On the parametric description of the French, German, Italian and Spanish city size distributions

Puente-Ajovin, Miguel and Ramos, Arturo

Universidad de Zaragoza, Universidad de Zaragoza

12 April 2014

On the parametric description of the French, German, Italian and Spanish city size distributions

MIGUEL PUENTE-AJOVÍN* ARTURO RAMOS†

April 12, 2014

Abstract

We study the parametric description of the city size distribution of four European countries: France, Germany, Italy and Spain. The parametric models used are the lognormal, the double Pareto lognormal, the normal-Box-Cox (defined in this paper) and the threshold double Pareto Singh–Maddala (introduced in a cited recent paper when studying US city size).

The results are quite regular. The preferred model is always the threshold double Pareto Singh–Maddala in the four countries. However, the dPln is not rejected always for the case of France, and in the case of Italy the dPln is the runner-up distribution. These results complement those obtained in a cited recent paper which study the US places' city size distribution.

JEL:C13, C16, R00.

Keywords: European city size distributions, population thresholds, lower and upper tails, new statistical distribution

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

1 Introduction

The study of city size distribution has been, since the contribution of Zipf (1949), of great importance in the field of Urban Economics. The so-called Zipf distribution, or the slightly more general form of Pareto distribution, has been extensively studied by many authors, we recall here, e.g., Black and Henderson (2003), Ioannides and Overman (2003), Soo (2005), Anderson and Ge (2005) and Bosker et al. (2008). More specifically, in recent times have appeared the important contributions of Eeckhout (2004), Giesen et al. (2010) and Ioannides and Skouras (2013). The first of these references introduces the need of considering the whole sample of cities when studying their size distribution, and proposes the lognormal distribution (see also Parr and Suzuki (1973)). The second continues a line of research initiated by Reed (2001, 2002, 2003); Reed and Jorgensen (2004) in which it is introduced the so-called double Pareto lognormal (dPln) distribution in the study of city size. This distribution has Pareto tails mixed (by means of a convolution) with a lognormal body and offers a good fit to the data, see Giesen et al. (2010) and also González-Val et al. (2013c). In turn, Ioannides and Skouras (2013) propose two distributions which have lognormal body and, above a certain *exact* threshold, a Pareto upper tail mixed or not (by means of a linear combination) with the lognormal. These two recently proposed distributions still do not outperform the dPln for US places in the year 2000, as Giesen and Suedekum (2013) indicate. In order to reconcile both tendencies, the recent work of Ramos et al. (2014) studies the US city size distribution with three types of data (incorporated, all places and CCA clusters), comparing the previously mentioned distributions and newly introduced ones. One of the main results of this last paper is that the parametric description of the size distribution of US places can be safely taken as a new one, called "threshold double Pareto Singh-Maddala" (tdPSM), which is a distribution with Pareto behavior in the lower and upper tails, and Singh-Maddala body. The transition between the tails and the body takes place at two exact thresholds, to be determined endogenously by the maximum likelihood (ML) estimation procedure. The new tdPSM greatly outperforms the lognormal, the dPln, and the distributions of Ioannides and Skouras (2013) in the case of US places.

In the previous articles of González-Val et al. (2013a,c) it has been used city population data of France, Italy and Spain without size restriction. In turn, Schluter and Trede (2013) use a dataset of all German municipalities or *Gemeinden* and propose a composition of the normal distribution with a Box-Cox transformation of the population data, with apparently quite good results. This will lead to a distribution which we will call normal-Box-Cox (nBC), to be defined below.

Thus it is our aim in this article to compare the lognormal, the dPln, the nBC and the tdPSM distributions for, generally decennial, samples of city size data of France, Germany, Italy and Spain without size restriction. The main results, which we advance here, is that the tdPSM is the preferred distribution almost always, and is a model clearly not rejected by the statistical tests we use below. Also, we obtain that the dPln is not rejected always in the case of France 1990-2009 and is very accurate as well in the parametric description of the size of Italian *comuni* in the period 1901-2011. We obtain thus a strong result: the parametric distribution of the city size of these four

European countries can be taken as the same as that of US places, namely the tdPSM. In all cases, the Pareto nature of the two tails seems to be an essential feature to be taken into account.

We will rely heavily on the previous paper of Ramos et al. (2014) so, for the sake of brevity and in order to avoid excessive repetitions, we will concentrate on the new results. The rest of the paper is organized as follows. Section 2 describes the databases used in this paper. Section 3 shows the definitions and main properties of the four distributions studied. Section 4 shows the detailed results, country by country. Finally, Section 5 concludes.

2 The databases

In this article we use population data, without size restriction, of four European countries: France, Germany, Italy and Spain.

For the case of France, as in González-Val et al. (2013a), we consider the lowest spatial subdivision, the *communes*, as listed by the *Institut national de la statistique et des études économiques* (`www.insee.fr`). We have data for the years 1990, 1999 and 2009. Note that Giesen and Suedekum (2012) use this kind of data for the year 2008.

For the case of Germany, Italy and Spain, the administrative urban unit of the data is the municipality (*Gemeinden* for the case of Germany). For Germany, we take data from two sources. The first is the data used in Schluter and Trede (2013), which has been kindly provided to us by Prof. Trede (the original source is the Federal German Statistical Office). We take the data of the years 1996 and 2006 in order to comprise a decennial period similarly to the data of the other considered countries. The second source is, directly, the cited statistical office through its web page `www.Destatis.de`. We use the data of the last available year 2011 for comparison purposes. For Italy, the data is obtained from the *Istituto Nazionale di Statistica* (`www.istat.it`), with all the Italian municipalities (*comuni*) for the period 1901-2011. We have used the Italian census for 1936 instead of 1941 because of the participation of Italy in the Second World War. The data for Spain is taken from the *Instituto Nacional de Estadística* (`www.ine.es`). They cover all the municipalities (*municipios*) along the period 1900-2010.

[Table 1 near here]

We offer in Table 1 the descriptive statistics of the used data for France, Germany, Italy and Spain. The information for Italy and Spain is the same as that in Table 1 of González-Val et al. (2013c).

3 Description of the distributions used

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

3.1 The lognormal distribution (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* the cities. The corresponding density is simply

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

where $\mu, \sigma > 0$ are respectively the mean and the standard deviation of $\ln x$. This is the first distribution we will consider in this study.

3.2 The double Pareto lognormal distribution (dPln)

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

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

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 González-Val et al. (2013c).

