Author/s:
Calderin-Ojeda, E

Title:
The distribution of all French communes: A composite parametric approach

Date:
2016-05-15

Citation:
Calderin-Ojeda, E. (2016). The distribution of all French communes: A composite parametric approach. Physica A: Statistical Mechanics and its Applications, 450, pp.385-394. https://doi.org/10.1016/j.physa.2016.01.018.

Persistent Link:
https://hdl.handle.net/11343/120650

The distribution of all French communes: A composite parametric approach

Enrique Calderín–Ojeda
Centre for Actuarial Studies, Department of Economics, The University of Melbourne, Australia

ABSTRACT

The distribution of the size of all French settlements (communes) from 1962 to 2012 is examined by means of a three–parameter composite Lognormal–Pareto distribution. This model is based on a Lognormal density up to an unknown threshold value and a Pareto density thereafter. Recent findings have shown that the untruncated settlement size data is in excellent agreement with the Lognormal distribution in the lower and central parts of the empirical distribution, but it follows a power law in the upper tail. For that reason, this probabilistic family, that nests both models, seems appropriate to describe urban agglomeration in France. The outcomes of this paper reveal that for the early periods (1962–1975) the upper quartile of the commune size data adheres closely to a power law distribution, whereas for later periods (2006–2012) most of the city size dynamics is explained by a Lognormal model.

Keywords: City size; Composite models; France; Lognormal; Pareto

Acknowledgements

The author would like to express his gratitude to the three anonymous Referees for their relevant and useful comments.

Address for correspondence: Enrique Calderín–Ojeda, Centre for Actuarial Studies, Department of Economics, University of Melbourne, Australia.
e–mail: enrique.calderin@unimelb.edu.au

The distribution of all French communes: A composite parametric approach

ABSTRACT

The distribution of the size of all French settlements (communes) from 1962 to 2012 is examined by means of a three–parameter composite Lognormal–Pareto distribution. This model is based on a Lognormal density up to an unknown threshold value and a Pareto density thereafter. Recent findings have shown that the untruncated settlement size data is in excellent agreement with the Lognormal distribution in the lower and central parts of the empirical distribution, but it follows a power law in the upper tail. For that reason, this probabilistic family, that nests both models, seems appropriate to describe urban agglomeration in France. The outcomes of this paper reveal that for the early periods (1962–1975) the upper quartile of the commune size data adheres closely to a power law distribution, whereas for later periods (2006–2012) most of the city size dynamics is explained by a Lognormal model.

Keywords: City size; Composite models; France; Lognormal; Pareto

1 Introduction

Traditionally, two important principles have been considered to analyze the empirical distribution of the city size. The Zipf's law (Zipf (1949)) and the Gibrat's law of proportionate growth of cities (Gibrat (1931)). On the one hand, the Zip's law in the city distribution system indicates that the second largest city is half the size of the largest, the third largest city a third the size of the largest and the nth is the nth size of the largest one. This implies that the city sizes follow a power law distribution. For example, Moura and Ribeiro (2006) studied the Zip's law for Brazilian cities and Gangopadhyay and Basu (2009) analyzed the size distributions of urban agglomerations for India and China by estimating the scaling exponent for Zip's law. Moreover, other different approaches based on generalization of the Pareto distribution have been suggested in the literature. In this regard, Sarabia and Prieto (2009) developed a model to describe Spanish city size data by

means of the Pareto Positive Stable distribution. Similarly, Gómez–Déniz and Calderín–Ojeda (2015) examined the arrangement of urban agglomerations in Australia and New Zealand by using the Pareto ArcTan distribution. On the other hand, the Gibrat's law asserts that the Lognormal distribution emerges when the size of the cities grows randomly but proportionately. In this sense, many papers have argued that the effect of fitting city size data by means of the Pareto distribution vanishes when the whole population range is included in the sample, without excluding medium and small cities. For instance, Anderson and Ge (2005) determined that the Lognormal model is preferable to the Pareto distribution by using size distribution of Chinese cities. Similarly, Eeckhout (2004) showed that the Lognormal distribution provides a good fit to the size of all cities in the US employing data from 2000 census. In addition, examination of the transition from Lognormal to Pareto has been considered in the urban economics literature. On this subject, Luckstead and Devadoss (2014) studied the city size distribution of China and India for seven decades; in their paper they concluded that the Chinese city distribution is explained by a Lognormal model between 1950–1990 and by a Pareto distribution in 2010. In contrast, the Indian cities fluctuate from Lognormal in the earlier periods to Zipf in the most recent periods.

Nevertheless, the assertion that the Lognormal distribution provides a better fit to city size data was disputed by Levy (2009) who claimed that in the top range of the largest cities, the size distribution diverges dramatically from the Lognormal distribution and it is in excellent agreement with a straight line. In this sense, the distribution of the settlement size can be divided into two regions: the bottom and middle ranges where the empirical data are explained by the Lognormal distribution, and the top range where the empirical distribution fits a power law distribution. Relying on this idea, Giesen et al. (2010) used the four–parameter Double Pareto Lognormal distribution, a distribution that is Pareto in the upper and lower tails and Lognormal in between, to explain the distribution of all cities by using untruncated city size data from eight countries. Additionally, González–Val et al. (2015) showed that Double Pareto Lognormal distribution provides the best fit to describe city size data in the US, Spain and Italy. However, this model in unable to estimate the thresholds where the lower tail ends and the upper tail begins. Recently, Ioannides and Skouras (2013) suggested the Lognormal upper tail Pareto distribution, a four–parameter continuous spliced model that combines the Lognormal distribution in the lower tail and middle part of the distribution and the Pareto distribution in the upper tail.

In this model the break point between the two components is endogenously estimated from the data. Puente–Ajovín and Ramos (2015) use the seven–parameter threshold double Pareto Singh–Madala distribution to describe the French, German, Italian and Spanish city size data. This is a distribution with Pareto behaviour in the lower and upper tails and Singh–Madala body.

