US city size distribution revisited: Theory and empirical evidence

Ramos, Arturo and Sanz-Gracia, Fernando

Universidad de Zaragoza, Universidad de Zaragoza

30 April 2015

US city size distribution revisited: Theory and empirical evidence

ARTURO RAMOS[a]   FERNANDO SANZ-GRACIA[b]

April 30, 2015

Abstract

We develop a urban economic model in which agents locate in cities of different size so as to maximize the net output of the whole system of cities in a country. From this model two new city size distributions are exactly derived. We call these functions "threshold double Pareto Generalized Beta of the second kind" and "double mixture Pareto Generalized Beta of the second kind". In order to test empirically the theory, we analyze the US urban system and consider three types of data (incorporated places from 1900 to 2000, all places in 2000 and 2010 and City Cluster Algorithm nuclei in 1991 and 2000). The results are encouraging because the new distributions clearly outperform the lognormal and the double Pareto lognormal for all data samples. We consider a number of different tests and statistical criteria and the results are robust to all of them. Thus, the new distributions describe accurately the US city size distribution and, therefore, support the validity of the theoretical model.

[a]Department of Economic Analysis, Universidad de Zaragoza (SPAIN) aramos@unizar.es. We thank Rafael González-Val and Fernando Pueyo for useful comments on previous versions of this paper. This work is supported by the Spanish *Ministerio de Economía y Competitividad* (ECO2013-45969-P) and *Gobierno de Aragón*, ADETRE Consolidated Group.

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

Keywords: human capital; congestion costs; lower tail, body, and upper tail; Pareto and Generalized Beta of the second kind distributions

JEL: C46, D39, R11, R12

1 Introduction

Cities are complex systems, which tend to self-organize and where everything is interconnected (Krugman, 1996; Batty, 2008, 2013; Bettencourt and West, 2010; Arcaute et al., 2014). Therefore, their study can and must be addressed from different points of view. This work aims to shed some light, both from a theoretical and from an empirical perspective, to the analysis of city size distributions. The empirical part focus on the US urban system, taking data from 1900 to 2010.

A first question which should be justified is why it is relevant to study city size distributions. In other words, why it is worth devoting effort to describe, as accurately as possible, the shape of the mentioned distributions. Following Malevergne et al. (2011) we can put forward three essential arguments. One, the shape can be very informative when trying to understand the mechanisms associated to the growth generating process. Two, the specific shape of the distribution (unimodal of not, skewed or not, leptokurtic or not) has consequences on many socio-economic aspects affecting citizens in the real world. And three, the upper tail is, by definition, quantitatively important in terms of population.

According to the content of the previous paragraph, the literature on city size distributions is ample. Without pretending to be exhaustive and citing only contributions of this century, we have Overman and Ioannides (2001); Black and Henderson (2003); Ioannides and Overman (2003, 2004); Eeckhout (2004); Soo (2005, 2014); Anderson and Ge (2005); Bosker et al. (2008); Giesen et al. (2010); Holmes and Lee (2010), Berry and Okulicz-Kozaryn (2012);[1] Ioannides and Skouras (2013); González-Val et al. (2015); Luckstead and Devadoss (2014a,b) and Berliant and Watanabe (2015). It is not easy to summarize the preceding papers. Notwithstanding, we will point out the main characteristics of this body of literature.

First, the most studied geographic area is that of the United States. Secondly, the

[1] Where you can find an excellent survey of the history of city size distributions in its Section 2.

two most studied distributions, by a far amount, are the Pareto I or power law (a particular case of this is the so-called Zipf's law or, in its deterministic version, the rank-size rule) and the lognormal. Thirdly, the definition of what is considered a city is not neutral to the results obtained finally. Indeed, researchers in this field usually have to take two decisions: on the one hand, the consideration or not of a cut-off or truncation point of the population variable (and, if affirmative, of what size) and, on the other hand, the specific definition of the objects of study.[2] Fourthly, there is a certain consensus (Desmet and Rappaport (2014) take it for granted to build up their article), about that the US city size distribution is lognormal and approximately Zipf at the upper tail, or at least, Pareto; see also Levy (2009).[3] Finally, a number of recent papers argue that for an excellent fit to the data for the whole range of possible sizes, it is necessary to consider more than a single functional form, since the different parts of the distribution present different behaviors. Indeed, Giesen et al. (2010) distinguish between the tails and the body, and Ioannides and Skouras (2013) do it between the upper tail and the rest of the distribution. We will return to this idea in Sections 2 and 4.

Against this background, the present paper attempts to define a new framework to analyse city size distributions, both from a theoretical and from an empirical point of view.

The theory will be addressed in Section 4, where we develop a urban model in which the population distributes itself so as to maximize the net output of the system of cities of the country. From this model the new density functions proposed in this paper, called "treshold double Pareto Generalized Beta of the second kind", (tdPGB2), and "double mixture Pareto Generalized Beta of the second kind", (dmPGB2) can be exactly deduced.

In the empirical part of this paper, we will compare the tdPGB2 and dmPGB2 with

[2]As an illustrative example, in the case of US, Rozenfeld et al. (2011) tell us that the distribution of places or legal cities is broadly lognormal, while the distribution of geography-based agglomerations (City Clustering Algorithm nuclei or CCA nuclei or simply, clusters, hereafter), is quasi-Zipf.

[3]An exception to this consensus can be found in Bee et al. (2013).

two well-known distributions previously considered when studying city size, namely the lognormal (lgn) and the double Pareto lognormal (dPln).[4] For the case of US places, we will consider as well the Generalized Beta of the second kind (GB2) distribution that is obtained directly from an economic model by Parker (1999), on which we base our further developments of Section 4. The densities newly introduced herein systematically improve the performance of all the distributions used up to now.

The rest of the paper is structured as follows. In Section 2 we will detail the principles underlying our approach. Section 3 defines the densities that are estimated later, namely, the lognormal, the double Pareto lognormal, the Generalized Beta of the second kind and the new tdPGB2 and dmPGB2. Section 4 develops the theoretical model in which the net output of the whole urban system is maximized, yielding as a result the new distributions. Section 5 describes the data sets used in the empirical application. Section 6 gives an account of the empirical results. Finally, we give some Conclusions.

2 Motivation of our approach

This paper is based on the following principles (one in each paragraph), many of them standard in the Urban Economics literature.

From our point of view, to find statistical distributions that fit the data much better than the ones known in the literature is an interesting contribution by itself. But it is even much more interesting if these new distributions are derived from a theoretical economic model, in which functions that have clear economic implications are defined, and explicitly, agents fulfill their rule of behaviour according to the usual optimization techniques. See, in this sense, Section 4.

The study of city size distributions should be, to the possible extent, a long-term

[4] We do not specifically show a separate study for the Pareto distribution as it is encoded at the tails of the new tdPGB2 and dmPGB2. Moreover, the new densities are overwhelmingly better than a single power law for all the range of city sizes.

analysis (Parr, 1985; Gabaix and Ioannides, 2004). In particular, we use different data sets concerning the United States from 1900 to 2010, although not all results are shown for the sake of brevity.

It seems that there is no single density function capable of providing adequate description of the distribution *for all values of the city size population variable*. This is an accepted statement in the literature on income size distribution.[5] We consider applicable this idea, in line with Schluter and Trede (2013) and Bee et al. (2013), also to the study of city size distribution. Consequently, in our approach, we divide the overall distribution in three parts: the lower tail, the body, and the upper tail. Each of these three parts has a specific treatment in the economic model that we develop in Section 4.

A principle that is derived directly from the previous paragraph is that big urban nuclei (upper tail) do matter and require a special attention. This is a generally accepted fact in the Urban Economics literature (Zanette (2015), for example, clearly highlights the problems associated with the modeling of city sizes for low ranks), where even the largest cities receive a special designation ("dragon kings", "king effect", "king plus vice-roy effect") and generally are considered to be outliers with respect to the the hypothesized distribution (see, e.g., Giesen and Suedekum (2012)).

Also, in parallel with the preceding two paragraphs, we have the certainty that small nuclei (lower tail) do matter and also require an specific treatment. Contrary to what is generally accepted for the larger cities, this approach is fairly overlooked, with the possible exception of Reed (2002, 2003). Therefore, we consider, as Eeckhout (2004) does in his pioneering work, all entities of population, without any truncation point. The importance of small nuclei has already been highlighted from a theoretical point of view in the models of Blank and Solomon (2000) and Lee and Li (2013). From

[5] Explicitly, Dagum (1979) points out in the page 16, citing (Macaulay, 1922, p. 393): "it seems unlikely that any useful mathematical law describing the entire distribution can ever be formulated" or (Mandelbrot, 1960, p. 82): "The above reason makes it unlikely that a [...] single empirical formula could ever represent all the data" and (Budd, 1970, p. 250): "it is virtually impossible to describe empirical distributions accurately by just one function".

an empirical perspective, small urban nuclei are not relevant for the percentage of the population they represent, but this is not the case with regards to the total number of nuclei. We will see in Section 6 that the empirical results obtained confirm the necessity of taking into account the lower tail specifically.

The parsimony in terms of the number of parameters of the distribution to be estimated is always a goal to be pursued. This is one of the reasons for the success of power laws and Zipf's law: they fit reasonably well the data, especially at the upper tail, and computationally they are not costly.[6] Notwithstanding this, the new distributions that we propose in this paper seem not to be particularly parsimonious at first sight. But we can defend this option based on four arguments. Firstly, with the currently available computing capabilities, to estimate a moderately high number of parameters is not particularly expensive. Secondly, the information criteria used in Section 6 in order to discriminate between the studied distributions, namely the Akaike Information Criterion (AIC) and the Bayesian or Schwarz Information Criterion (BIC), explicitly penalize the number of parameters of an hypothesized distribution. The penalty functions are rooted on solid information-theoretical arguments (Burnham and Anderson, 2002, 2004). Thirdly, there already exist examples in the literature where a mere increasing of the parameters of the distribution does not always lead to a better fit in information-theoretic terms.[7] And fourthly, as we will see in the empirical application of Section 6, the descriptive power of the new densities is extremely high, so a rather complex system like a complete urban structure, with thousands of urban nuclei and millions of people involved, can be accurately described by means of a distribution with ten parameters or less. The descriptive power of the parameters is, therefore, comparatively large in per capita terms.