3.3 The normal-Box-Cox (nBC)

In the article of Schluter and Trede (2013) it has been proposed the idea of composing the normal distribution with the well-known Box-Cox transformation for German city data. We include the distribution so obtained (normal-Box-Cox, nBC) in our study because it turns out that the nBC provides good results in the case of Germany.

The Box-Cox transformation is given by the well-known expression (Box and Cox, 1964)

$$g_\lambda(x) = \begin{cases} \dfrac{x^\lambda-1}{\lambda} & \text{if} \quad \lambda \neq 0 \\ \ln x & \text{if} \quad \lambda = 0 \end{cases}$$

The composition with the normal will be $g'_\lambda(x) f_{\rm n}(g_\lambda(x), \mu, \sigma)$, where $g'_\lambda(x)$ is the derivative of $g_\lambda(x)$ with respect to x and

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

is the normal density function. The case with $\lambda = 0$ leads to the lognormal, introduced in Subsection 3.1 and treated separately. Thus, for the case of $\lambda \neq 0$ we define the normal-Box-Cox (nBC) as the density

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

which has support $x \in (0, \infty)$. The quantities μ and σ are, respectively, the mean and standard deviation of $\dfrac{x^\lambda - 1}{\lambda}$.

3.4 The threshold double Pareto Singh–Maddala (tdPSM)

We recall here a distribution introduced in Ramos et al. (2014) when studying US city size, the so-called therein "threshold double Pareto Singh-Maddala" (tdPSM). It provided the best results, amongst the studied parametric distributions, for US places and in principle, it is our best candidate in the current study. The characteristics of this distribution are that the lower and upper tails are Pareto and the body is Singh-Maddala. The switch between the tails and the body occurs at two *exact* thresholds: $\epsilon > 0$ separates the lower tail from the body, and $\tau > \epsilon$ the body from the upper tail.

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

$$f_{\rm SM}(x, \mu, \sigma, \alpha) = \frac{\alpha\,(\mathrm{e}^{-\mu}x)^{1/\sigma}}{x\sigma(1 + (\mathrm{e}^{-\mu}x)^{1/\sigma})^{1+\alpha}} \tag{3}$$

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

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

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

Imposing continuity at the threshold points and overall normalization to unity, the

[1] The $f_{\rm SM}$ is directly related to the Burr Type XII distribution (Burr, 1942). See also Kleiber and Kotz (2003).

composite resulting density is (see Ramos et al. (2014) for details):

$$f_4(x,\rho,\epsilon,\mu,\sigma,\alpha,\tau,\zeta) = \begin{cases} b_4\, e_4\, l(x,\rho) & 0 < x < \epsilon \\ b_4\, f_{\rm SM}(x,\mu,\sigma,\alpha) & \epsilon \leq x \leq \tau \\ b_4\, a_4\, u(x,\zeta) & \tau < x \end{cases} \tag{6}$$

where

$$e_4 = \frac{f_{\rm SM}(\epsilon,\mu,\sigma,\alpha)}{l(\epsilon,\rho)} \tag{7}$$

$$a_4 = \frac{f_{\rm SM}(\tau,\mu,\sigma,\alpha)}{u(\tau,\zeta)} \tag{8}$$

$$b_4^{-1} = e_4\frac{\epsilon^\rho}{\rho} + {\rm e}^{\mu\alpha/\sigma}(({\rm e}^{\mu/\sigma}+\epsilon^{1/\sigma})^{-\alpha} - ({\rm e}^{\mu/\sigma}+\tau^{1/\sigma})^{-\alpha}) + \frac{a_4}{\zeta\,\tau^\zeta} \tag{9}$$

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

4 Results

For the sake of brevity, we will present the results country by country, and refer to Ramos et al. (2014) for a more detailed explanation of the maximum log-likelihood (ML) estimation, Kolmogorov–Smirnov (KS) and Cramér–Von Mises (CM) tests and AIC, BIC information criteria.

4.1 Results for France

We show in Table 2 the ML estimators of the studied distributions for the 1990, 1999 and 2009 French samples of *communes*. For the lognormal (lgn) the ML estimators are exact and equal to the mean and standard deviation of the log-population data. For the other three distributions (dPln, nBC and tdPSM) we provide the ML estimators and 95% confidence intervals.[2] The estimations appear to be rather precise in all cases.

[Table 2 near here]

In Table 3 we show the results of the KS and CM tests for the hypothesized distributions. These two tests are very powerful when the sample size is high or very high (Razali and Wah, 2011) as in our French samples, and non-rejections only occur if the deviations (statistics) are really small. We observe that the lgn is strongly rejected in all cases, and the nBC is rejected always as well, although with lower values of the tests' statistics. In turn, the dPln is not rejected by both tests 100% of the cases. And the tdPSM is not rejected always, too. The tests' statistics are always slightly lower for the tdPSM than for the dPln. According to these tests, the French *communes* size distribution can be taken as the excellent dPln, or even better, as the tdPSM. The excellent fit of the dPln for the French *communes* in the year 2008 has been anticipated by Giesen and Suedekum (2012).

[2] We have performed the estimations with MATLAB as in González-Val et al. (2013c) and Ramos et al. (2014).

[Table 3 near here]

In order to choose one of the hypothesized models according to information criteria, we show in Table 4 the results of the AIC and BIC, which are specially well suited to the maximum likelihood estimation we have performed before. Both of AIC and BIC favour the distribution with greater maximum likelihood, but there is a penalty for the number of parameters used in the distribution. The distribution with lowest AIC and/or BIC is preferred.

[Table 4 near here]

For the case of the French samples we observe that the lgn obtains always the greatest values of AIC and BIC, and that the lowest AIC and BIC occurs for the tdPSM in all cases. This result, jointly with the outcomes of the KS and CM tests yields that the French *communes* size distribution, can be very well described parametrically by our tdPSM, outperforming the dPln.

4.2 Results for Germany

We carry on now a similar analysis for our 1996, 2006, 2011 German samples of *Gemeinden*. First, we show in Table 5 the estimation results. The estimations are rather precise in this case as well. The obtained estimations of the parameter λ for the years 1996 and 2006 are consistent with the results of Schluter and Trede (2013).

[Table 5 near here]

In Table 6 we show the results of the KS and CM tests. The lgn, dPln and nBC are (strongly) rejected in all cases. In contrast, the tdPSM is not rejected always, with values of the tests' statistics really small.