In the last few years, composite distributions have been used in actuarial statistics to model loss data when the claims faced by insurers consist of a mixture of moderate and large claims. In this paper a composite Lognormal–Pareto distribution with unrestricted mixing weights is proposed to describe the distribution of the population size of all settlements (*communes*) in France for different years in the period between 1962 and 2012. This composite model, that was firstly introduced by Scollnik (2007), uses a Lognormal distribution up to an unknown threshold value, endogenously estimated from the data, and a two-parameter Pareto density thereafter. Next, continuity and differentiability conditions are imposed at the threshold to yield a smooth density function and to reduce from four to three the number of parameters to be estimated. The resulting model is similar in shape to the Lognormal distribution but with a thicker tail. This model is nested in the Lognormal upper tail Pareto distribution and it includes as particular cases the Lognormal, Pareto and Zipf distributions. Numerical results show that the upper quartile of the commune size distribution is close to Pareto for the earlier years, and the empirical data are better explained by the Lognormal distribution for the most recent years.

The reminder of the paper is organized as follows. In Section 2, the methodology used in this work is described by providing a short review of the Lognormal upper tail Pareto and composite Lognormal–Pareto models with unrestricted mixing weights together with brief comments on parameter estimation. Next, in Section 3 the data, numerical illustrations and graphical methods of model assessment based on Zipf's plots and log–log plots are presented. Finally, the paper ends with a concluding Section.

2 Methodology

Ioannides and Skouras (2013) proposed a Lognormal upper tail Pareto (LUTP) spliced model to describe the US city size distribution. This model combines a Lognormal distribution in the lower tail and main bulk of the distribution

and a Pareto distribution in the upper tail. The probability density function (pdf) of the LUTP distribution is given by

$$f(x) = \begin{cases} r\,\dfrac{f_1(x)}{F_1(\theta)}, & 0 < x \leq \theta \\ (1-r)\,f_2(x), & \theta \leq x < \infty \end{cases}. \tag{1}$$

In (1)

$$f_1(x) = \frac{1}{\sqrt{2\,\pi}\,x\,\sigma}\,\exp\left(-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right), \quad x > 0 \tag{2}$$

is the pdf of the Lognormal distribution, where $\mu \in \mathbb{R}$ is a location parameter, $\sigma > 0$ is a scale parameter and

$$F_1(\theta) = \Phi\left(\frac{\ln\theta - \mu}{\sigma}\right) \tag{3}$$

is the cumulative distribution function (cdf) of the Lognormal distribution evaluated at θ. Here $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. The upper tail Pareto has pdf

$$f_2(x) = \frac{\alpha\,\theta^{\alpha}}{x^{\alpha+1}}, \quad x \geq \theta, \tag{4}$$

where $\alpha > 0$ is a shape parameter and $\theta > 0$ is a scale parameter. The mixing weight r satisfies that $0 \leq r \leq 1$ to ensure that (1) integrates to one. In addition to this, in order to preserve the continuity of the density, r is defined as

$$r = \frac{f_2(\theta)\,F_1(\theta)}{f_2(\theta)\,F_1(\theta) + f_1(\theta)}. \tag{5}$$

Note that the threshold value θ is endogenously estimated from the data. Obviously, this model nests the Lognormal, Pareto and Zipf distributions.

2.1 The Composite Lognormal–Pareto

The LUTP distribution can be simplified if continuity and differentiability conditions at the break point θ are imposed simultaneously. If that is the

case, a three–parameter smooth density, the composite Lognormal–Pareto (CLP) distribution, is obtained. The CLP model with unrestricted mixing weights (Scollnik (2007)) uses adequate truncation of the Lognormal density up to an unknown threshold value, θ, and the Pareto distribution thereafter. This continuous family can be seen as a convex sum of two density functions in a form of a mixture model.

The pdf of the CLP distribution is provided by

$$f(x) = \begin{cases} r \left[\Phi\left(\dfrac{\ln\theta - \mu}{\sigma} \right) \right]^{-1} f_1(x), & 0 < x \leq \theta \\ (1-r)\, f_2(x), & \theta \leq x < \infty \end{cases} \tag{6}$$

where $0 \leq r \leq 1$ is the unrestricted mixing weight, $f_1(x)$and $f_2(x)$ are the densities of the Lognormal and Pareto distributions respectively and $\Phi(\cdot)$ denotes the cdf of the standard normal distribution. By allowing for continuity and differentiability at θ, we have that the mixing weight r and the shape parameter of the Pareto distribution α are given by

$$r = \frac{\sqrt{2\,\pi}\,\alpha\,\sigma\,\Phi(\alpha\,\sigma)\,\exp\left(\frac{1}{2}(\alpha\,\sigma)^2\right)}{\sqrt{2\,\pi}\,\alpha\,\sigma\,\Phi(\alpha\,\sigma)\,\exp\left(\frac{1}{2}(\alpha\,\sigma)^2\right) + 1} \quad \text{and} \tag{7}$$

$$\alpha = \frac{\ln\theta - \mu}{\sigma^2}. \tag{8}$$

respectively. Then (6) is only defined by means of the cut–off θ, i.e. the scale parameter of the Pareto distribution, and the two parameters of the Lognormal distribution. Note that (8) imposes a restriction on the tail index α of the Pareto distribution. Observe that this model is nested in the LUTP distribution and it also includes as special cases the Lognormal, Pareto and Zipf distributions.

