^{-\mu(t)}\epsilon(t))$$

Thus,

$$\mathrm{H}(x - \epsilon(t)) = \mathrm{H}(\mathrm{e}^{\ln x - \mu(t)} - \mathrm{e}^{-\mu(t)}\epsilon(t))$$

because $\mathrm{e}^{\mu(t)} > 0$ and the Heaviside function depends only on the sign of its argument. Then, we see that this function is of the form $j\left(\ln x - \int_0^t a(s)\,ds\right)$[18] if we choose

[18]The $1/x$ factor is included in the distributions that accompany the Heaviside functions. Also, the Heaviside function is discontinuous at only *one point*. However, our composite density functions are continuous at the threshold switching points.

$\mu(t) = \int_0^t a(s)\,ds$, and it follows that $\mathrm{e}^{-\mu(t)}\epsilon(t) = \mathrm{const}$. An analogous reasoning for the Heaviside function with $\tau(t)$ yields $\mathrm{e}^{-\mu(t)}\tau(t) = \mathrm{const}$. We move then to the f_{SM} term. From the definition (2) we see immediately that $b_4(t) f_{\mathrm{SM}}(x, \mu(t), \sigma(t), \alpha(t))$ is of the form $\frac{1}{x} j \left(\ln x - \int_0^t a(s)\,ds \right)$ choosing (consistently) $\mu(t) = \int_0^t a(s)\,ds$, and it is necessary that $\sigma(t) = \mathrm{const.}$, $\alpha(t) = \mathrm{const.}$ and $b_4(t) = \mathrm{const}$. Now, we analyze the lower tail term. Letting aside the b_4 factor, which as we have seen must be constant, we have

$$e_4(t)h(x, \rho(t)) = e_4(t)\frac{1}{x}x^{\rho(t)} = e_4(t)\frac{1}{x}\mathrm{e}^{\rho(t)(\ln x - \mu(t))}\mathrm{e}^{\rho(t)\mu(t)}$$

thus, in order to have again a function of the form $\frac{1}{x} j \left(\ln x - \int_0^t a(s)\,ds \right)$ it is necessary that $\mu(t) = \int_0^t a(s)\,ds$, $\rho(t) = \mathrm{const.}$ and $e_4(t)\mathrm{e}^{\rho(t)\mu(t)} = \mathrm{const}$. The reasoning for the upper tail part is analogous, yielding $\zeta(t) = \mathrm{const.}$ and $a_4(t)\mathrm{e}^{-\zeta(t)\mu(t)} = \mathrm{const}$.

Proof of Theorem 2. The result is obtained in a similar way as in the proof of Theorem 1.

Acknowledgements

This work is supported by Spanish Ministry of Economy and Competitiveness, project ECO2009-09332 and by Aragon Government, ADETRE Consolidated Group.

References

Anderson, G. and Ge, Y. (2005). The size distribution of Chinese cities. *Regional Science and Urban Economics*, 35(6):756–776.

Batty, M. (2006). Rank clocks. *Nature*, 444(7119):592–596.

Bee, M. (2012). Statistical analysis of the lognormal-Pareto distribution using probability weighted moments and maximum likelihood. Technical report, Department of Economics, University of Trento, Italia.

Beeson, P., DeJong, D., and Troesken, W. (2001). Population growth in US counties, 1840–1990. *Regional Science and Urban Economics*, 31(6):669–699.

Black, D. and Henderson, V. (1999). Spatial evolution of population and industry in the United States. *American Economic Review*, 89(2):321–327.

Black, D. and Henderson, V. (2003). Urban evolution in the USA. *Journal of Economic Geography*, 3(4):343–372.

Bosker, M., Brakman, S., Garretsen, H., and Schramm, M. (2008). A century of shocks: The evolution of the German city size distribution 1925-1999. *Regional Science and Urban Economics*, 38(4):330–347.

Brakman, S., Garretsen, H., and Schramm, M. (2004). The strategic bombing of cities in Germany in World War II and its impact on city growth. *Journal of Economic Geography*, 4:201–218.

Burnham, K. and Anderson, D. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer-Verlag.

Burnham, K. and Anderson, D. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33:261–304.

Burr, I. (1942). Cumulative frequency functions. *The Annals of Mathematical Statistics*, 13:215–232.

Champernowne, D. (1952). The graduation of income distributions. *Econometrica*, 20(4):591–615.

Cheshire, P. (1999). Trends in sizes and structure of urban areas. In Cheshire, P. and Mills, E., editors, *Handbook of Regional and Urban Economics*, volume 3, chapter 35. Elsevier, Amsterdam.

Combes, P., Duranton, G., Gobillon, L., Puga, D., and Roux, S. (2012). The productivity advantages of large cities: Distinghishing agglomeration from firm selection. *Econometrica*, 80:2543–2594.

Cooray, K. and Ananda, M. (2005). Modeling actuarial data with a composite lognormal-Pareto model. *Scandinavian Actuarial Journal*, 2005:321–334.