[Table 6 near here]

The results of the AIC and BIC information criteria are shown in Table 7. The lgn is the less preferred distribution always. Note as well that the nBC is preferred always to the dPln for the German samples. However, it is shown clearly that the preferred model (out of those studied in this paper) is the tdPSM in all instances and by both information criteria. Jointly with the results of the KS and CM tests, we conclude that the German city size distribution of *Gemeinden*, without size restriction, can be safely taken as the tdPSM.

[Table 7 near here]

4.3 Results for Italy

We have performed as well the ML estimation of the Italian samples of *comuni* in the period 1901-2011. The results are not shown here for the sake of brevity but are available from the authors upon request. We concentrate on the statistical tests and information criteria.

In Table 8 we show the results of the KS and CM tests for our four hypothesized distributions. The lgn is rejected always except in 2011. The dPln is not rejected always. The nBC is not rejected for the years 1981, 1991, 2001 and 2011. And the

tdPSM is not rejected always as well. The lowest values of the tests' statistics for the dPln and tdPSM alternate over time. Thus, it follows that the dPln and tdPSM are close competitors for the parametric description of Italian *comuni* size in the period 1901-2011.

[Table 8 near here]

We show in Table 9 the results of the AIC and BIC for the Italian samples. We obtain mixed results: according to the lowest AIC, the tdPSM is the preferred model most of the time (83.33% of the cases). Likewise, according to the lowest BIC, the dPln is the preferred model also 83.33% of the cases. And for both distributions there are two cases in which one or the other is clearly selected (1991, 2001 for the dPln and 1951, 1961 for the tdPSM). In the case of discrepancy of the outcomes of the AIC and BIC information criteria we follow Burnham and Anderson (2002, 2004) in preferring those of the AIC, based on theoretical and simulation arguments. Thus, for the Italian *comuni* without size restriction we obtain two excellent competing parametric models: the dPln and the tdPSM, with a preference for the second.

[Table 9 near here]

4.4 Results for Spain

Again, we have estimated the four distributions studied in this paper by ML for the samples of Spanish *municipios* in the period 1900-2010. The results are not shown for the sake of brevity but are available from the authors upon request. We concentrate on the results of the KS, CM tests and AIC, BIC criteria.

We show in Table 10 the results of the KS and CM tests. The lgn and the dPln are strongly rejected always. The nBC is rejected in almost all cases, with the exception of the KS test in 1981. In turn, the tdPSM is not rejected always, and with values of the tests' statistics quite low. The tdPSM reveals itself as a very good model for the Spanish city size.

[Table 10 near here]

In Table 11 we show the values of the AIC and BIC information criteria. The nBC is preferred to the dPln always for the Spanish *municipios*. And clearly, the preferred distribution is always the tdPSM for the whole period 1900-2010 of these urban units without size restriction. In short, the Spanish city size along the period 1900-2010 can be safely described in a parametric way by the tdPSM.

[Table 11 near here]

4.5 An informal graphical approximation

The use of graphical tools in assessing the fit of parametric distributions to empirical data has certain shortcomings to be taken into account, see, e.g., González-Val et al. (2013b). In this reference it has been shown that when representing the differences of the empirical and estimated $\ln(1-\mathrm{cdf})$'s, where cdf is the relevant cumulative density function, an amplification effect of the differences of the cdf's is obtained for the upper tail. A similar effect occurs for the $\ln(\mathrm{cdf})$'s and the lower tail. The amplification

effect increases as we approach infinity for the upper tail or zero for the lower tail, and it is difficult to quantify.

Also, the goodness-of-fit, as tested by the KS and CM tests, is strongly dependent on the number of observations in the sample. The graphical fit does not take into account, in an essential way, the number of observations.

We offer for completeness in this subsection some graphs corresponding to the studied cases. For France, we have taken the three samples and the best distribution obtained, the tdPSM. For Germany as well we present the graphs of the three samples used and the best parametric model, also the tdPSM. For Italy, we take the sample of 2001 and the corresponding chosen distribution by the information criteria, namely the dPln. For Spain, we take the sample of 1981, in which the tdPSM is specially well suited according to the very low statistics of the KS and CM tests.

For the French samples and the tdPSM the graphical results are also excellent: the fit in the lower tails is remarkable, and in the upper tails as well, maybe except for the biggest *communes*. For the densities, there are small discrepancies near the mode of the theoretical distributions.

For the German samples and the tdPSM, the lower tails of the 1996 and 2011 samples show slight discrepancies but for 2006 the lower tail fit is remarkable. For the upper tails the fit is visually excellent, and in particular for the four most populated German *Gemeinden*, namely Berlin, Hamburg, München and Köln, the fit is practically perfect for the three samples. The densities show very slight discrepancies but in a framework of overall excellent fit.

In the Italian case of 2001 and the dPln we observe some slight discrepancies in the lower tail and the six biggest cities in the upper tail deviate slightly from the estimated parametric model. However, the fit of the densities is visually excellent.

For the Spanish sample of 1981 and the tdPSM we observe an excellent fit in the lower tail. The upper tail fit is excellent with the possible exception of the biggest cities and the fit of the densities is remarkable.

In short, the graphical approximation in the selected cases by our formal criteria yields visually excellent fits in all the cases, with very slight discrepancies, if any, at the ends of the lower or upper tails, or at the mode of the theoretical densities.

[Figure 1 near here]

[Figure 2 near here]

[Figure 3 near here]

5 Conclusions

In this paper we have used population data corresponding to the lowest spatial subdivision of four European countries: France, Germany, Italy and Spain in different periods of the last and this centuries. We have used the data to study the parametric fit of four density functions: the lognormal (lgn) (Parr and Suzuki, 1973; Eeckhout, 2004), the double Pareto lognormal (dPln) (Reed, 2001, 2002, 2003; Reed and Jorgensen, 2004), the normal-Box-Cox (nBC) (Schluter and Trede, 2013) and the threshold double Pareto

Singh–Maddala (tdPSM) (Ramos et al., 2014).

We have estimated the four density functions by maximum likelihood (ML) for all the samples and have performed Kolmogorov–Smirnov (KS) and Cramér–Von Mises (CM) tests. We have studied as well the distributions according to the AIC and BIC information criteria.

The results are quite regular across different countries and periods. The tdPSM model is clearly the preferred model for the case of France, Germany and Spain, according to the lowest values of AIC and BIC, and the non-rejection of it by both KS and CM tests in the 100% of the cases. However, for France the dPln, although is not the preferred model, is not rejected always by the cited tests. For Italy the results are mixed: both of the dPln and tdPSM are not rejected always 100% of the cases, and according to the lowest AIC, the tdPSM is preferred in most cases. However, if one takes the lowest BIC, the dPln is preferred in most cases. Thus, for Italy the dPln and tdPSM offer quite similar performance, although we prefer the tdPSM according to the lowest values of AIC in case of discrepancy with the outcome of the BIC. As a side result, we have obtained as well that the nBC is always a preferred model to the dPln in the case of German and Spanish samples.