The results are more valid and powerful as they are robust to different alternatives.

[6]Although the determination of the optimal cut-off population value for the Hill (Hill, 1975) estimator takes a lot of calculations.

[7]See, for example, the case of Switzerland in Giesen et al. (2010), where the lognormal (two parameters) outperforms the double Pareto lognormal (four parameters), and other examples in González-Val et al. (2015) where the log-logistic (two parameters) also outperforms the dPln.

In the first place, we have used three definitions of US cities (they are not all of the existing ones, but we believe it is a number high enough): incorporated places, all places and CCA clusters of Rozenfeld et al. (2011). Secondly, we will consider a number of different criteria in order to assess the quality of the empirical fits. Indeed, we will use three different statistical tests which are very powerful for the large sample sizes at hand (Razali and Wah, 2011), and for which the non-rejections occur only if the deviations (statistics) are really small. They are the Kolmogorov–Smirnov (KS) test, Crámer–von Mises (CM) test and, as recommended by Cirillo (2013), the Anderson–Darling (AD) test. Furthermore, we reinforce the metrics with additional criteria very well suited to the study of city size distributions, namely the msd (mean squared differences) and pseudo-R^2 quantities of Duranton (2007). Also, and as already mentioned, in order to discriminate between the hypothesized theoretical distributions, we use the AIC and BIC information criteria. Finally, in the third place, one may wonder whether these results are also applicable to other geographical areas. The paper of Puente-Ajovín and Ramos (2015) carries out the empirical analysis for four major European countries: France, Germany, Italy and Spain. In all of these cases (for Italy and Spain data for more than one hundred years are employed), a particular case of one of our hypothesized distributions, the so called therein "tdPSM", is the one which offers the best performance amongst the studied ones.

We do not take into account that the city size distribution can be affected by geographical aspects. In other words, our approach is not spatial in nature. Although Behrens et al. (2014) have found that spacial frictions do not significantly influence the city size distribution, it is a potentially relevant subject which has already received attention in several recent articles (Favaro and Pumain, 2011; Rybski et al., 2013; Rauch, 2014; Hsu et al., 2014).

3 Description of the distributions used

In this Section we will introduce the distributions used along this paper.

3.1 The lognormal (lgn)

The well-known lognormal (lgn) distribution for the population of cities have been proposed in the field of Urban Economics by Parr and Suzuki (1973) and afterwards by Eeckhout (2004) when considering *all* cities. The corresponding density is simply

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

where $\mu, \sigma > 0$ are respectively the mean and the standard deviation of $\ln x$, x being the population of the urban units under study.

3.2 The double Pareto lognormal (dPln)

The second distribution in our study will be the double Pareto lognormal distribution (dPln), introduced by (Reed, 2002, 2003; Reed and Jorgensen, 2004):

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

where erf is the error function associated to the normal distribution and $\alpha, \beta, \mu, \sigma > 0$ are the four parameters of the distribution. It has the property that it approximates different power laws in each of its two tails: $f_{\mathrm{dPln}}(x) \approx x^{-\alpha-1}$ when $x \to \infty$ and $f_{\mathrm{dPln}}(x) \approx x^{\beta-1}$ when $x \to 0$, hence the name of double Pareto. The body is approximately lognormal, although it is not possible to exactly delineate the switch between the lognormal and the Pareto behaviors (Giesen et al., 2010). In this last reference it is shown that the dPln offers a good fit for a number of countries. In this line, see also the work of Giesen and Suedekum (2014) and González-Val et al. (2015).

3.3 The threshold double Pareto Generalized Beta of the second kind (tdPGB2)

We introduce here a new distribution. By construction, the tdPGB2 has a Generalized Beta of the second kind (GB2) 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 GB2 body, and $\tau > \epsilon$ separates the body from the Pareto power law in the upper tail.

The specific description is as follows. We first define the building block functions, setting

$$f_{\mathrm{GB2}}(x,a,b,p,q) = \frac{a x^{ap-1}}{b^{ap} B(p,q)\left(1+(x/b)^{a}\right)^{p+q}} \tag{1}$$

$$\mathrm{cdf}_{\mathrm{GB2}}(x,a,b,p,q) = \frac{1}{B(p,q)} B\left(\frac{(x/b)^{a}}{1+(x/b)^{a}}, p, q\right) \tag{2}$$

$$l(x,\rho) = x^{\rho-1} \tag{3}$$

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

The f_{GB2} ($\mathrm{cdf}_{\mathrm{GB2}}$) is the Generalized Beta of the second kind density (resp., cumulative distribution function, cdf) (McDonald, 1984; McDonald and Xu, 1995; Kleiber and Kotz, 2003), $B(z,p,q) = \int_0^z t^{p-1}(1-t)^{q-1}\,dt$ with $z \in [0,1]$ is the incomplete Beta function and $B(p,q) = B(1,p,q)$ is the Beta function. The four parameters a, b, p, q are positive and b is a scale parameter, and a, p, q are shape parameters. The functions $l(x,\rho)$ and $u(x,\zeta)$ will model, except for a multiplicative positive constant, the Pareto lower (l) and upper (u) tails of our distribution, where $\rho > 0$ and $\zeta > 0$ are the Pareto exponents.

We impose continuity of the composite density function at the two threshold points and overall normalization of the former to the unit. The resulting density reads

$$f_{\mathrm{tdPGB2}}(x,\rho,\epsilon,a,b,p,q,\tau,\zeta)=\begin{cases} b_4\, e_4\, l(x,\rho) & 0<x<\epsilon \\ b_4\, f_{\mathrm{GB2}}(x,a,b,p,q) & \epsilon\le x\le\tau \\ b_4\, a_4\, u(x,\zeta) & \tau<x \end{cases}$$

where the constants are given by (they are effectively constants because the functions below are evaluated at the specific values ϵ, τ and, therefore, do not depend on x)

$$e_4 = \frac{f_{\mathrm{GB2}}(\epsilon,a,b,p,q)}{l(\epsilon,\rho)} \tag{5}$$

$$a_4 = \frac{f_{\mathrm{GB2}}(\tau,a,b,p,q)}{u(\tau,\zeta)} \tag{6}$$

$$b_4^{-1} = e_4\frac{\epsilon^\rho}{\rho} + \mathrm{cdf}_{\mathrm{GB2}}(\tau,a,b,p,q) - \mathrm{cdf}_{\mathrm{GB2}}(\epsilon,a,b,p,q) + \frac{a_4}{\zeta\,\tau^\zeta} \tag{7}$$

This distribution depends on eight parameters $(\rho,\epsilon,a,b,p,q,\tau,\zeta)$ to be estimated.

3.4 The double mixture Pareto Generalized Beta of the second kind (dmPGB2)

The last distribution, also new, that we will consider in this paper is a variant of the previous tdPGB2. Now, the Pareto parts at the two tails are mixed (by means of convex linear combinations, see, e.g., Combes et al. (2012)) with the GB2 density, and the body is exclusively GB2. The thresholds ϵ and τ retain their roles and meanings.

We also impose continuity of the composite density function at the two threshold points and overall normalization of the former to the unit, as well as two weighting conditions for the components of the mixing at the tails (Ioannides and Skouras, 2013).

The resulting density reads

$$
\begin{aligned}
& f_{\mathrm{dmPGB2}}(x, \rho, \epsilon, \nu, a, b, p, q, \tau, \zeta, \theta) \\
& = \begin{cases}
b_5 [(1-\nu)\, d_5\, f_{\mathrm{GB2}}(x, a, b, p, q) + \nu\, e_5\, l(x, \rho)] & 0 < x < \epsilon \\
b_5\, f_{\mathrm{GB2}}(x, a, b, p, q) & \epsilon \leq x \leq \tau \\
b_5 [(1-\theta)\, c_5\, f_{\mathrm{GB2}}(x, a, b, p, q) + \theta\, a_5\, u(x, \zeta)] & \tau < x
\end{cases}
\end{aligned}
$$

In order to maintain good properties of the log-likelihood estimators we will only consider the case for which $\nu \in (0, 1)$ and $\theta \in (0, 1)$, that are new parameters with respect to the tdPGB2, representing the relative weight of the Pareto component in, respectively, the lower and the upper tail.

Now, the constants above are given by

$$
\begin{aligned}
d_5^{-1} &= 1 - \nu + \frac{\epsilon^{-\rho}\, \nu\, \rho\, \mathrm{cdf}_{\mathrm{GB2}}(\epsilon, a, b, p, q)\, l(\epsilon, \rho)}{f_{\mathrm{GB2}}(\epsilon, a, b, p, q)} \\
e_5^{-1} &= \frac{(1-\nu)\, \epsilon^{\rho}}{\rho\, \mathrm{cdf}_{\mathrm{GB2}}(\epsilon, a, b, p, q)} + \frac{\nu\, l(\epsilon, \rho)}{f_{\mathrm{GB2}}(\epsilon, a, b, p, q)} \\
c_5^{-1} &= 1 - \theta + \frac{\zeta\, \theta\, \tau^{\zeta}\, (1 - \mathrm{cdf}_{\mathrm{GB2}}(\tau, a, b, p, q))\, u(\tau, \zeta)}{f_{\mathrm{GB2}}(\tau, a, b, p, q)} \\
a_5^{-1} &= \frac{(1-\theta)\, \tau^{-\zeta}}{\zeta\, (1 - \mathrm{cdf}_{\mathrm{GB2}}(\tau, a, b, p, q))} + \frac{\theta\, u(\tau, \zeta)}{f_{\mathrm{GB2}}(\tau, a, b, p, q)} \\
b_5^{-1} &= e_5\, \frac{\epsilon^{\rho}}{\rho} + \mathrm{cdf}_{\mathrm{GB2}}(\tau, a, b, p, q) - \mathrm{cdf}_{\mathrm{GB2}}(\epsilon, a, b, p, q) + \frac{a_5}{\zeta\, \tau^{\zeta}}
\end{aligned}
$$

This distribution depends on ten parameters $(\rho, \epsilon, \nu, a, b, p, q, \tau, \zeta, \theta)$ to be estimated. Note that the tdPGB2 is a limiting case of the dmPGB2 when $\nu \to 1$ and $\theta \to 1$ (the GB2 components at the tails, multiplied by $(1-\nu)$ and $(1-\theta)$, simply disappear).