The survival function of (6), which will be used later to derive the Zipf's plots, is given by

$$S(x) = \begin{cases} 1 - r\,\Phi\left(\dfrac{\ln x - \mu}{\sigma} \right) \left[\Phi\left(\dfrac{\ln\theta - \mu}{\sigma} \right) \right]^{-1}, & 0 < x \leq \theta \\ (1-r) \left(\dfrac{\theta}{x} \right)^{\alpha}, & \theta \leq x < \infty. \end{cases} \tag{9}$$

From the latter expression the cdf of the CLP model can be easily derived. In addition, as (9) can be inverted, the quantile function is simply obtained as follows,

$$Q^{-1}(u) = \begin{cases} \exp\left\{\mu + \sigma\,\Phi^{-1}\left(\dfrac{u}{r}\,\Phi\left(\dfrac{\ln\theta-\mu}{\sigma}\right)\right)\right\}, & 0 < u \le r \\ \theta\left(\dfrac{1-u}{1-r}\right)^{-1/\alpha}, & r \le u < 1. \end{cases} . \quad (10)$$

Obviously, the inverse transformation method of simulation can be used to generate random variates from the CLP distribution.

2.2 Estimation

Let us assume that $\underline{x} = \{x_1, x_2, \ldots, x_k, x_{k+1}, \ldots, x_n\}$ is an ordered random sample selected from the distribution with probability density function (6). Let us also suppose that the unknown parameter θ satisfies $x_k \le \theta \le x_{k+1}$. Then, after writing the unrestricted mixing weight as $r(\mu, \sigma, \theta)$, the log–likelihood function is given by

$$\begin{aligned} &\ell(\mu,\,\sigma,\,\theta|\underline{x}) \\ &= k\left(\log r(\mu,\sigma,\theta) - \log\Phi\left(\frac{\ln\theta-\mu}{\sigma}\right) - \frac{1}{2}\log 2\pi - \log\sigma\right) + \sum_{i=1}^{k}\ln x_i \\ &- \frac{1}{2}\sum_{i=1}^{k}\left(\frac{\ln x_i-\mu}{\sigma}\right)^2 + (n-k)\,(\log(1-r(\mu,\sigma,\theta)) + \log(\ln\theta-\mu) \\ &- 2\log\sigma + \frac{\ln\theta-\mu}{\sigma^2}\log\theta\Big) - \left(\frac{\ln\theta-\mu}{\sigma^2}+1\right)\sum_{i=k+1}^{n}\log x_i. \end{aligned} \quad (11)$$

Then, after some tedious algebra, the score equations $\partial\ell/\partial\mu = 0$ and $\partial\ell/\partial\sigma = 0$ are derived. The maximum likelihood estimates of μ and σ are the simultaneous solution of these two equations. Clearly, these estimates cannot be obtained in closed–form and they must be computed numerically. As the log–likelihood function is not continuous with respect to the parameter θ, the maximum likelihood estimate of θ is calculated by segment–wise maximization.

3 Data analysis and numerical results

In this section investigation of French settlement size data is conducted by means of the Pareto, Lognormal, Lognormal upper tail Pareto (LUTP) and composite Lognormal–Pareto (CLP) distributions.

3.1 Data

The *commune* is the fourth–level administrative division of France and it corresponds to the lowest spatial subdivision of the country. Besides, it provides the best coverage since it basically represents the whole French population. Traditionally, French communes are based on pre–existing villages and have been designed with significant power to facilitate local governance. They still largely reflect the division of France into villages at the time of French Revolution. There is no parallelism between the considerably high number of communes in France than that of any other country. In general, the theories of modelling the city size distribution do not differentiate between an urban agglomeration and a rural one.
The datasets of the commune size distribution have been obtained from the national statistical office (`www.insee.fr`). In this work communes in the French overseas departments have not been examined. The sets of data considered in this manuscript comprise the estimated population for the years 1962, 1968, 1975, 1982, 1990, 1999, 2006, 2008, 2010 and 2012. It is important to point out that only communes with at least one inhabitant have been investigated in this work. In addition, across all the years considered, districts (*arrondissements*) in major cities (Marseille, Lyon and Paris) have been combined.
Some descriptive statistical measures for the empirical distribution of the size of the French communes are shown in Table 1. As it can be seen the total population of the Metropolitan France has been steadily growing over the last decades. Moreover, it can be noticed that the median is a small number as compared with the mean, due to the large proportion of communes with a low population and small proportion of settlements with a large population that makes the empirical distribution of the size of the communes positively skewed. The median shows that the vast majority of the French communes only has a few hundred inhabitants, with a steadily increasing value across the years studied. This fact is confirmed by the relatively large value of the standard deviation. Observe that the latter figure constantly reduces

Table 1: Number of communes and some descriptive statistical measures for the size of the French communes.

Year	Population	Comm.	Mean	S.D.	min	median	max
1962	46425393	36546	1270.3	16531.2	3	356.0	2790091
1968	49711853	36547	1360.2	16019.0	4	344.0	2590771
1975	52591584	36548	1439.0	14852.1	2	333.0	2299830
1982	54334871	36548	1486.7	14146.2	1	348.0	2176243
1990	56615155	36547	1549.1	13956.8	1	365.0	2152423
1999	58518395	36546	1601.2	13950.8	1	380.0	2125246
2006	62817120	36563	1718.1	14587.8	1	420.0	2203817
2008	63543618	36564	1737.9	14736.0	1	429.0	2233818
2010	64207050	36565	1756.0	14898.7	1	438.0	2268265
2012	64844825	36546	1774.3	14943.8	1	443.5	2265886

between the years 1962 and 1999 and later, it increases for the four last years under consideration. Note also that the minimum value is one for all the years examined except for 1962, 1968 and 1975. In the following, a thorough statistical analysis of these datasets will be carried out by using an untruncated settlement size.

3.2 Numerical results

In this subsection, parameter estimation is performed by the method of maximum likelihood (ML), which is implemented using the function "mle" / "mle2" in **R**. The ML estimates, across the years considered, for Pareto, Lognormal and CLP distributions, together with their corresponding standard errors, are reported in Table 2, for the distribution of the population size of the French communes after combining districts in major cities.

As it can be observed, the estimate of the Pareto exponent is well below one, thus departing from the Zipf's law; in addition its value is steadily decreasing ranging from 0.2181 in the year 1968 to 0.1609 in the year 2012. Moreover, the location parameter of the Lognormal distribuiton μ, firstly decreases in the earlier years, and then it increases over time, and the scale parameter σ steadily grows in the period of investigation. Therefore, except

Table 2: Parameter estimates obtained by ML estimation for Pareto, Lognormal and CLP distributions for the size of the French communes.