In all cases we see that the tdPSM model is the best studied model or amongst the best studied models (if one admits a preference for the outcomes of the BIC information criterium over the AIC), which conforms a quite strong empirical regularity for these European countries. In Ramos et al. (2014) it is shown that the same model is the one selected (by the same methodology) for the case of US places in the period 1900-2010, which is a surprising strong regularity for countries with, in principle quite different, historical processes of urbanization and different definitions of the urban units under study. The tdPSM implements Pareto upper and lower tails, and this feature seems to be essential in obtaining an excellent overall fit in all of the studied countries.

This suggests to study further the underlying processes behind the evolution of city size, comparing the US and these European countries, in order to obtain, if possible, further common regularities. We leave this for future research work.

Acknowledgements

This work is supported 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.

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.

Box, G. and Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26:211–252.

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.

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

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.

González-Val, R., Lanaspa, L., and Sanz-Gracia, F. (2013a). Gibrat's law for cities, growth regressions and sample size. *Economics Letters*, 118:367–369.

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

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.

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

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

Ramos, A., Sanz-Gracia, F., and González-Val, R. (2014). A new framework for US city size distribution: Empirical evidence and theory. *Working Paper available at Munich RePEC*, http://mpra.ub.uni-muenchen.de/53324/.

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.

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*.

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.

Zipf, G. (1949). *Human behavior and the principle of least effort*. Cambridge, Massachusetts: Addison-Wesley Press.

Table 1: Descriptive statistics of the samples of French, German, Italian and Spanish urban units used

	Urban units	Mean of pop.	SD of pop.	Minimum	Maximum
France					
1990	36,686	1,610	7,694	1	365,933
1999	36,685	1,677	7,886	1	398,423
2009	36,716	1,791	8,253	1	447,396
Germany					
1996	14,559	5,633	40,608	2	3,458,763
2006	12,312	6,686	44,043	7	3,404,037
2011	11,292	7,114	45,415	10	3,326,002
Italy					
1901	7,711	4,275	14,425	56	621,213
1911	7,711	4,648	17,393	58	751,211
1921	8,100	4,864	20,032	58	859,629
1931	8,100	5,067	22,560	93	960,660
1936	8,100	5,234	25,274	116	1,150,338
1951	8,100	5,866	31,138	74	1,651,393
1961	8,100	6,250	39,131	90	2,187,682
1971	8,100	6,684	45,582	51	2,781,385
1981	8,100	6,982	45,329	32	2,839,638
1991	8,100	7,010	42,450	31	2,775,250
2001	8,100	7,021	39,325	33	2,546,804
2011	8,094	7,490	41,505	34	2,761,477
Spain					
1900	7,800	2,282	10,178	78	539,835
1910	7,806	2,452	11,217	92	599,807
1920	7,812	2,622	13,501	82	750,896
1930	7,875	2,892	17,514	79	1,005,565
1940	7,896	3,181	20,100	11	1,088,647
1950	7,901	3,480	26,033	64	1,618,435
1960	7,910	3,802	33,652	51	2,259,931
1970	7,956	4,241	43,972	10	3,146,071
1981	8,034	4,701	45,995	5	3,188,297
1991	8,077	4,882	45,220	2	3,084,673
2001	8,077	5,039	43,079	7	2,938,723
2010	8,114	7,795	47,530	5	3,273,049

Table 2: Mean and standard deviation of the log-population of the French samples. Estimators and 95% confidence intervals of the parameters of the dPln, the nBC and tdPSM for the French samples

France	lgn		dPln				
	μ	σ	α	β	μ	σ	
1990	6.07	1.34	0.97±0.02	2.85±0.31	5.38±0.04	0.80±0.03	
1999	6.11	1.35	0.97±0.02	2.98±0.38	5.42±0.04	0.83±0.03	
2009	6.21	1.35	1.00±0.02	3.32±0.25	5.52±0.03	0.88±0.02	
	nBC						
	μ	σ	λ				
1990	3.95±0.06	0.53±0.02	-0.15±0.01				
1999	4.01±0.06	0.54±0.02	-0.14±0.01				
2009	4.14±0.07	0.56±0.02	-0.14±0.01				
	tdPSM						
	ρ	ϵ	μ	σ	α	τ	ζ
1990	2.16±0.08	67±0	5.37±0.05	0.62±0.02	0.58±0.03	56,259±1,237	1.78±0.25
1999	2.18±0.08	68±0	5.43±0.05	0.64±0.02	0.60±0.04	52,947±1,172	1.77±0.23
2009	2.16±0.08	77±0	5.62±0.05	0.69±0.03	0.68±0.04	57,969±1,311	1.78±0.24

Table 3: Results of the Kolmogorov–Smirnov (KS) and Cramér–Von Mises (CM) tests for French samples and the used density functions. Non-rejections are marked in boldface

	lgn		dPln	
France	KS	CM	KS	CM
1990	0 (0.05)	0 (27.87)	**0.054 (0.0082)**	**0.17 (0.27)**
1999	0 (0.05)	0 (23.21)	**0.10 (0.0075)**	**0.15 (0.28)**
2009	0 (0.04)	0 (18.55)	**0.18 (0.0067)**	**0.16 (0.28)**
	nBC		tdPSM	
	KS	CM	KS	CM
1990	0 (0.014)	0 (2.00)	**0.14 (0.0071)**	**0.26 (0.206)**
1999	0 (0.012)	0.005 (0.998)	**0.26 (0.0062)**	**0.54 (0.109)**
2009	0 (0.011)	0.009 (0.815)	**0.56 (0.0048)**	**0.63 (0.091)**

Table 4: Maximum log-likelihoods, AIC and BIC for French samples. The lowest values of AIC and BIC for each sample are marked in boldface

	lgn			dPln		
France	log-likelihood	AIC	BIC	log-likelihood	AIC	BIC
1990	-285,530	571,063	571,080	-284,137	568,281	568,315
1999	-287,456	574,915	574,932	-286,161	572,331	572,365
2009	-291,228	582,460	582,477	-290,114	580,236	580,270
	nBC			tdPSM		
	log-likelihood	AIC	BIC	log-likelihood	AIC	BIC
1990	-284,273	568,552	568,578	-284,098	**568,209**	**568,269**
1999	-286,268	572,541	572,567	-286,113	**572,240**	**572,299**
2009	-290,183	580,372	580,398	-290,075	**580,164**	**580,224**

