Accepted Manuscript

Comparisons of log-normal mixture and Pareto tails, GB2 or log-normal body of Romania's all cities size distribution

Irina Băncescu, Luminiţa Chivu, Vasile Preda, Miguel Puente-Ajovín, Arturo Ramos

PII: S0378-4371(19)30627-2
DOI: https://doi.org/10.1016/j.physa.2019.04.253
Reference: PHYSA 21017

To appear in: *Physica A*

Received date : 3 September 2018
Revised date : 10 March 2019

Please cite this article as: I. Băncescu, L. Chivu, V. Preda et al., Comparisons of log-normal mixture and Pareto tails, GB2 or log-normal body of Romania's all cities size distribution, *Physica A* (2019), https://doi.org/10.1016/j.physa.2019.04.253

***Highlights (for review)**

Highlights

- Empirical evidence for the fitting of cities size of Romania both from models having Pareto tails and from models which do not, situations that occur at the same time
- Comparisons of Pareto tails and log-normal or generalized beta of second kind (GB2) body distributions to mixtures of log-normal models
- Statistical equivalence of mixture of log-normal models to Pareto tails and different bodies
- Introduction of the threshold double Pareto GB2 model (tdPGB2)

Available online at www.sciencedirect.com
ScienceDirect

ELSEVIER

Physica A

Physica A 00 (2019) 1–14

Comparisons of log-normal mixture and Pareto tails, GB2 or log-normal body of Romania's all cities size distribution

Irina Băncescu[a,*], Luminiţa Chivu[a,1], Vasile Preda[a,1], Miguel Puente-Ajovín[b,1], Arturo Ramos[b,1]

[a] "Costin C. Kiriţescu" National Institute of Economic Research, Bucharest, Romania
[b] Department of Economic Analysis, Universidad de Zaragoza, Zaragoza, Spain

Abstract

Modeling demographic data has been on the agenda of statisticians for many years. Some of the distributions used are Pareto, reverse Pareto, q-exponential and log-normal models. An approach to this problem is to consider three statistical models: one for the upper tail, one for the middle range, and another for the lower tail. This paper deals with the size distribution of urban and rural agglomerations in Romania for the 1992-2017 period, by comparing the recently introduced three log-normal mixture (3LN), Pareto tails log-normal (PTLN), and threshold double Pareto Generalized Beta of second kind (tdPGB2) models. The tdPGB2 statistical model has the PTLN distribution as a limiting case. The maximum likelihood estimates of the distributions are computed, and goodness-of-fit tests are performed using the Kolmogorov-Smirnov (KS), Cramér-von Mises (CM) and Anderson-Darling (AD) statistics. Also, we use the Vuong and Bayes factor log-likelihood tests. Using both graphical and formal statistical tests, our results rigorously confirm that the 3LN model is statistically equivalent to PTLN and tdPGB2 distributions, the preferred model being the PTLN probability law. Both the PTLN and tdPGB2 distributions have Pareto tails but the 3LN model does not. All the three models prove to be very well suited parameterizations of Romania's city size data.

Keywords: Mixture of log-normal models, Pareto upper and lower tails, GB2 distribution, City-size distribution

1. Introduction

Although natural phenomena are complex processes, they frequently display macroscopic regularities. Statisticians observe these patterns and try to describe them by different probability laws. One such complex system is represented by the distribution of cities and villages, in different countries or regions.

City size distribution has been studied extensively for several decades [1, 2, 3, 4]. The first studies considered only big cities, presumably due to lack of data. However, owing to advances in technology and statistical tools, data for small cities have been available for researchers.

Despite the vast research conducted so far, the fitting of the whole population of cities, both small and big, remains difficult. Some studies have attempted to combine the log-normal body and the upper-tail Pareto into a unified distribution to analyze the distribution of all cities [5], introducing, among others, the Pareto tails log-normal (PTLN)

*Corresponding author
E-mail address: irina_adrianna@yahoo.com
[1] chivu@ince.ro (L. Chivu), vasilepreda0@gmail.com (V. Preda), mpajovin@unizar.es (M. Puente-Ajovín), aramos@unizar.es (A. Ramos)

distribution, by modeling lower and upper tails with Pareto and middle range with log-normal, and identifying the transition points both from lower tail to log-normal and from log-normal to upper tail [6, 7]. Other distributions, such as the reverse Pareto and reverse generalized Pareto, are used in analyzing the lower tail cities size [8]. Another probability law used in describing the distribution of cities for all ranges of populations is the q-exponential distribution, which reproduces the Zipf-Mandelbrot law. This function is related to the generalized non-extensive statistical mechanics, obeying an anomalous decay equation [9, 10]. In 2018, the q-exponential distribution was generalized, the resulted distribution being used to describe urban data [11].