Year	Pareto	Lognormal		CLP		
	$\widehat{\alpha}$	$\widehat{\mu}$	$\widehat{\sigma}$	$\widehat{\mu}$	$\widehat{\sigma}$	$\widehat{\theta}$
1962	0.2042	5.9969	1.1506	5.7856	0.8950	782.84
	(0.0011)	(0.0060)	(0.0043)	(0.0070)	(0.0055)	(16.122)
1968	0.2181	5.9708	1.2058	5.7379	0.9273	754.14
	(0.0011)	(0.0063)	(0.0045)	(0.0087)	(0.0068)	(20.098)
1975	0.1680	5.9529	1.2708	5.6972	0.9738	746.21
	(0.0009)	(0.0066)	(0.0047)	(0.0061)	(0.0048)	(11.115)
1982	0.1667	5.9970	1.3002	5.7468	1.0151	833.58
	(0.0009)	(0.0068)	(0.0048)	(0.0043)	(0.0034)	(12.422)
1990	0.1655	6.0415	1.3241	5.8099	1.0601	969.10
	(0.0009)	(0.0069)	(0.0049)	(0.0063)	(0.0049)	(15.185)
1999	0.1646	6.0770	1.3345	5.8357	1.0655	988.04
	(0.0009)	(0.0070)	(0.0049)	(0.0061)	(0.0047)	(14.096)
2006	0.1621	6.1688	1.3348	5.9598	1.0965	1232.82
	(0.0008)	(0.0070)	(0.0049)	(0.0105)	(0.0084)	(46.188)
2008	0.1616	6.1887	1.3342	5.9843	1.1015	1284.59
	(0.0008)	(0.0070)	(0.0049)	(0.0058)	(0.0044)	(15.206)
2010	0.1611	6.2055	1.3344	6.0067	1.1079	1343.06
	(0.0008)	(0.0070)	(0.0049)	(0.0062)	(0.0048)	(20.359)
2012	0.1609	6.2161	1.3375	6.0222	1.1161	1393.32
	(0.0008)	(0.0070)	(0.0049)	(0.0067)	(0.0052)	(26.248)

for the years 1968 and 1975 when the mean declines, both the mean and variance of the Lognormal distribution increase over time. This shows firstly that the population of the communes grows and secondly that the differences among the population of the settlements widen. Next, the parameters of the CLP model consistently rise across the years considered (except $\hat{\mu}$ for 1962 and 1968), indicating the prevalence with time of the Lognormal distribution over the Pareto distribution in the composite model. This is confirmed in Table 3 by the values of the unrestricted mixing weight r. In this sense, for example, the distribution of the French communes for 1975 follows a power law distribution ($\hat{\alpha} = 0.9678$) from percentile 75 onwards; in addition, communes with a population between 1 and 746 inhabitants grows randomly and proportionately following a Lognormal distribution and the settlements with a population size greater than 746 are better represented by a power law. Similarly, for the year 2012, the city size distribution follows a Pareto law beyond percentile 81 ($\hat{\alpha} = 0.9772$). Here the distribution of cities between 1 and 1393 inhabitants adopts a Lognormal distribution, again the growth of cities of larger size are better described by the Pareto distribution. These results are supported by the Zipf's plots (see Figure 1 and 2). In a similar fashion, the estimate of the tail index $\hat{\alpha}$ is very close to unity for all the years considered. Finally, as the estimates of the parameters of the LUTP model are almost indistinguishable from those ones derived from the CLP, they are not displayed in Table 2. The values of the unrestricted mixing weight r and the estimate of the tail index $\hat{\alpha}$ for CLP and LUTP distributions are given in Table 3

Model assessment is presented from a theoretical plausibility justified by means of Kullback–Leibler divergence, suggesting using an information–criterion based approach. In this paper, the negative of the log–likelihood (NLL) and the Hanann–Quinn information criterion (HQIC) have been chosen as measures of model validation. The HQIC (see Hannan and Quinn (1979)) is evaluated as $\text{HQIC} = -2\,\text{NLL} + 2(d+1)\log(\log n)$, where d is the number of parameters in the model and n is the sample size. Note that for these two information criteria described above, smaller values indicate a better fit of the model to the data. As it can be seen in the results shown in Table 4, the CLP probabilistic family outperforms the Pareto and Lognormal distributions consistently across the different years analyzed in this work for French population data. In terms of the NLL, the LUTP provides slightly better fit to data than the CLP model; however, when using the HQIC as measure of model selection, the CLP distribution yields a lower value for all

Table 3: Estimated values of tail index $\hat{\alpha}$ and unrestricted mixing weight r for the size of the French communes for CLP and LUTP models.

	1962	1968	1975	1982	1990
			CLP		
$\hat{\alpha}$	1.0953	1.0323	0.9678	0.9499	0.9495
r	0.7687	0.7591	0.7528	0.7621	0.7796
			LUTP		
$\hat{\alpha}$	1.0943	1.0332	0.9574	0.9502	0.9491
r	0.7675	0.7605	0.7498	0.7618	0.7788
	1999	2006	2008	2010	2012
			CLP		
$\hat{\alpha}$	0.9338	0.9626	0.9676	0.9743	0.9772
r	0.7746	0.7978	0.8015	0.8064	0.8104
			LUTP		
$\hat{\alpha}$	0.9464	0.9623	0.9660	0.9725	0.9770
r	0.7835	0.7984	0.7982	0.8036	0.8102

the years examined since it penalizes the number of parameters in the model. In addition to all these models, for the sake of comparison, we have included in Table 4 the threshold double Pareto Singh–Madala (PSMP) distribution. Although the convergence to the optimal estimates is very sensitive to the initial values, the PSMP provides across all the years considered the best fit to data among all the models examined in terms of the two measures of model validation.