Table 5: Mean and standard deviation of the log-population of the German samples. Estimators and 95% confidence intervals of the parameters of the dPln, the nBC and tdPSM for the German samples

Germany	lgn		dPln				
	μ	σ	α	β	μ	σ	
1996	7.18	1.49	0.92±0.02	4.74±0.01	6.30±0.01	1.05±0.01	
2006	7.43	1.50	1.18±0.03	4.11±0.01	6.82±0.01	1.21±0.01	
2011	7.51	1.51	1.34±0.05	3.82±0.01	7.03±0.01	1.29±0.01	
	nBC						
	μ	σ	λ				
1996	4.78±0.14	0.62±0.04	-0.12±0.01				
2006	5.61±0.19	0.84±0.06	-0.08±0.01				
2011	6.06±0.22	0.97±0.07	-0.06±0.01				
	tdPSM						
	ρ	ϵ	μ	σ	α	τ	ζ
1996	2.16±0.11	199±1	5.78±0.16	0.69±0.12	0.35±0.10	13,872±336	1.26±0.07
2006	1.95±0.14	164±1	6.07±0.13	0.61±0.06	0.26±0.06	13,029±190	1.28±0.06
2011	1.89±0.15	151±1	6.21±0.17	0.67±0.07	0.27±0.07	12,846±298	1.30±0.06

Table 6: Results of the Kolmogorov–Smirnov (KS) and Cramér–Von Mises (CM) tests for German samples and the used density functions. Non-rejections are marked in boldface

	lgn		dPln	
Germany	KS	CM	KS	CM
1996	0 (0.04)	0 (8.82)	0 (0.02)	0 (3.05)
2006	0 (0.03)	0 (3.00)	0 (0.02)	0 (1.51)
2011	0 (0.03)	0 (1.86)	0 (0.02)	0.004 (1.02)
	nBC		tdPSM	
	KS	CM	KS	CM
1996	0 (0.02)	0 (1.82)	**0.76 (0.0059)**	**0.84 (0.056)**
2006	0.01 (0.01)	0.01 (0.75)	**0.96 (0.0048)**	**0.95 (0.037)**
2011	0.02 (0.015)	0.03 (0.56)	**0.87 (0.0059)**	**0.96 (0.033)**

Table 7: Maximum log-likelihoods, AIC and BIC for German samples. The lowest values of AIC and BIC for each sample are marked in boldface

	lgn			dPln		
Germany	log-likelihood	AIC	BIC	log-likelihood	AIC	BIC
1996	-130,962	261,928	261,944	-130,697	261,402	261,432
2006	-113,895	227,795	227,810	-113,803	227,615	227,645
2011	-105,474	210,952	210,967	-105,426	210,860	210,889
	nBC			tdPSM		
	log-likelihood	AIC	BIC	log-likelihood	AIC	BIC
1996	-130,634	261,275	261,297	-130,506	**261,026**	**261,079**
2006	-113,775	227,556	227,578	-113,729	**227,471**	**227,523**
2011	-105,411	210,828	210,850	-105,382	**210,777**	**210,828**

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

	lgn		dPln	
Italy	KS	CM	KS	CM
1901	0 (0.03)	0 (2.42)	**0.40 (0.0106)**	**0.34 (0.167)**
1911	0 (0.03)	0 (2.42)	**0.26 (0.0119)**	**0.42 (0.142)**
1921	0 (0.03)	0 (2.24)	**0.21 (0.0122)**	**0.34 (0.167)**
1931	0 (0.03)	0.02 (1.88)	**0.10 (0.0140)**	**0.29 (0.190)**
1936	0 (0.03)	0 (1.66)	**0.21 (0.0122)**	**0.30 (0.184)**
1951	0 (0.03)	0 (1.59)	**0.11 (0.0140)**	**0.18 (0.254)**
1961	0 (0.03)	0 (2.10)	**0.16 (0.0129)**	**0.20 (0.239)**
1971	0 (0.03)	0 (2.05)	**0.11 (0.0140)**	**0.43 (0.138)**
1981	0 (0.02)	0 (1.52)	**0.52 (0.0094)**	**0.84 (0.056)**
1991	0.002 (0.02)	0.006 (0.94)	**0.83 (0.0072)**	**0.94 (0.039)**
2001	0.005 (0.02)	0.008 (0.84)	**0.94 (0.0061)**	**0.99 (0.024)**
2011	**0.10 (0.014)**	**0.06 (0.43)**	**0.98 (0.0055)**	**0.95 (0.036)**

	nBC		tdPSM	
	KS	CM	KS	CM
1901	0 (0.02)	0 (0.98)	**0.81 (0.0075)**	**0.91 (0.044)**
1911	0 (0.02)	0.01 (0.80)	**0.94 (0.0063)**	**0.95 (0.037)**
1921	0.01 (0.02)	0.02 (0.65)	**0.87 (0.0069)**	**0.98 (0.030)**
1931	0 (0.02)	0.02 (0.59)	**0.23 (0.0120)**	**0.48 (0.125)**
1936	0.01 (0.019)	0.01 (0.70)	**0.73 (0.0080)**	**0.83 (0.058)**
1951	0 (0.02)	0.01 (0.88)	**0.64 (0.0086)**	**0.84 (0.057)**
1961	0 (0.02)	0.01 (0.88)	**0.87 (0.0069)**	**0.89 (0.048)**
1971	0.02 (0.018)	0.03 (0.55)	**0.45 (0.0099)**	**0.96 (0.035)**
1981	**0.06 (0.015)**	**0.08 (0.38)**	**0.69 (0.0082)**	**0.98 (0.029)**
1991	**0.17 (0.013)**	**0.14 (0.29)**	**0.81 (0.0074)**	**0.89 (0.047)**
2001	**0.44 (0.010)**	**0.27 (0.20)**	**0.70 (0.0082)**	**0.98 (0.028)**
2011	**0.85 (0.007)**	**0.59 (0.10)**	**0.99 (0.0052)**	**0.95 (0.038)**