4 The theoretical models generating the new distributions

The most common functions used to describe city size distributions have all an underlying theoretical model from which they are derived. Thus, Gabaix (1999) and Córdoba (2008) deduce power laws and, more specifically, Zipf's law. The same law, although in a very different setting, is also obtained by Hsu (2012), while, in turn, Eeckhout (2004) proposes a model for the lognormal. The more recent double Pareto lognormal comes from the theoretical models proposed by Reed (2002), Reed and Jorgensen (2004) and Giesen and Suedekum (2014). Other theoretical work in this line that should be mentioned, although they do not generate a specific statistical distribution, but do address theoretical aspects of Gibrat's law and Zipf's law, are Duranton (2006, 2007) and Rossi-Hansberg and Wright (2007).

The model presented in this Section is an adaptation of that of Parker (1999). It is not of a statistical nature as they are, to some extent, the productivity random shocks model of Eeckhout (2004) and, specially, the random growth-Gibrat model of Gabaix (1999). As a result, we move away from the type of approaches for which the resulting city size distribution is the steady state of a stochastic process of growth that, generally, is governed by a Brownian motion.

In Parker (1999), within a neoclassical labour market model where firms maximize profits, it is exactly deduced the density of the Generalised Beta of the second kind distribution presented above. What it is interesting about this model is that it allows, *mutatis mutandis*, to apply it to the case of urban nuclei to get our new distributions of Section 3: the tdPGB2 and dmPGB2. This shows the flexibility and versatility of the approach of Parker (1999).

We separate the study of city size distributions on three different regions: lower tail, body and upper tail. The lower tail will be *exactly delineated* from the body at the threshold value $\epsilon > 0$. Likewise, the body and the upper tail will be *exactly delineated* at the threshold value $\tau > 0$ (and, of course, $\tau > \epsilon$).

We will denote the number of cities (of our baseline model, tdPGB2) on the interval of population values $x \in (0, \epsilon)$ as $n_1(x)$; that on the interval $x \in [\epsilon, \tau]$ as $n_2(x)$, and finally that on the interval $x \in (\tau, \infty)$ as $n_3(x)$. The corresponding cumulative numbers of cities are, on the respective intervals:

$$\begin{aligned} N_1(x) &= \int_0^x n_1(x)\,dx\,, \quad x \in (0, \epsilon) \\ N_2(x) &= N_1(\epsilon) + \int_\epsilon^x n_2(x)\,dx\,, \quad x \in [\epsilon, \tau] \\ N_3(x) &= N_2(\tau) + \int_\tau^x n_3(x)\,dx\,, \quad x \in (\tau, \infty) \end{aligned}$$

The total number of cities, $N_3(\infty)$, is obviously a constant and it is assumed to have a finite upper bound, denoted by Θ, so that $N_3(\infty) < \Theta$.

The theoretical model that we will develop shortly determines endogenously the number of cities on each of the previous intervals relying on certain economic conditions. If we want to obtain an overall continuous probability density function we have to:

i) Assuming, as it is usual in the field, x to be a continuous variable,[8] obtain the continuity of $n_i(x)$, $i = 1, 2, 3$ on the respective intervals where they are defined. This is achieved by the model in a natural way (the equations defining $n_i(x)$, $i = 1, 2, 3$ in the next subsections correspond to continuous functions on their respective intervals).

ii) Impose continuity of the previous functions at the threshold points, namely

$$\lim_{x \to \epsilon} n_1(x) = \lim_{x \to \epsilon} n_2(x)\,, \quad \lim_{x \to \tau} n_2(x) = \lim_{x \to \tau} n_3(x) \tag{8}$$

[8] This assumption is based on the high number of cities in the samples (see Table 1) and proves to be very accurate in practice: see Section 6 on the empirical results.

iii) Divide the number of cities $n_i(x)$, $i = 1, 2, 3$ by the *total number of cities* $N_3(\infty)$, so that $n_i(x)/N_3(\infty)$, $i = 1, 2, 3$ give the correct *densities* of cities of population x on the respective intervals and also at the threshold points ϵ and τ.

At the end, this process will lead exactly to the generation of the previously defined tdPGB2.

We develop accordingly the model considering the three regions separately; afterwards, we will consider the joint results.[9]

4.1 Model for the lower tail (variables and parameters with subindex 1) of the tdPGB2

The model consists on maximizing the net output function in monetary units of the whole urban system of a country at a given time. The number of cities is finite, so, as already mentioned, there exists an upper bound Θ so that the total number of cities is strictly less than Θ.

As stated before, the size (population) of each city is denoted by x. The human capital level of each city depends on the population according to the function $\psi_1(x)$, being $\psi_1(0) = 0$. We assume it to be positive and increasing. Let $n_1(x)$ be the number of cities of population x. Each inhabitant supplies one unit of labour inelastically. The gross output of the cities of population x is $F_1[n_1(x), \psi_1(x)]$, in such a way that it depends on the number of cities of population x and on its level of human capital. There are diminishing returns to the number of cities, i.e., $\dfrac{\partial F_1}{\partial n_1} > 0$, $\dfrac{\partial^2 F_1}{\partial n_1^2} < 0$ at all population levels. There are also monetary congestion costs $c_1(x)$ associated to a city of population x. These costs reduce the gross output of each urban settlement. We assume that $c_1(x) > 0$ and $c_1'(x) > 0$ (see Subsection 4.4 for an economic justification

[9] Obviously, the maximum of a sum of three addends (lower tail, body, and upper tail) is the sum of the maxima of each of them.

of all these assumptions).

Thus, the net output of the cities of population $x \in (0, \epsilon)$ is $F_1[n_1(x), \psi_1(x)] - c_1(x)n_1(x)$, and the net output of all cities with populations between 0 and ϵ (the lower tail) is the corresponding definite integral of this last quantity. In order to specify more the problem, we assume further that $F_1[n_1(x), \psi_1(x)] = \psi_1(x)n_1(x)^\beta$, where $\beta \in (0, 1)$ in order to satisfy the signs of the derivatives (see above) of the gross output with respect to the number of cities.

Therefore, the cities' optimal control problem for the lower tail, where the output price has been normalised to unity, can be stated as

$$\begin{aligned}
\max_{n_1} \quad & \int_0^\epsilon (\psi_1(x)n_1(x)^\beta - c_1(x)n_1(x))\, dx \\
\text{subject to :} \quad & \frac{dN_1(x)}{dx} = n_1(x) \\
& N_1(0) = 0 \\
& N_1(\epsilon) = \int_0^\epsilon n_1(x)\, dx < \Theta \\
& n_1(x) \in (0, \infty)
\end{aligned}$$

where the state variable is $N_1(x)$, which represents the cumulative number of cities up to a population of x, and the control is $n_1(x)$. Note that the variable x is not time but population.[10]

This problem can be solved by using Pontryagin's Maximum Principle (Pontryagin et al., 1962). The associated Hamiltonian function is simply (we take the same convention as Parker (1999) for the Lagrange multiplier)

$$H_1(x, N_1, n_1, \lambda_1) = \psi_1(x)n_1(x)^\beta - c_1(x)n_1(x) - \lambda_1(x)n_1(x)$$

[10]This problem can be formulated as well as an ordinary variational problem, see Kamien and Schwartz (1991) and Chiang (1992), but in the presented way the economic meaning of the imposed conditions is clearer.

The state and costate equations are the following[11]

$$\frac{dN_1(x)}{dx} = -\frac{\partial H_1}{\partial \lambda_1} = n_1(x)$$
$$\frac{d\lambda_1(x)}{dx} = \frac{\partial H_1}{\partial N_1} = 0$$

and thus $\lambda_1(x) = \lambda_1 = \text{Constant}$. The control to be chosen $n_1(x)$ is the one which maximizes the Hamiltonian and belongs to an open interval, so no corner solutions may arise. The first order condition is just

$$\frac{\partial H_1}{\partial n_1} = \psi_1(x)\beta n_1(x)^{\beta-1} - c_1(x) - \lambda_1 = 0 \tag{9}$$

The second order derivative is

$$\frac{\partial^2 H_1}{\partial n_1^2} = \psi_1(x)\beta(\beta-1)n_1(x)^{\beta-2} < 0\,, \quad x \in (0,\epsilon)$$

and therefore the first order condition becomes necessary and sufficient for a strict global maximum. From equation (9) we can solve for $n_1(x)$ as follows

$$n_1(x) = \left(\frac{\beta\psi_1(x)}{c_1(x)+\lambda_1}\right)^{1/(1-\beta)}\,, \quad x \in (0,\epsilon)$$

It is time now, according to the behaviour imposed above to the human capital and costs functions of a city of size x, to define specific functional forms for both functions: $\psi_1(x) = A_1 x^{\gamma_1}$, $c_1(x) = k_1 x^{b_1}$, where $A_1 > 0$, $\gamma_1 > 0$, $k_1 > 0$ and $b_1 > 0$. Consequently, we have

$$n_1(x) = \left(\frac{\beta\, A_1\, x^{\gamma_1}}{k_1 x^{b_1}+\lambda_1}\right)^{1/(1-\beta)}\,, \quad x \in (0,\epsilon)$$

We want $n_1(x)$ to be a pure Pareto power law, that is, to be proportional to a power

[11] It is not necessary to impose the transversality conditions because $N_1(\epsilon) < \Theta$.

function. For that, it is necessary and sufficient that $\lambda_1 = 0$.[12] This condition has an economic meaning. Indeed, (minus) the Lagrange multiplier $-\lambda_1$ is "the marginal valuation of the associated state variable $N_1(x)$ at population x. If there were an exogenous, tiny increment to the state variable at population x and if the problem were modified optimally thereafter, the increment in the total value of the objective function would be at the rate $-\lambda_1$". (Kamien and Schwartz, 1991, p. 138, adapted). If the maximum net output of the cities does not vary when the total number of cities up to population x increases by a small amount, then the mentioned marginal valuation vanishes. This is the economic characterization of the pure Pareto lower tail according to our model.