Davis, D. and Weinstein, D. (2002). Bones, bombs and break points: The geography of economic activity. *American Economic Review*, 92:1269–1289.

Eeckhout, J. (2004). Gibrat's law for (all) cities. *American Economic Review*, 94(5):1429–1451.

Eeckhout, J. (2009). Gibrat's law for (all) cities: Reply. *American Economic Review*, 99:1676–1683.

Fisk, P. (1961). The graduation of income distributions. *Econometrica*, 29:171–185.

Frank, T. (1991). *Nonlinear Fokker–Planck equations*. Springer.

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

Gabaix, X. (2009). Power laws in Economics and finance. *Annu. Rev. Econ.*, 2009:255–293.

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

Gabaix, X. and Ioannides, Y. (2004). The evolution of city size distributions. In Henderson, V. and Thisse, J. F., editors, *Handbook of Regional and Urban Economics*, volume 4, chapter 53, pages 2341–2378. Elsevier.

Giesen, K. and Suedekum, J. (2012). The French overall city size distribution. *Région et Développement*, 36:107–126.

Giesen, K. and Suedekum, J. (2013). City age and city size. Conference paper, ECONSTOR.

Giesen, K., Zimmermann, A., and Suedekum, J. (2010). The size distribution across all cities-double Pareto lognormal strikes. *Journal of Urban Economics*, 68(2):129–137.

Glaeser, E. and Shapiro, J. (2002). Cities and warfare: The impact of terrorism on urban form. *Journal of Urban Economics*, 51(2):205–224.

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

González-Val, R., Ramos, A., and Sanz-Gracia, F. (2013a). The accuracy of graphs to describe size distributions. *Applied Economics Letters*, 20(17):1580–1585.

González-Val, R., Ramos, A., Sanz-Gracia, F., and Vera-Cabello, M. (2013b). Size distribution for all cities: Which one is best? *Papers in Regional Science*, Forthcoming. doi:10.1111/pirs.12037.

Grasman, J. and van Herwaarden, O. (1999). *Asymptotic methods for the Fokker-Planck equation and the exit problem in applications*. Springer.

Ioannides, Y. and Overman, H. (2003). Zipf's law for cities: An empirical examination. *Regional Science and Urban Economics*, 33(2):127–137.

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

Kim, S. (2000). Urban development in the United States, 1690-1990. *Southern Economic Journal*, 66(4):855–880.

Kleiber, C. and Kotz, S. (2003). *Statistical size distributions in Economics and actuarial sciences*. Wiley-Interscience.

Levy, M. (2009). Gibrat's law for (all) cities: Comment. *American Economic Review*, 99:1672–1675.

Malevergne, Y., Pisarenko, V., and Sornette, D. (2011). Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities. *Physical Review E*, 83:1–11.

Miguel, E. and Roland, G. (2011). The long-run impact of bombing Vietnam. *Journal of Development Economics*, 96:1–15.

Myint-U, T. and Debnath, L. (2007). *Linear partial differential equations for scientists and engineers*. Birkhäuser.

Pareto, V. (1896). *Cours d'economie politique*. Geneva: Droz.

Parr, J. and Suzuki, K. (1973). Settlement populations and the lognormal distribution. *Urban Studies*, 10(3):335–352.

Payne, H. (1967). *The response of nonlinear systems to stochastic excitation*. PhD thesis, California Institute of Technology, Pasadena, California.

Razali, N. and Wah, Y. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2:21–33.

Reed, W. (2001). The Pareto, Zipf and other power laws. *Economics Letters*, 74:15–19.

Reed, W. (2002). On the rank-size distribution for human settlements. *Journal of Regional Science*, 42:1–17.

Reed, W. (2003). The Pareto law of incomes–an explanation and an extension. *Physica A*, 319:469–486.

Reed, W. and Jorgensen, M. (2004). The double Pareto-lognormal distribution–a new parametric model for size distributions. *Communications in Statistics-Theory and Methods*, 33(8):1733–1753.

Rozenfeld, H., Rybski, D., Andrade, J., Batty, M., Stanley, H., and Makse, H. (2008). Laws of population growth. *Proceedings of the National Academy of Sciences*, 105(48):18702–18707.

Rozenfeld, H., Rybski, D., Gabaix, X., and Makse, H. (2011). The area and population of cities: new insights from a different perspective on cities. *American Economic Review*, 101:2205–2225.

Sanso-Navarro, M., Sanz-Gracia, F., and Vera-Cabello, M. (2013). The impact of the American Civil War on city growth. Mimeo.

Scollnik, D. (2007). On composite lognormal-Pareto models. *Scandinavian Actuarial Journal*, 2007:20–33.

Singh, S. and Maddala, G. (1976). A function for size distribution of incomes. *Econometrica*, 44(5):963–970.

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

Toda, A. (2012). The double power law in income distribution: Explanations and evidence. *Journal of Economic Behavior & Organization*, 84:364–381.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, Massachusetts: Addison-Wesley Press.