Table 9: Maximum log-likelihoods, AIC and BIC for Italian samples. The lowest values of AIC and BIC for each sample are marked in boldface

	lgn			dPln		
Italy	log-likelihood	AIC	BIC	log-likelihood	AIC	BIC
1901	-70,325	140,654	140,668	-70,148.4	140,305	**140,333**
1911	-70,871.9	141,748	141,762	-70,698.2	141,404	**141,432**
1921	-74,657.4	149,319	149,333	-74,474.5	148,957	**148,985**
1931	-74,918.2	149,840	149,854	-74,757.6	149,523	**149,551**
1936	-75,091.6	150,187	150,201	-74,942.3	149,893	**149,921**
1951	-75,830.9	151,666	151,680	-75,689.6	151,387	151,415
1961	-75,836.7	151,677	151,691	-75,675.3	151,359	151,387
1971	-75,951.9	151,908	151,922	-75,798	151,604	**151,632**
1981	-76,390.6	152,785	152,799	-76,284.1	152,576	**152,604**
1991	-76,653.1	153,310	153,324	-76,583.2	**153,174**	**153,202**
2001	-76,865.2	153,734	153,748	-76,818.1	**153,644**	**153,672**
2011	-77,390.1	154,784	154,798	-77,359.4	154,727	**154,755**

	nBC			tdPSM		
	log-likelihood	AIC	BIC	log-likelihood	AIC	BIC
1901	-70,201.5	140,409	140,430	-70,138.8	**140,292**	140,340
1911	-70,743.5	141,493	141,514	-70,689.1	**141,392**	141,441
1921	-74,511.5	149,029	149,050	-74,465.6	**148,945**	148,994
1931	-74,786	149,578	149,599	-74,747.9	**149,510**	149,559
1936	-74,973.1	149,952	149,973	-74,931.2	**149,876**	149,925
1951	-75,719.6	151,445	151,466	-75,672.3	**151,359**	**151,408**
1961	-75,702.3	151,411	151,432	-75,659.9	**151,334**	**151,383**
1971	-75,818	151,642	151,663	-75,791	**151,596**	151,645
1981	-76,297.1	152,600	152,621	-76,278.6	**152,571**	152,620
1991	-76,594.1	153,194	153,215	-76,580.5	153,175	153,224
2001	-76,827.6	153,661	153,682	-76,815.4	153,645	153,694
2011	-77,365.7	154,737	154,758	-77,356	**154,726**	154,775

Table 10: Results of the Kolmogorov–Smirnov (KS) and Cramér–Von Mises (CM) tests for Spanish samples and the used density functions. Non-rejections are marked in boldface

	lgn		dPln	
Spain	KS	CM	KS	CM
1900	0 (0.06)	0 (7.13)	0 (0.03)	0 (1.46)
1910	0 (0.05)	0 (6.42)	0 (0.03)	0 (1.72)
1920	0 (0.06)	0 (7.23)	0 (0.03)	0 (1.76)
1930	0 (0.05)	0 (7.27)	0 (0.03)	0 (2.07)
1940	0 (0.05)	0 (6.75)	0 (0.03)	0 (1.94)
1950	0 (0.06)	0 (7.43)	0 (0.03)	0 (2.01)
1960	0 (0.06)	0 (7.15)	0 (0.03)	0 (2.38)
1970	0 (0.05)	0 (5.48)	0 (0.03)	0 (1.37)
1981	0 (0.05)	0 (4.51)	0.001 (0.02)	0.002 (1.14)
1991	0 (0.05)	0 (4.91)	0 (0.02)	0 (1.39)
2001	0 (0.05)	0 (6.21)	0 (0.03)	0 (2.20)
2010	0 (0.05)	0 (5.17)	0 (0.03)	0 (2.76)
	nBC		tdPSM	
	KS	CM	KS	CM
1900	0 (0.02)	0.01 (0.85)	**0.45 (0.0101)**	**0.72 (0.076)**
1910	0 (0.02)	0 (1.23)	**0.92 (0.0065)**	**0.89 (0.048)**
1920	0 (0.02)	0.01 (0.91)	**0.83 (0.0073)**	**0.93 (0.040)**
1930	0 (0.02)	0 (1.51)	**0.55 (0.0093)**	**0.84 (0.056)**
1940	0 (0.02)	0 (1.20)	**0.93 (0.0063)**	**0.93 (0.041)**
1950	0 (0.02)	0 (1.23)	**0.90 (0.0067)**	**0.89 (0.048)**
1960	0 (0.02)	0 (1.34)	**0.71 (0.0081)**	**0.96 (0.035)**
1970	0 (0.02)	0.01 (0.72)	**0.46 (0.0099)**	**0.87 (0.052)**
1981	**0.06 (0.015)**	0.04 (0.50)	**0.95 (0.0060)**	**0.98 (0.029)**
1991	0.02 (0.017)	0.01 (0.82)	**0.88 (0.0068)**	**0.93 (0.040)**
2001	0.02 (0.022)	0 (1.29)	**0.81 (0.0074)**	**0.69 (0.080)**
2010	0 (0.03)	0 (1.82)	**0.51 (0.0095)**	**0.53 (0.111)**

Table 11: Maximum log-likelihoods, AIC and BIC for Spanish samples. The lowest values of AIC and BIC for each sample are marked in boldface

	lgn			dPln		
Spain	log-likelihood	AIC	BIC	log-likelihood	AIC	BIC
1900	-65,873.6	131,751	131,765	-65,627.3	131,263	131,290
1910	-66,413.5	132,831	132,845	-66,169.3	132,347	132,374
1920	-66,762.6	133,529	133,543	-66,520.7	133,049	133,077
1930	-67,782.4	135,569	135,583	-67,552.4	135,113	135,141
1940	-68,291.6	136,587	136,601	-68,042.6	136,093	136,121
1950	-68,656.2	137,316	137,330	-68,403.7	136,815	136,843
1960	-68,762	137,528	137,542	-68,514.3	137,037	137,065
1970	-68,529.4	137,063	137,077	-68,341.7	136,691	136,719
1981	-68,568.1	137,140	137,154	-68,424.2	136,856	136,884
1991	-68,592.2	137,188	137,202	-68,453.7	136,915	136,943
2001	-68,833.3	137,671	137,685	-68,687.1	137,382	137,410
2010	-69,911.2	139,826	139,840	-69,795.7	139,599	139,627
	nBC			tdPSM		
	log-likelihood	AIC	BIC	log-likelihood	AIC	BIC
1900	-65,579.8	131,166	131,186	-65,536.8	**131,088**	**131,136**
1910	-66,119.1	132,244	132,265	-66,075.3	**132,165**	**132,213**
1920	-66,468.5	132,943	132,964	-66,423.2	**132,860**	**132,909**
1930	-67,496.8	135,000	135,021	-67,441.7	**134,897**	**134,946**
1940	-68,003	136,012	136,033	-67,943	**135,900**	**135,949**
1950	-68,350.5	136,707	136,728	-68,296.1	**136,606**	**136,655**
1960	-68,458.6	136,923	136,944	-68,396.2	**136,806**	**136,855**
1970	-68,304.5	136,615	136,636	-68,273.3	**136,561**	**136,610**
1981	-68,398.3	136,803	136,824	-68,377.4	**136,769**	**136,818**
1991	-68,416.8	136,840	136,861	-68,386.1	**136,786**	**136,835**
2001	-68,629.9	137,266	137,287	-68,580.5	**137,175**	**137,224**
2010	-69,729.8	139,466	139,487	-69,659.8	**139,334**	**139,383**