Then, with $\lambda_1 = 0$ we simply have $n_1(x) = \left(\frac{\beta A_1}{k_1}\right)^{1/(1-\beta)} x^{\frac{\gamma_1 - b_1}{1-\beta}}$ so in order to have a pure Pareto lower tail we require that the corresponding Pareto exponent ρ (in the notation of Section 3) satisfies $\rho = \frac{\gamma_1 - b_1}{1-\beta} + 1 > 0$. The assumptions made so far about the values of the parameters β, b_1 and γ_1 are compatible with the fulfillment of this expression and the empirical analysis confirms that the estimations of ρ are always positive.

4.2 Model for the body (variables and parameters with subindex 2) of the tdPGB2

In the body of the tdPGB2 distribution we assume a similar model than for the lower tail on the corresponding interval $[\epsilon, \tau]$. Therefore, the number of cities in the body can be found to be

$$n_2(x) = \left(\frac{\beta \psi_2(x)}{c_2(x) + \lambda_2}\right)^{1/(1-\beta)} , \quad x \in [\epsilon, \tau]$$

[12]As already indicated, the proper probability density function on the interval $x \in (0, \epsilon)$ is $n_1(x)/N_3(\infty)$ (see also Parker (1999)). Since $N_3(\infty)$ is a finite positive constant we have a Pareto distribution in the lower tail if and only if, as stated in the text, $\lambda_1 = 0$. This footnote also applies to the power law in the upper tail with the corresponding Lagrange multiplier, see Subsection 4.3.

For the body, we do not expect it to be pure Pareto, so that we will consider that $\lambda_2 > 0$. Also, we define $\psi_2(x) = A_2 x^{\gamma_2}$, $c_2(x) = k_2 x^{b_2}$, where now we assume that $A_2 > 0$, $\gamma_2 > 0$, $k_2 > 0$, $b_2 > 0$, so that

$$n_2(x) \quad = \quad \left(\frac{\beta\, A_2 x^{\gamma_2}}{k_2 x^{b_2} + \lambda_2} \right)^{1/(1-\beta)} \quad , \quad x \in [\epsilon, \tau]$$

Comparing this last expression with the definition of the GB2 distribution (see eq. (1)) both functions can be properly related, so that $n_2(x)/N_3(\infty)$ is, up to a positive multiplicative constant, the expression of $f_{\mathrm{GB2}}(x, a, b, p, q)$ of (1). Indeed, we simply have

$$n_2(x) \quad = \quad \left(\frac{\beta A_2}{\lambda_2} \right)^{1/(1-\beta)} B(p,q) \frac{b^{ap}}{a} f_{\mathrm{GB2}}(x, a, b, p, q)\,, \quad x \in [\epsilon, \tau]$$

with the identifications of the parameters

$$\begin{aligned} a &= b_2 \\ b &= \left(\frac{\lambda_2}{k_2} \right)^{1/b_2} \\ p &= \frac{1}{b_2} \left(1 + \frac{\gamma_2}{1-\beta} \right) \\ q &= \frac{1}{1-\beta} - \frac{1}{b_2} \left(1 + \frac{\gamma_2}{1-\beta} \right) \end{aligned}$$

Since $\beta \in (0,1)$, let us point out that it should happen that $p + q > 1$ according to our economic model.

4.3 Model for the upper tail (variables and parameters with subindex 3) of the tdPGB2

In the upper tail we assume a similar model than for the lower tail and body, but now defined on the interval (τ, ∞). The corresponding number of cities in the upper tail can be found to be

$$n_3(x) \quad = \quad \left(\frac{\beta\psi_3(x)}{c_3(x)+\lambda_3}\right)^{1/(1-\beta)} \quad , \quad x \in (\tau, \infty)$$

Now, we take $\psi_3(x) = A_3 x^{\gamma_3}$, $c_3(x) = k_3 x^{b_3}$, where $A_3 > 0$, $\gamma_3 > 0$, $k_3 > 0$ and $b_3 > 0$. Thus

$$n_3(x) = \left(\frac{\beta\, A_3 x^{\gamma_3}}{k_3 x^{b_3} + \lambda_3}\right)^{1/(1-\beta)} \quad , \quad x \in (\tau, \infty)$$

In this case, we want to obtain again a pure Pareto upper tail. Thus, we require that $\lambda_3 = 0$, with analogous economic interpretation to that in the lower tail. Then, we have $n_3(x) = \left(\frac{\beta\, A_3}{k_3}\right)^{1/(1-\beta)} x^{\frac{\gamma_3 - b_3}{1-\beta}}$ and it must happen that the Pareto exponent ζ (in the notation of Section 3) satisfies $-\zeta = \frac{\gamma_3 - b_3}{1-\beta} + 1 < 0$. The assumptions made so far about the values of the parameters β, b_3 and γ_3 are compatible with the fulfillment of this expression and the empirical analysis confirms that the estimations of ζ are always positive.[13]

Finally, for the overall distribution, we can impose the following natural conditions, namely, continuity of the human capital ($\psi(x)$) and *effective cost functions* ($c(x) + \lambda$) at the threshold values ϵ and τ:

$$\lim_{x\to\epsilon} A_1 x^{\gamma_1} = \lim_{x\to\epsilon} A_2 x^{\gamma_2} \,, \quad \lim_{x\to\tau} A_2 x^{\gamma_2} = \lim_{x\to\tau} A_3 x^{\gamma_3} \tag{10}$$

$$\lim_{x\to\epsilon} k_1 x^{b_1} = \lim_{x\to\epsilon} k_2 x^{b_2} + \lambda_2 \,, \quad \lim_{x\to\tau} k_2 x^{b_2} + \lambda_2 = \lim_{x\to\tau} k_3 x^{b_3} \tag{11}$$

These conditions have the immediate consequence that (8) holds. And, as stated previously, dividing $n_i(x)$, $i = 1, 2, 3$ by the total number of cities $N_3(\infty)$ provides the exact probability density function on each interval and the threshold values corre-

[13]The empirical evidence corroborates that Zipf's law holds especially for the larger cities. Thus, it is obtained for the upper tail when $\zeta = 1$ or, in other words, $\frac{b_3 - \gamma_3}{1-\beta} = 2$. In the estimations shown in Table 2, ζ is statistically equal to one at the 5% significance level only for the sample of incorporated places in 1900.

sponding to the tdPGB2 of Subsection 3.3, and of course taking these quotients implies overall normalization to the unit.[14] Let us remark that in Subsection 3.3, the definition of the e_4, a_4, b_4 quantities by eqs. (5), (6) and (7) reflects exactly the conditions of the overall probability density function to be continuous at the threshold values ϵ and τ and to be normalized to the unit.

The distribution dmPGB2 can be obtained in a similar way, with the only distinction that at the respective tails there are two types of cities: one whose marginal valuation vanish (the corresponding Lagrange multiplier vanishes), and others whose marginal valuation of the net output with respect to variations of the cumulative number of cities is negative. The first type of these are formally equal to the $n_1(x)$ (resp. $n_3(x)$) above for the lower (resp. upper) tail. They are combined convex and linearly with the $n_2(x)$ above (but extended in domain to both tails). At the body, the $n_2(x)$ in the dmPGB2 is the same as that of the tdPGB2. The conditions (10) and (11) (and, of course, (8) as well) can be taken to hold also in this case. The cumulative numbers of cities differ slightly from those of the tdPGB2 but can be computed in a straightforward and analogous way.[15]

As a final outcome, we have demonstrated that we can obtain the tdPGB2 and dmPGB2 probability distributions from theoretical economic models in which the population of a country locates in cities of different sizes so as to maximize the net output of the overall urban system.

Different types of data might be best described by different density functions. As we will see on Section 6, the tdPGB2 is very appropriate for describing US places and the dmPGB2 for US CCA clusters. For places, we will compare as well the tdPGB2 with the baseline distribution GB2 obtained from the model of Parker (1999) in order to see how the assumption of pure Pareto tails leads to a remarkable improvement.

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

[15]More details are available from the authors upon request, and we provide a MATHEMATICA notebook with the corresponding Hamiltonians and optimization equations as supplementary material.

For CCA clusters, the estimated parameters p, q of the GB2 are such that $p + q < 1$ so it cannot be derived directly from the economic model of Parker (1999) and the corresponding results are omitted.

4.4 Economic explanations for the shape of the functions $\psi(x)$ and $c(x)$

There are two functions that define the most important characteristics of our model. On the one hand, $\psi(x)$ relates the stock of human capital in each city with its size x. On the other hand, $c(x)$ provides the cost function for each nucleus. With regards to the first, the proposed functional form and values of the γ parameters in the previous subsections, make it to be an increasing function ($\gamma > 0$). It is obvious that it has to be increasing: larger cities have, on average, more human capital. It can be convex ($\gamma > 1$), linear ($\gamma = 1$), or concave ($\gamma < 1$). In the first (third) case, if a city has a size that is, for example, twice the size of another city, its human capital stock will be larger (smaller) than twice the one of the smaller city. Our theoretical model is compatible with the three options and, therefore, is an empirical question, out of the scope of this paper, to determine which of them holds.

Of course, from an economic point of view, the most interesting scenario is that in which human capital accumulates at rates that are increasing with respect to the size of the urban settlements. We can justify this behaviour according to two arguments.