Table 4: NLL (above) and HQIC (below) values evaluated at ML estimates for Pareto, Lognormal, CLP, LUTP and PSMP distributions for the size of the French communes.

Year	Pareto	Lognormal	CLP	LUPT	PSMP
1962	313775	276146	274284	274284	274266
	627560	552307	548587	548592	548570
1968	310413	276914	275006	275006	274980
	620836	553842	550032	550036	549998
1975	319311	278184	276410	276410	276374
	638630	556382	552840	552846	552785
1982	321194	280362	279161	279161	279117
	642398	561277	558341	558346	558272
1990	323081	282917	281695	281694	281649
	646171	565848	563408	563413	563335
1999	324584	284492	283363	283361	283300
	649177	568999	566744	566746	566637
2006	328638	287987	286994	286993	286928
	657285	575987	574006	574011	573893
2008	329494	288707	287753	287753	287689
	658997	577428	575524	575529	575416
2010	330214	289335	288422	288422	288359
	660438	578684	576863	576868	576756
2012	330492	289657	288776	288776	288709
	660994	579328	577570	577575	577456

The upper θ_u and lower θ_L cut–off points for the CLP, LUTP and PSMP distributions are provided in Table 5. Note that the values of θ_U for each

year correspond to the estimated values of θ in the expressions (6) for CLP and (1) for LUTP. In the following, model validation is also presented from

Table 5: Upper and lower tail thresholds for the different spliced models considered for the size of the French communes.

Year	CLP	LTUP	PSMP	
	$\widehat{\theta}_U$	$\widehat{\theta}_U$	$\widehat{\theta}_L$	$\widehat{\theta}_U$
1962	782.84	779.50	42.00	1246.50
1968	754.14	758.34	3.99	1168.06
1975	746.21	740.68	1.92	985.81
1982	833.58	832.32	64.99	1045.85
1990	969.10	966.54	1.99	1222.03
1999	988.04	1035.60	27.99	1278.08
2006	1232.82	1235.32	1.99	1310.70
2008	1284.59	1262.47	1.99	1341.62
2010	1343.06	1324.15	23.01	1574.84
2012	1393.32	1391.71	29.01	1534.81

a practical point of view. In this sense, the "distance" between the empirical distribution function constructed from the data and the cumulative distribution function of the fitted models can be quantified. In particular, it is suggested to use a measure of goodness-of-fit based on the empirical distribution function (EDF) to quantify this distance, the Kolmogorov-Smirnov (KS) test statistic. Denote the cumulative distribution function of the fitted model by $\hat{F}$, the original data by $x_1, ..., x_N$ and the ordered data in increasing magnitude by $x_{(1)}, ..., x_{(N)}$, then the KS test statistic is defined as $D = max(D^+, D^-)$, where

$$D^+ = \max_{1 \leq j \leq N} \left\{ \frac{j}{N} - \hat{F}(x_{(j)}) \right\}, D^- = \max_{1 \leq j \leq N} \left\{ \hat{F}(x_{(j)}) - \frac{j-1}{N} \right\}.$$

For this model assessment measure, a smaller value indicate a better fit of the distribution to the empirical data. Here, only Pareto, Lognormal and CLP models are examined. Results are summarized in Table 6.
The Kolmogorov–Smirnov test statistic is clearly much larger for the Pareto distribution than for the other two models. By comparing the Lognormal

Table 6: Kolmogorov–Smirnov test statistic (KS) and its corresponding p–values (in brackets) for Pareto, Lognormal and CLP distributions for the size of the French communes.

Year	Pareto	Lognormal	CLP
1962	0.4387 (0.000)	0.0547 (0.000)	0.0073 (0.038)
1968	0.4192 (0.000)	0.0583 (0.000)	0.0073 (0.041)
1975	0.3932 (0.000)	0.0576 (0.000)	0.0084 (0.012)
1982	0.4519 (0.000)	0.0523 (0.000)	0.0099 (0.002)
1990	0.4477 (0.000)	0.0478 (0.000)	0.0112 (0.000)
1999	0.4473 (0.000)	0.0452 (0.000)	0.0114 (0.000)
2006	0.4493 (0.000)	0.0424 (0.000)	0.0129 (0.000)
2008	0.4492 (0.000)	0.0419 (0.000)	0.0125 (0.000)
2010	0.4486 (0.000)	0.0412 (0.000)	0.0128 (0.000)
2012	0.4482 (0.000)	0.0404 (0.000)	0.0132 (0.000)

and CLP distributions, the values of the test statistic are greater for the former model than for the latter one consistently across all the years considered. However, the numerical value of the test statistic declines with time for the Lognormal distribution whereas for the CLP model increases. The test statistic also allows us to perform hypothesis testing for model validation purposes if it is assumed that all parameters are specified completely. An extremely small p–value may lead to a confident rejection of the null hypothesis that the data come from the proposed model. Observe that the p–value is lower than 0.001 in all cases except for the years 1962, 1968 and 1975 in the CLP model. However, it is relevant to mention that the KS–test would reject the Pareto and Lognormal distribution earlier than the CLP model for all the years examined. The p-values were computed via Monte Carlo methods using a simulation size of 10000 repetitions.

The linear relationship between the population of the settlements and their corresponding ranks on a log–log plot (Zipf's plots) is found to be a power law, where the absolute value of this linear expression is the exponent of the power law. Figure 1 and 2 illustrate Zipf's plots of actual and predicted values of the Pareto, Lognormal, LUTP and CLP for the sample cities of the French communes for the years 1962, 1968, 1975, 1982, 1990 and 1999,

and 2006, 2008, 2010 and 2012 respectively. These two sets of graphical representation correspond to the plots, in log–log scale, of the complementary of the cumulative distribution function against the observed ordered data for these four models together with the empirical distribution. Logarithmic of empirical quantiles appears as scatter points on the chart, the logarithm of the theoretical quantiles are given by lines: Pareto (dotted), Lognormal (dahsed), LUTP (dotdashed) and CLP (solid). Note that for the Pareto case a straight line with slope $1/\alpha$ is obtained. Observe that for all the years considered the Pareto distribution does not perform well. The Lognormal distribution does not explain the empirical data in the years considered, however, it can be observed that the fit to data improves slightly in the second set of years (see Figure 2), thus corroborating the results given in Table 6. In addition, the CLP and LUTP distributions stay closer to the empirical data across all the years considered. This effect is even more evident in the years 1962, 1968 and 1975. Moreover, observe that the plots of the LUTP and CLP distributions are almost indistinguishable from each other.