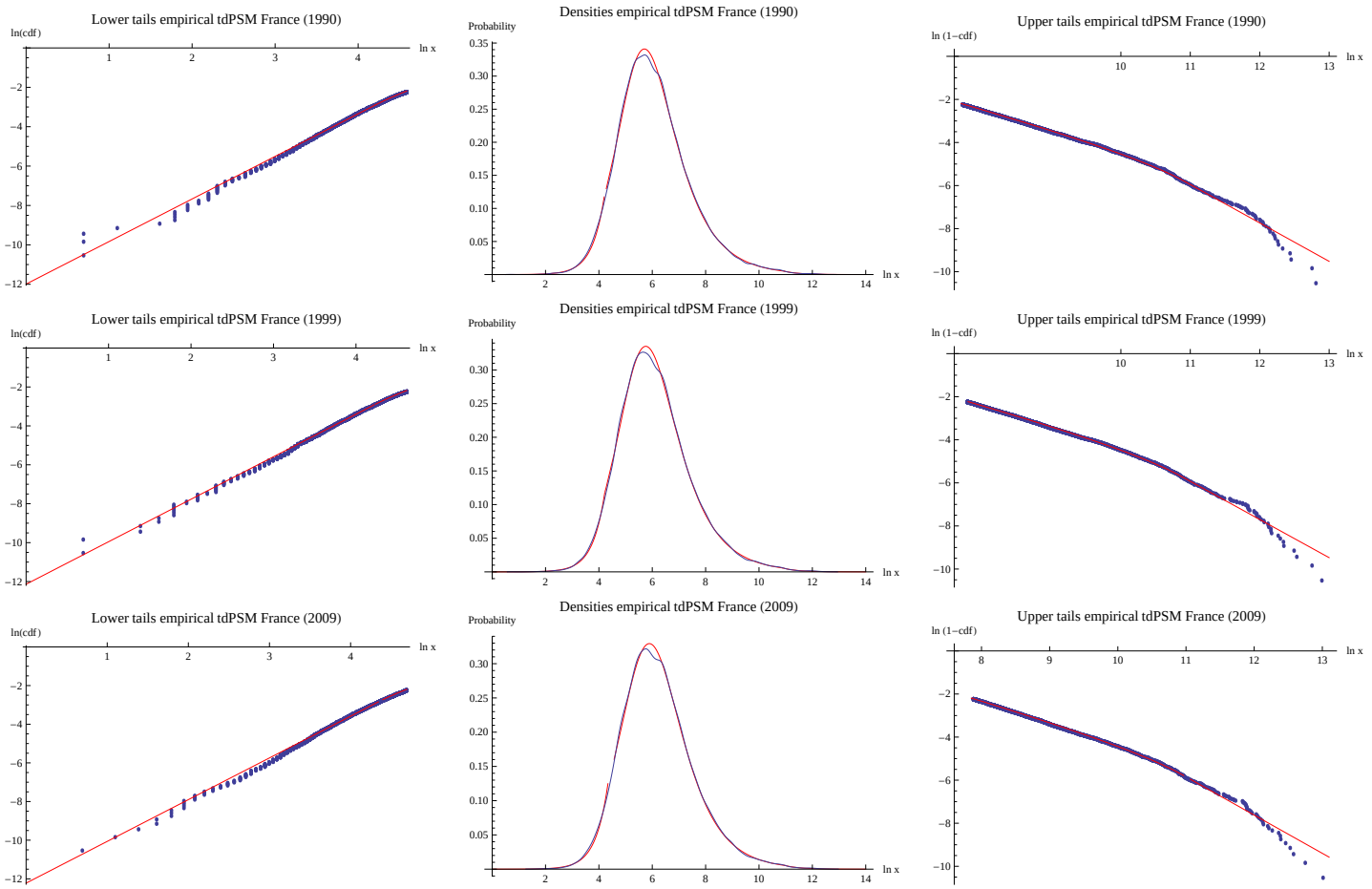

Figure 1: For the French city sizes on 1990, 1999 and 2009. Left-hand column: Empirical and estimated tdPSM $\ln(\mathrm{cdf})$ for the lower tail (empirical in blue, estimated in red). Center column: Empirical (Gaussian adaptive kernel density) and estimated tdPSM density functions (empirical in blue, estimated in red). Right-hand column: Empirical and estimated tdPSM $\ln(1-\mathrm{cdf})$ for the upper tail (empirical in blue, estimated in red).