First, "there is some evidence suggesting that human capital accumulates more quickly in urban areas" (Glaeser and Resseger, 2010). This empirical evidence is also corroborated in Moretti (2004). In this last reference the nineteen MSAs of USA with larger (and smaller) percent of college graduates in 2000, ranging from the 43.6% of San Francisco, CA, to the 11.3% of Danville, VA, are presented. The average size of the first (the calculation is ours) in 2000 is 1,087,157 inhabitants, while the latter is nearly a quarter, namely 289,863 inhabitants. This supports that, in general, there is a positive relationship between accumulation of human capital and size of the city.

Secondly, the existence of agglomeration economies. Estimates of their magnitude deduce that doubling the size or density of an urban area increases its productivity between 2% to 8% (see the excellent and comprehensive surveys about the topic of Rosenthal and Strange (2004), Melo et al. (2009), Puga (2010) and Combes and Gobillon (2015)). In our model, these productivity gains associated with a larger size are due to human capital accumulation with respect to the population of the city: an urban nucleus with twice the labour force than another has more than twice human capital provided that $\gamma > 1$. This is one of the inputs in the production function, and therefore causes that the productivity per worker is higher in the larger city. The greater accumulation in absolute and percentage terms of human capital in the biggest cities, obviously, is not the only cause which may give rise to the emergence of agglomeration economies and productivity gains (see the classic micro-fondations of sharing, matching, and learning in Duranton and Puga (2004)), but it is one of the more plausible and it has been empirically corroborated by Rauch (1993) (raising the average education level of a metropolitan area by one year increases the total factor productivity by 2.8%).[16]

It is time now to justify the shape adopted for the cost function $c(x)$, which is increasing with respect to city size ($b > 0$). We label these costs, without loss of generality, as congestion costs. Now, we are thinking about the factors related to all the "bad" that is traditionally associated with bigger cities. This is why these costs reduce the gross output of each urban nucleus and give rise to the obtention of the net output, available for consumers and maximized in our model. Some of these factors are the following: crime, traffic, diseases, pollution and housing prices. All of them tend

[16]It is outside the scope of this work to deepen into the economic mechanisms that explain why larger cities might accumulate human capital more quickly and are, on average, more productive. Useful references to this regard are the seminal article by Lucas Jr. (1988), in particular its sixth section devoted to "Cities and growth"; Glaeser (1998), which emphasizes the role of cities as nodes for interaction and learning, and Glaeser (2011), where you can find an informative exposition, with many actual examples, of all these ideas. See also Yankow (2006) and Combes et al. (2008), which analyze human capital divergences as a key driver of earning differences (wage premium in bigger cities).

to increase in absolute terms with city size. In per capita or relative terms things are not so clear. Bettencourt and West (2010) report that the magnitude of crime, traffic, and certain diseases is multiplied by 2.3 if the population of a city is doubled; in this case $b > 1$. Regarding pollution there is a certain consensus about the fact that larger cities are, on average, greener (Glaeser, 2011), see also Glaeser and Kahn (2010) for the intensity of CO_2 emissions; in this case $b < 1$. Finally, The connection between city size and housing prices is a complex topic which depends on local geography, regulatory policies and internal spatial structure of the cities: see Saiz (2010), Glaeser et al. (2012), and references therein for an overview; in this case $b > 1$ or $b < 1$. Again, as for the $\psi(x)$ function, our theoretical model is compatible with all the possibilities and, therefore, is an empirical question to determine whether $c(x)$ increases with city size at a more than proportional rate ($b > 1$) or not ($b < 1$). It is not the aim of this paper to deal specifically with this interesting applied topic.

We want to remark, as a reflection, an important feature of our theoretical model. That is, the parameters of the overall distribution so obtained, the tdPGB2, depend on the elasticities of the gross output with respect to the number of cities with a given population (β), of the human capital stock with respect to population (γ_i, $i = 1, 2, 3$) and of the city costs with respect to the population as well (b_i, $i = 1, 2, 3$). In particular, the Pareto exponents at the lower and upper tails are related to the previous parameters. These elasticities might vary over time, mainly due to economic reasons, so we obtain that city size distribution is explained at a given time by the economic conditions that determine it. Therefore, this model may help in explaining the observed persistence of the city size distribution in the short term (Black and Henderson, 1999; Kim, 2000; Beeson et al., 2001), because the previously mentioned elasticities probably have a slow time variation. However, in the long term the variations can be quite remarkable, as Batty (2006) points out.

The content of the last paragraph lead us to two important outcomes. First, the urban policy implications of the previous discussion are, in our opinion, very important. Secondly, the interpretation of city size distribution as a steady state (by definition, with no time changes at all) of a stochastic process seems to be questionable.

Finally, Section 6 shows that our new distributions describe the actual data with an extremely high precision. This is an example of how the interaction and feedback between empirics and economic modeling might lead to a new theoretical understanding, in this case, of the way city size distributes.

5 The databases

In this article, we use data about US urban centers from three sources. The first is the decennial data of the US Census Bureau of "incorporated places" without any size restriction, for the years 1900, 1950 and 2000. These 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.[17] This data was first introduced in González-Val (2010), see therein for details, and later used in other works like González-Val et al. (2015).

The second source consists of all US urban places, unincorporated and incorporated, and without size restrictions, also provided by the US Census Bureau for the years 2000 and 2010. The data for the year 2000 was first used in Eeckhout (2004) and later in Levy (2009), Eeckhout (2009), Giesen et al. (2010), Ioannides and Skouras (2013) and Giesen and Suedekum (2014). The two samples were also used in González-Val et al. (2015).

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

[17] `http://www.census.gov/prod/www/decennial.html` Last accessed: April 30$^{\text{th}}$, 2015.

Algorithm" (CCA) to get "an automated and systematic way of building population clusters based on the geographical location of people." (*op. cit.*) We use their US clusters data based on the radii of 2 and 3 km. (the samples with higher sample sizes) and for the years 1991 and 2000. This kind of data has been used in Ioannides and Skouras (2013) and Giesen and Suedekum (2014).

We do not consider, on the other hand, types of data like Economic Areas of Berry and Okulicz-Kozaryn (2012); Core Based Statistical Areas (CBSA), popularized by Duranton (2007) and also employed by Lee and Li (2013); MSAs (see, e.g., Ioannides and Skouras (2013) for their definition) and used in many papers. These three types violate our principle that the small nuclei also matter and that there should be no truncation point: there are only 366 MSAs, 940 CBSAs, and less than 200 Economic Areas in 2010.

The descriptive statistics of the data sets used in this paper can be seen in Table 1.[18]

6 Results

We show briefly in this Section how our new distributions, the tdPGB2 and the dmPGB2, perform in the fit of the size of US places (incorporated and all) and CCA nuclei, compared to well-known distributions of city size as the lognormal (lgn) and the double Pareto lognormal (dPln). For the case of US places, we show also the corresponding results of the Generalized Beta of the second kind (GB2).

Firstly, we show in Table 2 (places) and 3 (CCA nuclei) the estimation results for the used distributions. We observe that the estimations are rather precise in all cases.[19]

[18] The results for the remaining years of incorporated places in the period 1900-2000, and for the CCA nuclei with radii 4 and 5 km. are similar and are not shown for the sake of brevity. They are avaliable from the authors upon request. The previous statement also applies to all of the Tables in Section 6.

[19] We show the estimators and the standard errors (SE) computed according to the procedure of McCullough and Vinod (2003) and Efron and Hinkley (1978). Also, we have checked that all of our estimations satisfy the requirements raised on the first of these references. The dmPGB2 cannot be always be estimated for places samples, and the tdPGB2 yields worse results than the dmPGB2 for CCA clusters. This means that the two types of data (places, clusters) are essentially different. These differences have already been

We show in Table 4 the results of the Kolmogorov–Smirnov (KS), Crámer–von Mises (CM) and Anderson–Darling (AD) tests for the studied samples and used density functions. The AD test is very appropriate when one wants to assess the adequacy of the distributions at the tails, see, e.g., Cirillo (2013). The first remarkable result is that the lognormal (lgn) is strongly rejected always, so this specification seems not to be as good as a parametric description in practice, at least for US places and CCA clusters.[20] The second observation is that a similar thing happens for the double Pareto lognormal (dPln) and GB2: they are rejected almost always, with the only exception of the sample of incorporated places in 1900. The third observation is that for places, the GB2 offers always lower values of the tests' statistics than the dPln, meaning that an immediate adaptation of the economic model of Parker (1999) may yield a better city size description than that of the latter distribution (see also, to corroborate this outcome, the values of the AIC and BIC information criteria for the dPln and GB2 in Table 6).

In addition, at the same time and with the same techniques, the herein proposed tdPGB2 (for places) and dmPGB2 (for clusters) are not rejected in the 100% of the cases by the three tests, and not by a slight margin precisely. To this respect, the differences in the statistics of the used tests are huge when going from the lognormal to the dPln and then to either the tdPGB2 or the dmPGB2, depending on the type of data. This means that both the tdPGB2 and dmPGB2 are two robust and excellent parametric specifications for the size of US places and CCA clusters, respectively. We also see that our modification of the economic model of Parker (1999) in order to account for pure Pareto tails (for places), is crucial for obtaining an excellent fit, versus the consideration of a single GB2 for all the range of city sizes.

As a complementary analysis we have computed also the metrics defined in Duranton (2007) in order to see "how good approximation the used parametric distributions

pointed out by Giesen and Suedekum (2014).

[20]This last fact has been previously highlighted by Giesen and Suedekum (2014).

give in quantitative terms". We take all of the US cities on our study (Duranton uses 232 and 922 French and US nuclei, respectively), but nevertheless it is possible to compute the quantities msd and the pseudo R^2 as follows:

$$\begin{aligned} \mathrm{msd} &= \frac{1}{m}\sum_{j=1}^{m}[\text{Actual log Size}(j) - \text{Mean log Simulated Size}(j)]^2 \\ R^2 &= 1 - \frac{\mathrm{msd}}{\mathrm{var}} \end{aligned}$$

where var is the empirical variance for log city sizes and m is the number of cities $(j = 1, \ldots, m)$ in the empirical sample and msd is (see the formula) the mean of the squared differences between the actual log-sizes data and the means of the log simulated sizes by the corresponding distribution. After performing simulations generating 100 random samples[21] we show the corresponding results in Table 5. We see that the msd quantities are overwhelmingly lower for the tdPGB2 and dmPGB2 than for the other studied distributions. This is also reflected in the R^2 values, which also favour these distributions.