Thus for these three years, it can be concluded that the upper quartile of the commune size distribution adheres closely to a power law distribution. Then, as time goes by, this effect vanishes. In particular for the period 2006–2012, although the French population has been steadily increasing, only the range covering communes from medium–to–large size are described by the Zip's law. For these years (see Figure 2), the Pareto distribution tends to overestimate the empirical distribution for the communes of large–to–extreme size. These results are also supported by the values of KS–statistic, the Zipf's plots and the value of the mixing weight r. Note that for the last years, the Lognormal distribution explains more than 80% of the composite model. Although is not considered here, the Zipf's plot could be improved by including more complex models such as the PSMP distribution.

Finally, graphical validation of the model is also illustrated by means of two more sets of scatter plots. Figure 3 and Figure 4 show the differences between empirical and fitted quantiles in log scale for the two set of years aforementioned. Here the Pareto distribution is no longer considered. As it can be observed the CLP and LUTP provide the best fit to data over all years under consideration and the solid line (CLP) and dotdashed line (LUTP) stay closer to the horizontal axis (i.e. empirical log quantiles) than the Lognormal distribution (dashed line). Note that for the top part of the distribution the spliced models tend to overestimate the empirical distribution of the commune size whereas the Lognormal distribution underestimates

the empirical data.

4 Conclusion

In this paper, it has been shown that the size distribution of all communes in France can be easily modelled with a three–parameter composite Lognormal–Pareto distribution. This continuous family uses a Lognormal distribution up to an unknown threshold value, estimated from the data, and a Pareto density thereafter. This feature is consistent with the empirical findings by Levy (2009). Additionally, the settlement size distribution has been examined for ten different years between 1962 and 2012. From the numerical results, it has been determined that in the early periods, the Pareto distribution adheres closely to the upper quartile of the empirical data, whereas for the more recent years this distribution tends to overestimate the population of the largest cities.

References

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

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

Gangopadhyay, K., Basu, B. (2009). City size distributions for India and China. *Physica A* 388 (13), 2682–2688.

Gibrat, R. (1931). Les inégalités économiques, Librairie du Recueil Sirey, Paris.

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

Gómez–Déniz, E. and Calderín–Ojeda, E. (2015). On the use of the Pareto ArcTan distribution for describing city size in Australia and New Zealand. *Physica A* 436, 821–832.

González–Val, R., Ramos, A., Sanz–Gracia, F. and Vera–Cabello, M. (2015). Size distributions for all cities: Which one is best? *Papers in Regional Science* 94(1), 177–196.

Hannan, E. J.and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, B* 41, 190–195.

Ioannides, Y. and Skouras, S. (2013). US city size distribution: Robustly Pareto, but only in the tail. *Journal of Urban Economics* 73(1), 18–29.

Levy, M. (2009). Gibrat's Law for (all) cities: Comment. *American Economic Review* 99(4), 1672–1675.

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

Moura, N. J., and Ribeiro, M. B. (2006). Zipf law for Brazilian cities. *Physica A* 367, 441–448.

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.

Sarabia, J.M. and Prieto, F. (2009). The Pareto–positive stable distribution: A new descriptive model for city size data. *Physica A* 388, 4179–4191.

Scollnik, D. P. M. (2007). On composite Lognormal–Pareto models. *Scandinavian Actuarial Journal* 1, 20–33.

Zipf, G,K. (1949). Human Behavior and the Principle of Least Effort. Addison–Wesley Press. Cambridge, MA.