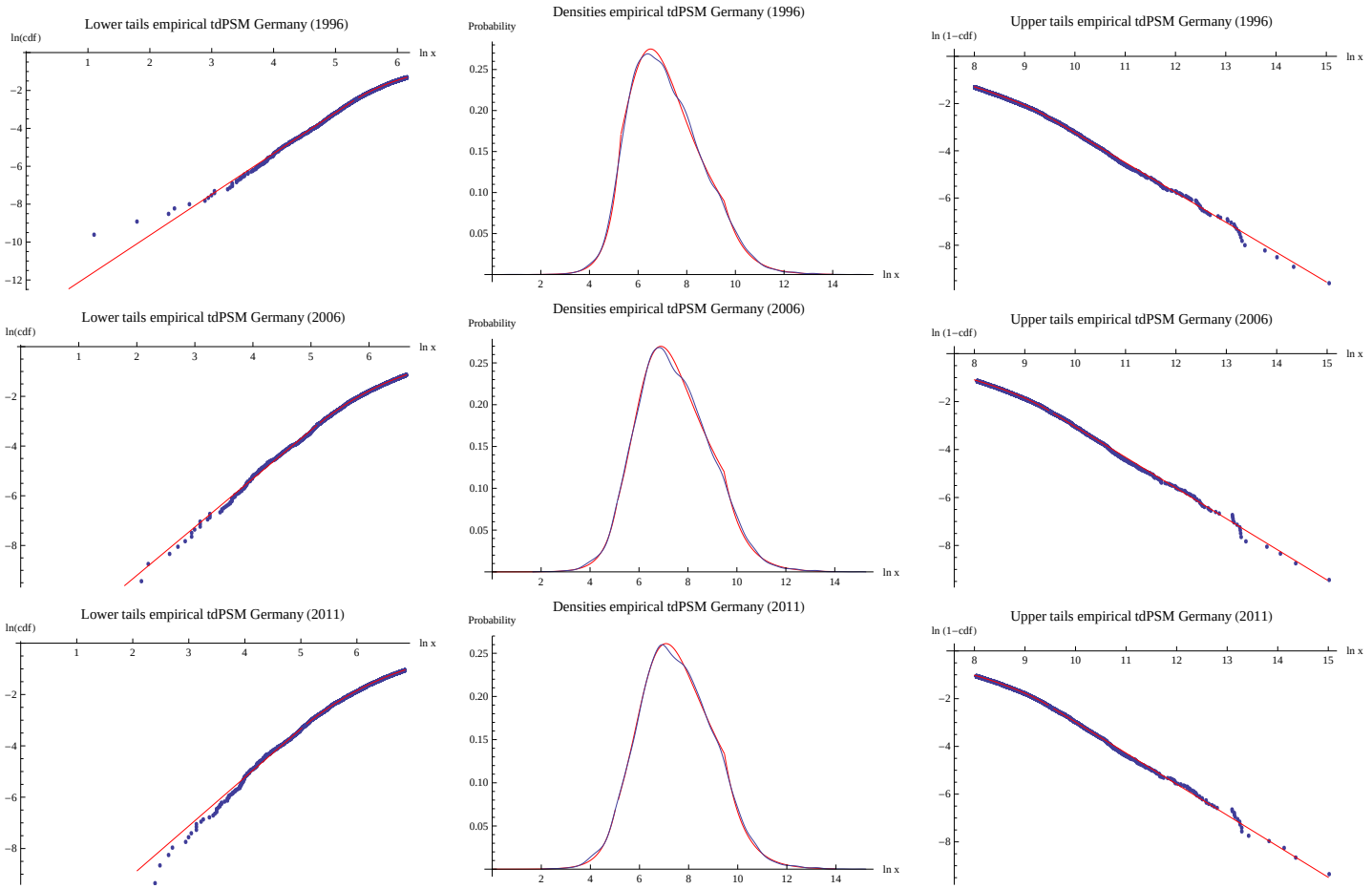

Figure 2: For the German city sizes on 1996, 2006 and 2011. Left-hand column: Empirical and estimated tdPSM $\ln(\mathrm{cdf})$ for the lower tail (empirical in blue, estimated in red). Center column: Empirical (Gaussian adaptive kernel density) and estimated tdPSM density functions (empirical in blue, estimated in red). Right-hand column: Empirical and estimated tdPSM $\ln(1-\mathrm{cdf})$ for the upper tail (empirical in blue, estimated in red).

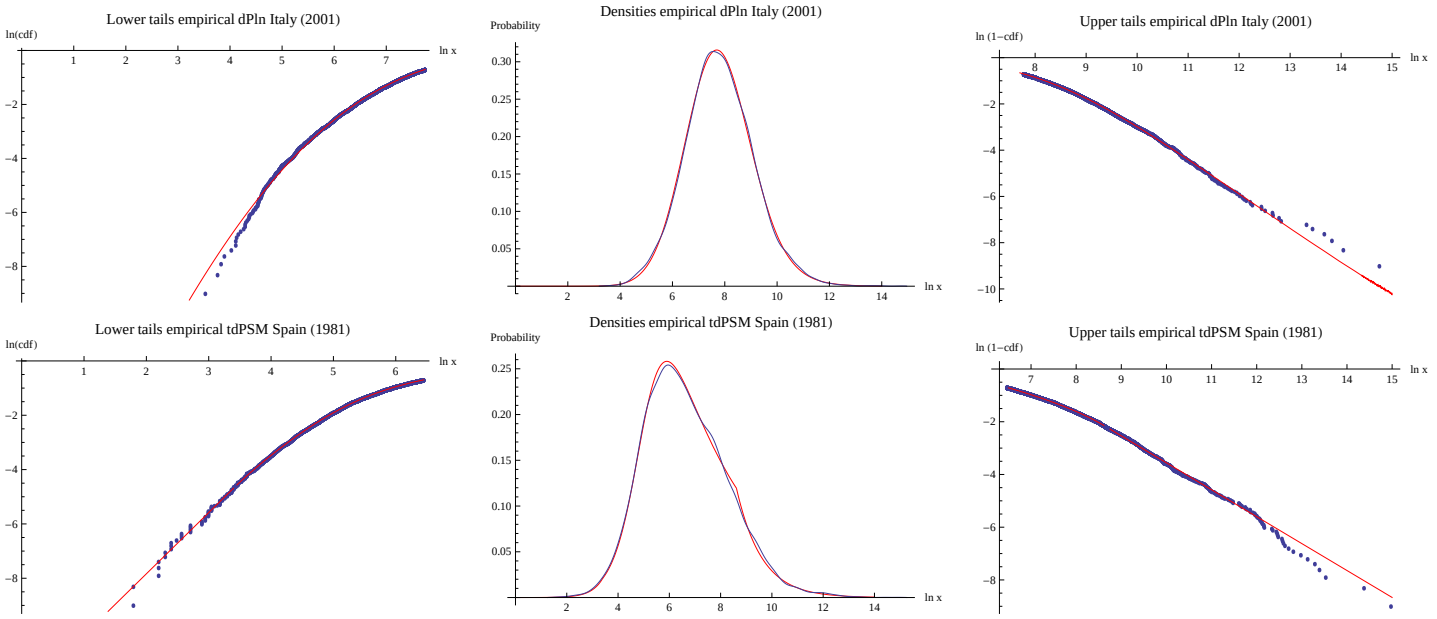

Figure 3: For the Italian and Spanish city sizes on 2001 and 1981, respectively. dPln for Italy 2001 and tdPSM for Spain 1981. Left-hand column: Empirical and estimated $\ln(\mathrm{cdf})$ for the lower tail (empirical in blue, estimated in red). Center column: Empirical (Gaussian adaptive kernel density) and estimated density functions (empirical in blue, estimated in red). Right-hand column: Empirical and estimated $\ln(1-\mathrm{cdf})$ for the upper tail (empirical in blue, estimated in red).