Finally, we show in Table 6 the results of the Akaike Information Criterion (AIC) and the Schwarz or Bayesian Information Criterion (BIC), which are standard in the literature, in order to choose between the proposed distributions. We see that for places the selected specification is clearly the tdPGB2 by both AIC and BIC criteria. For CCA clusters, a similar thing happens for the dmPGB2.

In short, by all of the studied criteria, we obtain a very strong result: the US city size distribution of places can be safely taken as our new tdPGB2. For CCA clusters, an analogous statement holds for our new dmPGB2.[22]

To end the empirical part of this paper we offer an informal graphical approxima-

[21]Each of these samples is of the sample size of the empirical data. The total generated observations range from about 1,005,000 to 3,002,000 for each case under study. We have chosen a number of generated samples reasonably high enough while maintaining computational feasibility.

[22]We have also compared the distributions introduced in Ioannides and Skouras (2013): they are also outperformed by our new ones. Likewise, our results also outperform those of Berliant and Watanabe (2015). More details are available from the authors upon request.

tion to our best distributions (Figure 1). We have chosen for all places the last available Census data on 2010, and for CCA clusters the sample of year 2000, 3km. For other possible cases the results are very similar. We see that the fits at the tails and of the entire densities are, also visually, extremely good.

7 Conclusions

This work tries to contribute, both from a theoretical perspective and from an empirical approach, to the literature on city size distributions.

To summarize, the contributions from a theoretical point of view are the following. The main result is that the new statistical distributions introduced in this paper, namely the "treshold double Pareto Generalized Beta of the second kind" (tdPGB2) and "double mixture Pareto Generalized Beta of the second kind" (dmPGB2) are exactly deduced using a simple model in which the population of a country locates in cities of different sizes so as to maximize the net output of the system of cities. There are four basic features of this model. Firstly, that it is built up piecewise, taking into account the specific particularities of the lower tail, the body and the upper tail. Secondly, the production function is increasing and concave in the number of cities, so that it complies with the law of diminishing returns. Thirdly, the human capital stock of a city is increasing with respect to city size. And fourthly, the congestion costs that lessen the gross output of each urban unit are also increasing with respect to cities population.

The theoretical parameters of the overall distribution are given explicitly, at any given time, in terms of the elasticities of the gross output with respect the number of cities, of human capital stock with respect to city size, and of costs with respect to city size, as well. Economic conditions may change and accordingly the associated elasticities, thus determining the resulting city size distribution. This fact opens the door for urban economic policy recipes trying to govern the economic conditions previously mentioned. Therefore, our approach is rooted in economic modeling, rather than in pure statistical reasoning.

Empirically, the data sets we consider are those of the US incorporated places in 1900, 1950, and 2000. Also, all US places in 2000, 2010 and the CCA Clusters of Rozenfeld et al. (2011) for the years 1991, 2000 and radii of 2 and 3 km. As mentioned, we have introduced the tdPGB2 and dmPGB2 distributions. The first is pure Pareto at both tails and Generalized Beta of the second kind (GB2) on the body. The second has convex linear combinations of Pareto and GB2 at the tails, and again, GB2 on the body. Both of these two new density functions systematically outperform the most widely used ones in the literature, namely, the Pareto, the lognormal and the double Pareto lognormal (dPln). In fact, the tdPGB2 is the distribution chosen to describe US places, incorporated and all. In turn, the dmPGB2 is the one chosen to describe US CCA clusters. These results are robust to a battery of different independent criteria: Kolmogorov–Smirnov, Crámer–von Mises, Anderson–Darling tests; msd (mean squared differences) and pseudo-R^2 metrics of Duranton (2007); Akaike Information Criterion and Bayesian Information Criterion. All of these results point out clearly to the fact that the US city size distribution can be safely taken by either the tdPGB2 (places) or by the dmPGB2 (clusters).

In this sense, from an empirical point of view the main contributions of the paper are the following:

i) A classical distribution, with underlying theoretical model (Eeckhout, 2004) as the lognormal is widely surpassed by new ones like the tdPGB2 or dmPGB2, depending on the type of data.

ii) A newer distribution, as it is the dPln, also with underlying theoretical foundations (Giesen and Suedekum, 2014), is outperformed by new ones, again the tdPGB2 or dmPGB2.

iii) The mentioned new distributions confirm something that has been known for a long time: that the upper tail can be taken as Pareto (pure or mixed). Moreover, also the lower tail can be taken as Pareto (pure or mixed).

iv) Against all evidence accumulated so far, the best body is not lognormal but Generalized Beta of the second kind, and this distinction has an economic theoretical origin.

The empirical results are in an extremely good agreement with the theoretical models developed, based on economic foundations. Both theory and strong empirical support may lead to a new way of looking at city size distributions.

References

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

Arcaute, E., Hatna, E., Ferguson, P., Youn, H., Johansson, A., and Batty, M. (2014). Constructing cities, deconstructing scaling laws. *Journal of the Royal Society Interface*, 12.

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

Batty, M. (2008). The size, scale, and shape of cities. *Science*, 319:769–771.

Batty, M. (2013). *The new science of cities*. The MIT Press.

Bee, M., Riccaboni, M., and Schiavo, S. (2013). The size distribution of US cities: not Pareto, even in the tail. *Economics Letters*, 120:232–237.

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

Behrens, K., Mion, G., Murata, Y., and Südekum, J. (2014). Spatial frictions. *DICE Discussion Paper No. 160*.

Berliant, M. and Watanabe, H. (2015). Explaining the size distribution of cities: Extreme economies. *Quantitative Economics*, 6:153–187.

Berry, B. J. L. and Okulicz-Kozaryn, A. (2012). The city size distribution debate: resolution for US urban regions and megalopolitan areas. *Cities*, 29:517–523.

Bettencourt, L. and West, G. (2010). A unified theory of urban living. *Nature*, 4667:912–913.

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.

Blank, A. and Solomon, S. (2000). Power laws in cities population, financial markets and internet sites (scaling in systems with a variable number of components). *Physica A*, 289:279–288.

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.

Budd, E. C. (1970). Distribution issues: Trends and policies. Postwar changes in the size distribution of income in the U. S. . *American Economic Review*, 60(2):247–260.

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

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

Chiang, A. C. (1992). *Elements of dynamic optimization*. McGraw-Hill.

Cirillo, P. (2013). Are your data really Pareto distributed? *Physica A*, 392:5947–5962.

Combes, P. P., Duranton, G., and Gobillon, L. (2008). Spatial wage disparities: sorting matters! *Journal of Urban Economics*, 63(2):723–742.

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

Combes, P. P. and Gobillon, L. (2015). The empirics of agglomeration economies. In Duranton, G., Henderson, V., and Strange, W. C., editors, *Handbook of Regional and Urban Economics, forthcoming*, volume 5. Elsevier.

Córdoba, J. C. (2008). On the distribution of city sizes. *Journal of Urban Economics*, 63(1):177–197.

Dagum, C. (1979). Generating systems and properties of income distribution models. In *Conference of Canadian Incomes, May 10-12*.

Desmet, K. and Rappaport, J. (2014). The settlement of the United States, 1800 to 2000: the long transition towards Gibrat's law. *The Federal Reserve Bank of Kansas City Research Workin Papers*.

Duranton, G. (2006). Some foundations for Zipf's law: product proliferation and local spillovers. *Regional Science and Urban Economics*, 36:542–563.

Duranton, G. (2007). Urban evolutions: The fast, the slow, and the still. *American Economic Review*, 97(1):197–221.

Duranton, G. and Puga, D. (2004). Micro-foundations of urban agglomeration economies. In Henderson, V. and Thisse, J. F., editors, *Handbook of Regional and Urban Economics*, volume 4, chapter 48, pages 2063–2117. Elsevier.

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.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, 65(3):457–482.

Favaro, J. M. and Pumain, D. (2011). Gibrat revisited: an urban growth model incorporating spatial interactions and innovation cycles. *Geographical Analysis*, 43:261–286.

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

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. (2014). City age and city size. *European Economic Review*, 71:193–208.

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. L. (1998). Are cities dying? *Journal of Economic Perspectives*, 12(2):139–160.

Glaeser, E. L. (2011). *Triumph of the city*. The Penguin Press.

Glaeser, E. L., Gottlieb, J. D., and Tobio, K. (2012). Housing booms and city centers. *NBER Working Paper No. 17914*.

Glaeser, E. L. and Kahn, M. E. (2010). The greenness of cities: carbon dioxide emissions and urban development. *Journal of Urban Economics*, 67(3):404–418.

Glaeser, E. L. and Resseger, M. G. (2010). The complementarity between cities and skills. *Journal of Regional Science*, 50(1):221–244.

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., Sanz-Gracia, F., and Vera-Cabello, M. (2015). Size distribution for all cities: Which one is best? *Papers in Regional Science*, (1):177–197.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

Holmes, T. J. and Lee, S. (2010). *Agglomeration Economics*, chapter Cities as six-by-six-mile squares: Zipf's Law? The University of Chicago Press.

Hsu, W.-T. (2012). Central place theory and city size distribution. *The Economic Journal*, 122:903–932.

Hsu, W.-T., Mori, T., and Smith, T. (2014). Spatial patterns and size distribution of cities. *Kier Discussion Paper No. 882*.

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

Ioannides, Y. M. and Overman, H. G. (2004). Spatial evolution of the US urban system. *Journal of Economic Geography*, 4(2):131–156.

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

Kamien, M. I. and Schwartz, N. L. (1991). *Dynamic optimization. The calculus of variations and optimal control in Economics and Management*. Elsevier.