In 2015, Puente-Ajovín and Ramos [12] concluded that the threshold double Pareto Singh-Maddala distribution (tdPSM) is the preferred model in four countries: France, Germany, Italy, and Spain. The tdPSM distribution considers Pareto behavior in the lower and upper tails, and a Singh-Maddala body. This distribution has also been used to model the US city size [13]. This type of statistical model also considers two transition points or thresholds between the tails and the body which can be determined endogenously by maximum log-likelihood estimation.

In this paper, we analyze empirically the population distribution of municipalities, towns, and rural administrative units for Romania. It is well known that population size is influenced by three factors: natural growth, internal and external migrations. According to various studies conducted Romania shows a large mobility of individuals [14, 15].

During the first years of the transition from a state-socialist society to a market economy based, democratic society (1990-1994), internal migration went through certain transformations. In 1990, the rural-urban flow has reached the high share of 70 percent of all migrants, and dropped to only 30.5 percent in 1994. Nowadays, in Romania the urban-rural migration is higher than the traditional reverse flow; from 1992 more people started to move from towns to rural areas and in 1997 the migration from urban to rural areas became higher than the reverse flow [16, 17].

According to data provided by the National Institute of Statistics (INS), in the year 2017, on average, 11.3 out of 1,000 of urban residents changed their residential status to rural, while the average annual flow of internal migration from rural to urban areas was 7 people out of 1,000 inhabitants. In 2017, the rural-urban flow share was 22.90% of all migrants. However, despite the change of internal migration flows direction, there is a significant external migration at the national level, from both urban and rural areas abroad to other countries, which lead to a massive depopulation of the rural areas [18]. According to recent studies, over 3 million active individuals from Romania (approximately 15% of the total population), most of whom belonging to the 25-45 age segment are graduates from high school or university and live abroad [19, 20, 21].

During the 2007-2017 period, the urban population decreased from 55.44% in 2007 to 53.60% in 2017, the urban population being higher in the larger towns. At present, only six towns, except for the capital city, exceed 300,000 inhabitants: Iaşi, Timişoara, Cluj-Napoca, Constanţa, Galaţi and Craiova. The capital Bucharest itself currently counts with over 2 million inhabitants. In 2017, the urban-urban flow share was 29.35% of all migrants.

Using the three log-normal mixture (3LN), Pareto tails log-normal (PTLN), and threshold double Pareto Generalized Beta of second kind distributions (tdPGB2), we prove that Romanian cities' size distribution (considering all cities) suits well to these statistical models, the preferred model being the PTLN probability law. In our analysis we used the information Tempo online INS database regarding usually resident population from urban and rural areas, from 1992 to 2017, organized into administrative-territorial units (UATs).

In 2017, by its residential population of 19.64 millions of inhabitants, Romania was ranked the 7th among the 28 Member States of the European Union, after Germany, France, the UK, Italy, Spain and Poland, that is about 3.8% of the total EU 28 population. In the whole EU 28, from 2007 to 2017, the total residential population increased by approximately 13.2 millions of inhabitants (2.7%). Despite a deceleration in population growth is registered in the entire European continent, what is registered in Romania is much worse. From 2007 to 2017, in Romania, the total residential population decreased by 1.49 million people (-7.0%).

In 2017, out of a total number of 3,181 UATs in Romania, 320 (10% of the total) are located in the urban area (municipalities and towns) and 2,861 (90%) in the rural area (rural administrative units). These 320 municipalities and towns are structured in terms of the size of the population according to the following scheme:

- Less than 10,000 inhabitants, which comprises 36.8% of the total UATs from urban area and approximately 6.4% from the population in this area.
- Between 10,000 and 99,999 inhabitants, that is 55.31% of the UATs and 38% of the urban population.
- More than 100,000 inhabitants, that is 7.81% UATs, and 55.6% of the urban population.