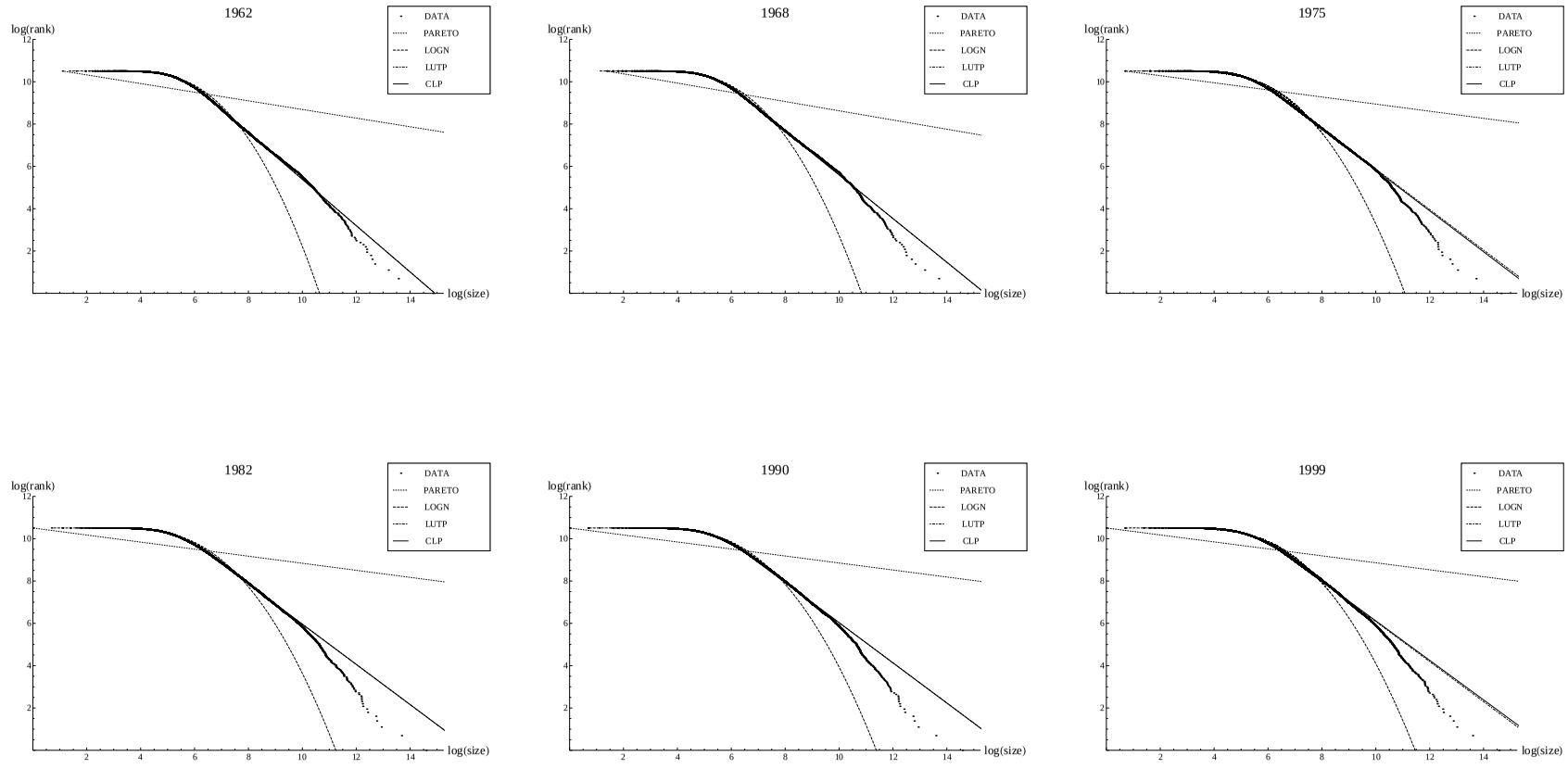

Figure 1: Zipf's plots for the size of the French communes (years 1962–1999).

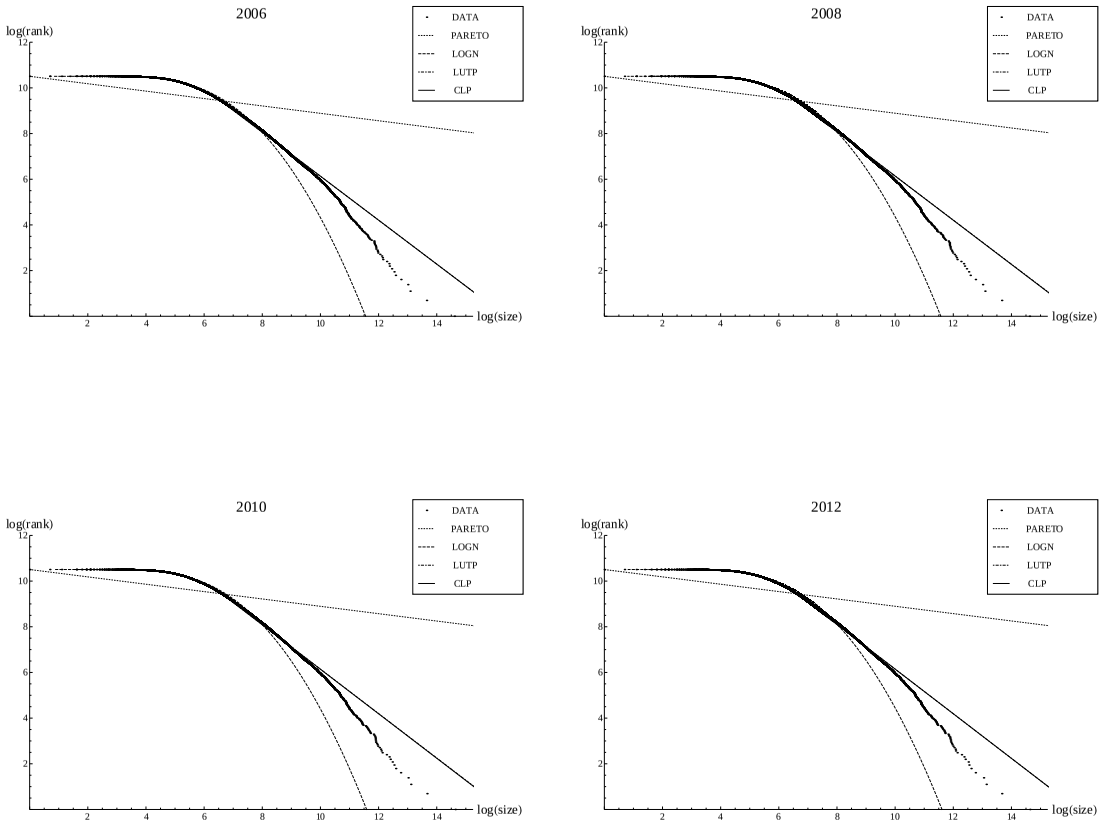

Figure 2: Zipf's plots for the size of the French communes (years 2006–2012).