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.

Krugman, P. R. (1996). *The self-organizing economy*. Blackwell Publishers. Cambridge, MA.

Lee, S. and Li, Q. (2013). Uneven landscapes and city size distributions. *Journal of Urban Economics*, 78:19–29.

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

Lucas Jr., R. E. (1988). On the mechanics of economic development. *Journal of Monetary Economics*, 22:3–42.

Luckstead, J. and Devadoss, S. (2014a). A comparison of city size distributions for China and India from 1950 to 2010. *Economics Letters*, 124:290–295.

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

Macaulay, F. R. (1922). *Income in the United States, its amount and distribution, 1909-1919, Volume II: Detailed Report*, chapter The personal distribution of income in the United States. National Bureau of Economic Research.

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.

Mandelbrot, B. (1960). The Pareto-Lévy law and the distribution of income. *International Economic Review*, I:79–106.

McCullough, B. D. and Vinod, H. D. (2003). Verifying the solution from a nonlinear solver: A case study. *American Economic Review*, 93(3):873–892.

McDonald, J. B. (1984). Generalized functions for the size distribution of income. *Econometrica*, 52(3):647–665.

McDonald, J. B. and Xu, Y. J. (1995). A generalization of the beta distribution with applications. *Journal of Econometrics*, 66:133–152.

Melo, P., Graham, D., and Noland, R. (2009). A meta-analysis of estimates of urban agglomeration economies. *Regional Science and Urban Economics*, 39(3):332–342.

Moretti, E. (2004). Human capital externalities in cities. In Henderson, V. and Thisse, J. F., editors, *Handbook of Regional and Urban Economics*, volume 4, chapter 51, pages 2243–2291. Elsevier.

Overman, H. G. and Ioannides, Y. M. (2001). Cross-Sectional evolution of the US city size distribution. *Journal of Urban Economics*, 49(3):543–566.

Parker, S. C. (1999). The generalised beta as a model for the distribution of earnings. *Economics Letters*, 62:197–200.

Parr, J. B. (1985). A note on the size distribution of cities over time. *Journal of Urban Economics*, 18:199–212.

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

Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V., and Mishchenko, E. F. (1962). *The mathematical theory of optimal processes*. Wiley Interscience Publishers.

Puente-Ajovín, M. and Ramos, A. (2015). On the parametric description of the French, German, Italian and Spanish city size distributions. *The Annals of Regional Science.*, 54(2):489–509.

Puga, D. (2010). The magnitude and causes of agglomeration economies. *Journal of Regional Science*, 50(1):203–219.

Rauch, F. (2014). Cities as spatial clusters. *Journal of Economic Geography*, 14(4):759–773.

Rauch, J. E. (1993). Productivity gains from geographic concentration of human capital: evidence from the cities. *Journal of Urban Economics*, 34(3):380–400.

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

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

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

Reed, W. J. 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.

Rosenthal, S. S. and Strange, W. C. (2004). Evidence on the nature and sources of agglomeration economies. In Henderson, V. and Thisse, J. F., editors, *Handbook of Regional and Urban Economics*, volume 4, chapter 49, pages 2119–2171. Elsevier.

Rossi-Hansberg, E. and Wright, M. L. J. (2007). Urban structure and growth. *Review of Economic Studies*, 74(2):597–624.

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

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

Rybski, D., García-Cantú-Ros, A., and Kropp, J. P. (2013). Distance-weighted city growth. *Physical Review E*, 87:042114.

Saiz, A. (2010). The geographic determinants of housing supply. *The Quarterly Journal of Economics*, 125(3):1253–1296.

Schluter, C. and Trede, M. (2013). Gibrat, Zipf, Fisher and Tippet: City size and growth distributions reconsidered. *Working Paper 27/2013 Center for Quantitative Economics WWU Münster*.

Soo, K. T. (2005). Zipf's Law for cities: A cross-country investigation. *Regional Science and Urban Economics*, 35(3):239–263.

Soo, K. T. (2014). Zipf, Gibrat and geography: evidence from China, India and Brazil. *Papers in Regional Science*, 93(1):159–182.

Yankow, J. J. (2006). Why do cities pay more? an empirical examination of some competing theories of the urban wage premium. *Journal of Urban Economics*, 60(2):139–161.

Zanette, D. H. (2015). Zipf's law and city sizes: a short tutorial review of multiplicative processes in urban growth. *Advances in Complex Systems. Forthcoming*.

Table 1: Descriptive statistics of the US data samples used

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 1950	17,113	63.48	5,613	76,064	1	7,891,957
Inc. places 2000	19,296	61.49	8,968	78,015	1	8,008,278
All places 2000	25,358	73.98	8,232	68,390	1	8,008,278
All places 2010	29,461	74.31	7,826	65,494	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 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

Table 2: Estimators and standard errors (SE) of the parameters of the dPln, GB2 and tdPGB2 for the US places samples. The estimators for the lognormal are the mean and the standard deviation of the logarithm of population data

	lgn			
	μ	σ		
Inc. places 1900	6.65	1.26		
Inc. places 1950	6.84	1.50		
Inc. places 2000	7.18	1.78		
All places 2000	7.28	1.75		
All places 2010	7.11	1.82		
	dPln			
	α	β	μ	σ
Inc. places 1900	0.92 (0.01)	2.64 (0.06)	5.95 (0.01)	0.58 (0.01)
Inc. places 1950	0.80 (0.01)	2.15 (0.04)	6.06 (0.01)	0.78 (0.01)
Inc. places 2000	0.87 (0.01)	3.62 (0.09)	6.31 (0.01)	1.36 (0.01)
All places 2000	1.22 (0.01)	3.15 (0.08)	6.78 (0.01)	1.52 (0.01)
All places 2010	1.12 (0.01)	3.03 (0.07)	6.54 (0.01)	1.55 (0.01)
	GB2			
	a	b	p	q
Inc. places 1900	1.879 (0.015)	276 (3)	1.479 (0.018)	0.509 (0.005)
Inc. places 1950	1.242 (0.007)	252 (3)	1.922 (0.017)	0.688 (0.005)
Inc. places 2000	0.370 (0.001)	13 (0)	15.460 (0.067)	3.168 (0.013)
All places 2000	0.1645 (0.0002)	0.0107 (0.0001)	96.358 (0.171)	14.223 (0.025)
All places 2010	0.2480 (0.0005)	8 (0)	22.506 (0.056)	6.813 (0.017)
	tdPGB2			
	ρ	ϵ	a	b
Inc. places 1900	1.88 (0.17)	58 (7)	2.350 (0.030)	242 (3)
Inc. places 1950	1.43 (0.07)	62 (2)	1.063 (0.005)	49 (1)
Inc. places 2000	1.51 (0.07)	43 (4)	0.699 (0.003)	36 (1)
All places 2000	1.52 (0.03)	102 (4)	0.473 (0.002)	223 (3)
All places 2010	1.32 (0.02)	132 (3)	0.497 (0.003)	318 (5)
	tdPGB2			
	p	q	τ	ζ
Inc. places 1900	1.271 (0.018)	0.305 (0.005)	2,472 (227)	1.001 (0.018)
Inc. places 1950	7.882 (0.068)	0.679 (0.006)	15,968 (1,164)	1.082 (0.028)
Inc. places 2000	7.201 (0.055)	0.895 (0.007)	55,595 (2,537)	1.380 (0.043)
All places 2000	4.592 (0.027)	2.182 (0.011)	54,016 (3,094)	1.445 (0.042)
All places 2010	3.068 (0.022)	1.878 (0.010)	56,703 (3,238)	1.430 (0.040)

Table 3: Estimators and standard errors (SE) of the parameters of the dPln and dmPGB2 for the US CCA clusters samples. The estimators for the lognormal are the mean and the standard deviation of the logarithm of population data

	lgn				
	μ	σ			
CCA 1991 (2000m)	8.33	0.85			
CCA 1991 (3000m)	8.32	0.89			
CCA 2000 (2000m)	8.44	0.87			
CCA 2000 (3000m)	8.43	0.91			
	dPln				
	α	β	μ	σ	
CCA 1991 (2000m)	1.95 (0.01)	1.85 (0.01)	8.360 (0.003)	0.138 (0.007)	
CCA 1991 (3000m)	1.76 (0.02)	1.86 (0.02)	8.291 (0.006)	0.11 (0.01)	
CCA 2000 (2000m)	1.86 (0.01)	1.82 (0.01)	8.449 (0.004)	0.183 (0.007)	
CCA 2000 (3000m)	1.66 (0.02)	1.83 (0.02)	8.371 (0.006)	0.16 (0.01)	
	dmPGB2				
	ρ	ϵ	ν	a	b
CCA 1991 (2000m)	0.570 (0.025)	1,470 (37)	0.342 (0.015)	4.317 (0.021)	4,287 (14)
CCA 1991 (3000m)	0.585 (0.032)	1,302 (43)	0.344 (0.018)	4.317 (0.024)	4,184 (15)
CCA 2000 (2000m)	0.556 (0.025)	1,669 (47)	0.293 (0.013)	3.583 (0.017)	4,524 (16)
CCA 2000 (3000m)	0.566 (0.032)	1,374 (45)	0.320 (0.018)	3.756(0.021)	4,358 (17)
	dmPGB2				
	p	q	τ	ζ	θ
CCA 1991 (2000m)	0.595 (0.004)	0.626 (0.004)	17,034 (387)	0.967 (0.031)	0.788 (0.021)
CCA 1991 (3000m)	0.578 (0.004)	0.632 (0.004)	16,178 (302)	0.864 (0.024)	0.894 (0.017)
CCA 2000 (2000m)	0.728 (0.004)	0.685 (0.004)	21,205 (654)	0.931 (0.036)	0.668 (0.026)
CCA 2000 (3000m)	0.663 (0.005)	0.630 (0.004)	20,539 (551)	0.862 (0.027)	0.817 (0.022)