In 2017, 56.33% of the Romanians lived in these 320 urban areas UATs. As for the inhabitants living in the rural areas, the structure of the 2,861 rural administrative units, by the size population, is the following:

- Less than 5,000 inhabitants, comprising 83.01% of the number of UATs and 65.2% of the total rural population.
- Between 5,000 and 10,000, that is 15.55% of the total UATs in the rural area and 29.6% of the total rural population.
- More than 10,000, but not more than 32,000 inhabitants, that is 1.4% of the UATs and 5.2% of the total rural population.

Thus, analyzing the size distribution of all Romanian cities, during the 1992-2017 time span and focusing on the years 1992, 2007 and 2017, provides an essential insight into the organization of living areas in Romania.

In 1992, in Romania there were 260 towns and 2,686 rural administrative units, while the country had their first general election after the communist era. In 2007, Romania became an EU member.

This paper is organized as follows. In Section 2, we present the three log-normal mixture (3LN), the Pareto tails log-normal (PTLN) and threshold double Pareto Generalized Beta of second kind (tdPGB2) distributions. Empirical analysis of Romania's towns and rural administrative units population is performed in Section 3, while Section 4 concludes the paper.

2. Methodology

Some characteristics of the data sets considered such as maximum and minimum values, number of observations, measures of skewness and kurtosis, standard deviations, and means are shown in Table 1. We notice that the measure of kurtosis is very high for each data set, suggesting a heavy tail distribution. Also, the skewness is high for these data sets. In Figs. 1 and 2 we display the empirical density function of Romania log city sizes for 1992, 2007 and 2017.

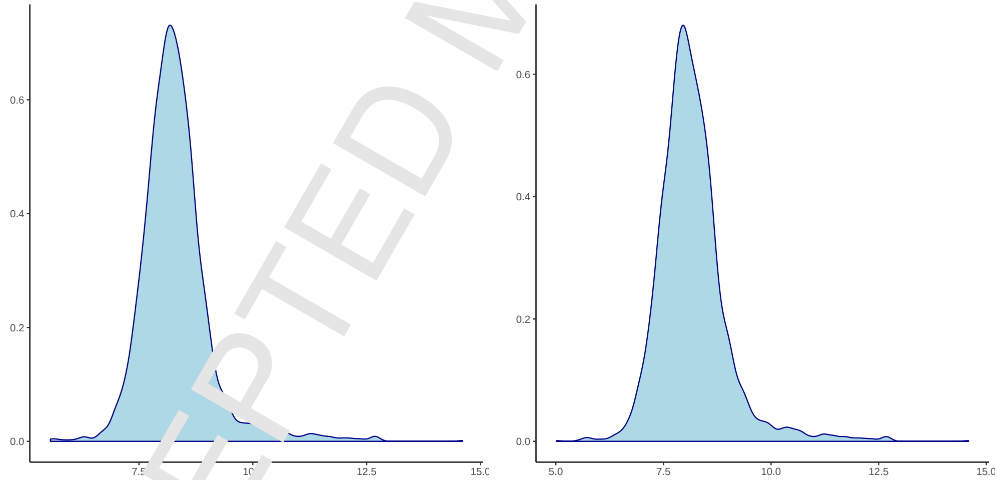

Figure 1. Density of the log of Romanian cities 1992 and 2007.

The three log-normal mixture distribution (3LN) [22] is defined by the following density function

$$f_{3LN}(x;\mu_1,\sigma_1,\mu_2,\sigma_2,\mu_3,\sigma_3,\pi_1,\pi_2,\pi_3) = \sum_{i=1}^{3} \pi_i f_{LN}(x;\mu_i,\sigma_i) \tag{1}$$

where $x > 0$, $0 \leq \pi_i \leq 1$, $\pi_1 + \pi_2 + \pi_3 = 1$, and f_{LN} is the density function of log-normal model of parameters $\mu, \sigma > 0$, that is,