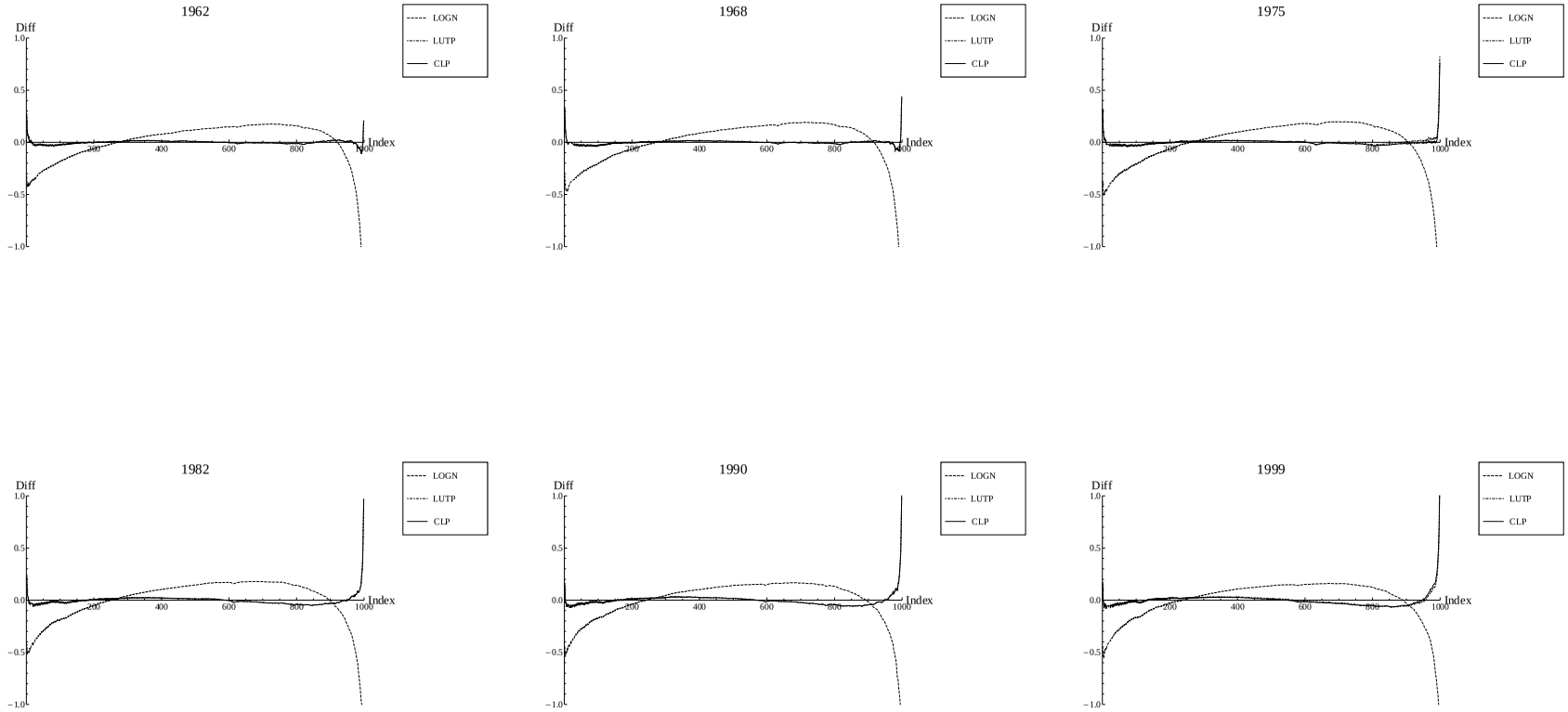

Figure 3: Difference of empirical and fitted quantiles in log scale (years 1962–1999).

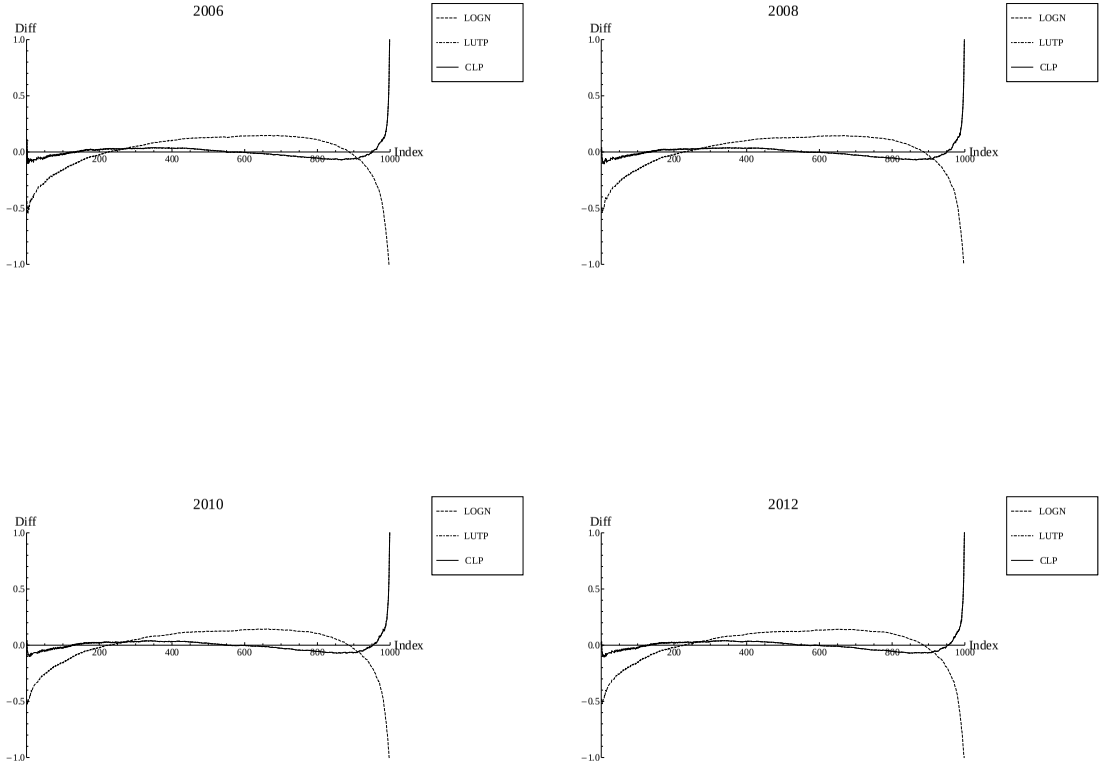

Figure 4: Difference of empirical and fitted quantiles in log scale (years 2006–2012).