Table 4: p-values (statistics) of the Kolmogorov–Smirnov (KS), Cramér–Von Mises (CM) and Anderson–Darling (AD) tests for the US places and CCA clusters samples and the density functions used. Non rejections at the 5% significance level are in bold

	lgn			dPln		
	KS	CM	AD	KS	CM	AD
Inc. places 1900	0 (0.07)	0 (17.22)	0 (100.47)	**0.17 (0.01)**	**0.11 (0.34)**	**0.10 (1.97)**
Inc. places 1950	0 (0.06)	0 (17.56)	0 (104.90)	0 (0.02)	0 (1.40)	0 (10.48)
Inc. places 2000	0 (0.04)	0 (9.40)	0 (53.66)	0 (0.02)	0 (1.95)	0 (12.63)
All places 2000	0 (0.02)	0 (3.03)	0 (19.12)	0 (0.02)	0 (1.45)	0 (8.98)
All places 2010	0 (0.02)	0 (4.57)	0 (29.48)	0 (0.02)	0 (1.73)	0 (11.73)
	GB2			tdPGB2		
	KS	CM	AD	KS	CM	AD
Inc. places 1900	**0.568 (0.008)**	**0.358 (0.161)**	**0.276 (1.177)**	**0.978 (0.005)**	**0.971 (0.032)**	**0.989 (0.205)**
Inc. places 1950	0.010 (0.013)	0.013 (0.695)	0.002 (5.191)	**0.994 (0.003)**	**0.990 (0.025)**	**0.989 (0.206)**
Inc. places 2000	0.001 (0.015)	0.001 (1.153)	0 (8.403)	**0.986 (0.004)**	**0.974 (0.031)**	**0.986 (0.214)**
All places 2000	0.012 (0.011)	0.012 (0.717)	0.005 (4.482)	**0.969 (0.003)**	**0.936 (0.039)**	**0.971 (0.249)**
All places 2010	0 (0.018)	0 (1.389)	0 (11.639)	**0.899 (0.004)**	**0.814 (0.060)**	**0.769 (0.478)**
	lgn			dPln		
	KS	CM	AD	KS	CM	AD
CCA 1991 (2000m)	0 (0.09)	0 (92.68)	0 (567.01)	0 (0.02)	0 (1.71)	0 (18.47)
CCA 1991 (3000m)	0 (0.10)	0 (89.36)	0 (536.38)	0 (0.03)	0 (2.25)	0 (23.18)
CCA 2000 (2000m)	0 (0.09)	0 (73.24)	0 (450.94)	0 (0.02)	0 (1.18)	0 (13.67)
CCA 2000 (3000m)	0 (0.09)	0 (66.57)	0 (405.72)	0 (0.02)	0.001 (1.26)	0 (13.94)
	dmPGB2					
	KS	CM	AD			
CCA 1991 (2000m)	**0.858 (0.004)**	**0.867 (0.051)**	**0.874 (0.374)**			
CCA 1991 (3000m)	**0.842 (0.004)**	**0.869 (0.051)**	**0.895 (0.351)**			
CCA 2000 (2000m)	**0.876 (0.004)**	**0.817 (0.060)**	**0.817 (0.432)**			
CCA 2000 (3000m)	**0.839 (0.004)**	**0.860 (0.053)**	**0.917 (0.326)**			

Table 5: Values of the msd (in units of 10^{-3}) and of the pseudo R^2 of Duranton (2007) for the US places and CCA clusters samples and the used distributions. The most favoured values are marked in boldface.

	lgn		dPln		GB2		tdPGB2	
	msd	R^2	msd	R^2	msd	R^2	msd	R^2
Inc. places 1900	72.40	0.9545	24.32	0.9847	25.69	0.9838	**0.40**	**0.9997**
Inc. places 1950	127.40	0.9435	16.34	0.9928	11.06	0.9951	**0.39**	**0.9998**
Inc. places 2000	34.00	0.9893	24.13	0.9924	8.42	0.9974	**0.21**	**0.9999**
All places 2000	9.60	0.9969	4.29	0.9986	2.99	0.9990	**0.19**	**0.9999**
All places 2010	15.20	0.9954	7.34	0.9978	6.35	0.9981	**0.39**	**0.9999**
	lgn		dPln		dmPGB2			
	msd	R^2	msd	R^2	msd	R^2		
CCA 1991 (2000m)	103.74	0.8345	40.64	0.9433	**3.70**	**0.9948**		
CCA 1991 (3000m)	116.08	0.8529	45.43	0.9424	**3.10**	**0.9960**		
CCA 2000 (2000m)	94.22	0.8744	36.37	0.9515	**3.22**	**0.9957**		
CCA 2000 (3000m)	105.89	0.8723	39.21	0.9527	**2.37**	**0.9971**		

Table 6: Maximum log-likelihoods, AIC and BIC for the distributions used and the US places and CCA clusters data. The lowest values of AIC and BIC for each sample are in bold

	lgn			dPln		
	log-likelihood	AIC	BIC	log-likelihood	AIC	BIC
Inc. places 1900	-87,943	175,891	175,905	-87,254	174,516	174,545
Inc. places 1950	-148,254	296,512	296,528	-147,593	295,194	295,225
Inc. places 2000	-177,127	354,258	354,274	-176,931	353,870	353,901
All places 2000	-234,773	469,550	469,566	-234,710	469,428	469,461
All places 2010	-268,748	537,499	537,516	-268,657	537,323	537,356
	GB2			tdPGB2		
	log-likelihood	AIC	BIC	log-likelihood	AIC	BIC
Inc. places 1900	-87,246	174,500	174,529	-87,230	**174,476**	**174,535**
Inc. places 1950	-147,566	295,140	295,171	-147,471	**294,958**	**295,020**
Inc. places 2000	-176,871	353,751	353,782	-176,770	**353,556**	**353,619**
All places 2000	-234,680	469,367	469,400	-234,628	**469,272**	**469,337**
All places 2010	-268,616	537,239	537,273	-268,520	**537,056**	**537,122**
	lgn			dPln		
	log-likelihood	AIC	BIC	log-likelihood	AIC	BIC
CCA 1991 (2000m)	-289,460	578,923	578,940	-284,288	568,584	568,617
CCA 1991 (3000m)	-226,140	452,284	452,300	-221,851	443,711	443,743
CCA 2000 (2000m)	-293,311	586,627	586,643	-288,879	577,765	577,798
CCA 2000 (3000m)	-229,171	458,347	458,363	-225,494	450,996	451,028
	dmPGB2					
	log-likelihood	AIC	BIC			
CCA 1991 (2000m)	-283,583	**567,186**	**567,270**			
CCA 1991 (3000m)	-221,216	**442,453**	**442,533**			
CCA 2000 (2000m)	-288,307	**576,634**	**576,717**			
CCA 2000 (3000m)	-225,019	**450,057**	**450,138**			

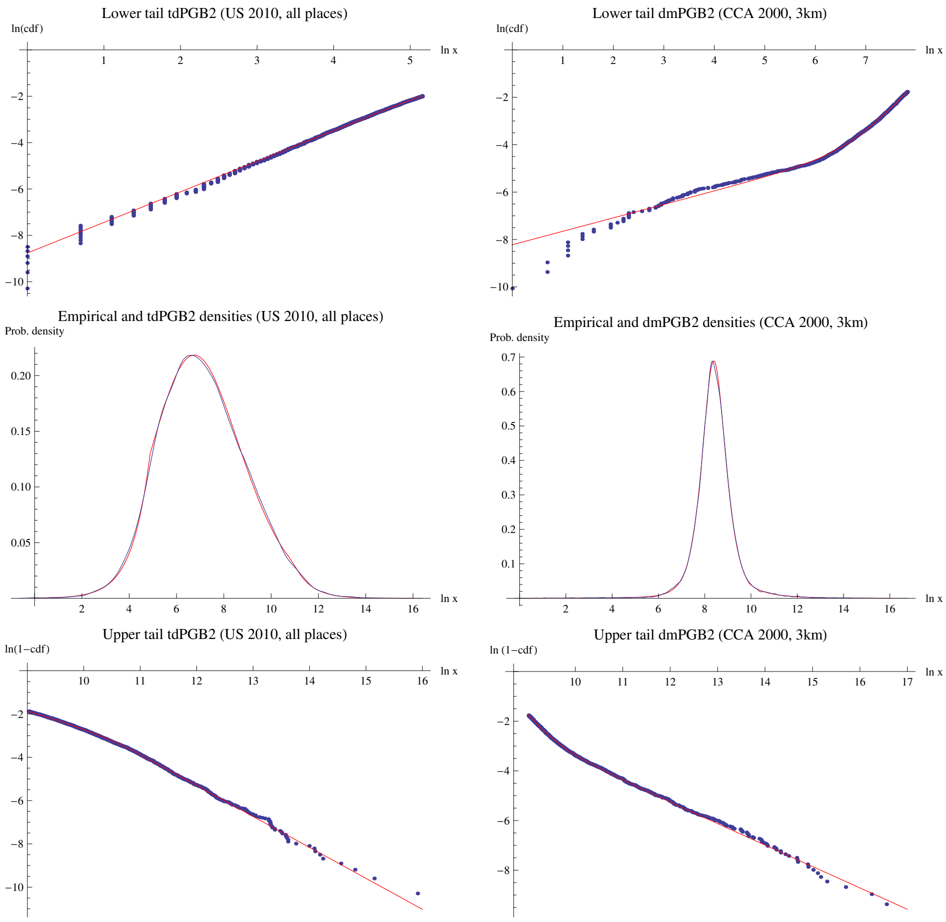

Figure 1: First row: Empirical and estimated tdPGB2 and dmPGB2 $\ln(\text{cdf})$ for the lower tail. Second row: Empirical (Gaussian adaptive kernel density) and estimated tdPGB2 and dmPGB2 density functions. Third row: Empirical and estimated tdPGB2 and dmPGB2 $\ln(1-\text{cdf})$ for the upper tail. First column: US all places (2010). Second column: US CCA clusters (2000, 3km).