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

Table 1. Descriptive statistics of Romania cities' population

Year	Nr. of obs.	Mean	SD	Mean (log scale)	SD (log scale)	Min	Max	Skewness	Kurtosis
1992	2,946	7,850	45,517	8.317	0.774	259	2,191,176	38.[illegible]0	1,796
1993	2,946	7,841	45,656	8.307	0.779	247	2,195,490	38.3[illegible]	1,788
1994	2,946	7,834	45,665	8.300	0.784	241	2,193,1[illegible]	38.[illegible]7	1,780
1995	2,946	7,819	45,617	8.293	0.789	240	2,188,462	38.19	1,774
1996	2,948	7,789	45,411	8.286	0.792	233	2,177,[illegible]	38.09	1,767
1997	2,948	7,769	45,263	8.280	0.795	219	2,1[illegible],149	38.02	1,762
1998	2,948	7,756	45,131	8.277	0.798	209	2,160,94[illegible]	37.98	1,760
1999	2,951	7,735	44,981	8.273	0.801	187	2,154,19[illegible]	38.01	1,762
2000	2,951	7,729	44,919	8.273	0.802	179	2,152,178	38.01	1,762
2001	2,951	7,723	44,873	8.271	0.804	17[illegible]	2,14[illegible],763	38.06	1,765
2002	2,955	7,698	44,844	8.266	0.806	171	2,151,408	38.22	1,779
2003	2,983	7,611	44,610	8.252	0.807	169	2,151,527	38.54	1,807
2004	3,133	7,232	43,503	8.194	0.804	165	2,151,552	39.49	1,896
2005	3,164	7,150	43,268	8.180	0.805	1[illegible]	2,151,601	39.71	1,916
2006	3,173	7,121	43,242	8.174	0.808	155	2,154,487	39.83	1,926
2007	3,176	7,106	43,242	8.171	0.809	151	2,156,978	39.93	1,933
2008	3,180	7,089	43,203	8.169	0.809	159	2,158,816	40.03	1,940
2009	3,180	7,082	43,220	8.166	0.8[illegible]	153	2,160,627	40.08	1,944
2010	3,181	7,071	43,224	8.162	0.81[illegible]	151	2,162,037	40.14	1,949
2011	3,181	7,055	43,116	8.159	0.819	147	2,157,282	40.17	1,951
2012	3,181	7,042	43,000	8.158	0.820	145	2,151,758	40.10	1,946
2013	3,181	7,029	42,837	8.153	0.82[illegible]	142	2,140,816	39.95	1,934
2014	3,181	7,010	42,362	8.149	0.827	137	2,110,752	39.68	1,914
2015	3,181	6,997	42,215	8.145	0.831	127	2,100,519	39.68	1,914
2016	3,181	6,989	42,230	8.1[illegible]1	0.835	125	2,103,251	39.67	1,913
2017	3,181	6,977	42,366	8.1[illegible]5	0.839	120	2,112,483	39.80	1,922

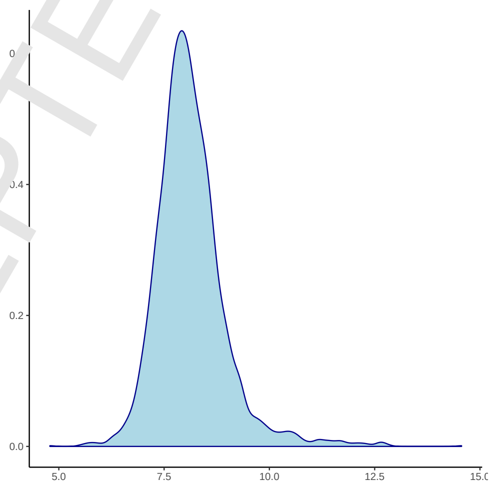

Figure 2. Density of the log of Romanian cities 2017.

In addition, the cumulative distribution function (CDF) of the 3LN model is simply

$$F_{3LN}(x; \mu_1, \sigma_1, \mu_2, \sigma_2, \mu_3, \sigma_3, \pi_1, \pi_2, \pi_3) = \sum_{i=1}^{3} \pi_i \Phi(x; \mu_i, \sigma_i) \tag{2}$$

where Φ is the CDF corresponding to the log-normal density function, that is,

$$\Phi(x; \mu, \sigma) = \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{\ln(x) - \mu}{\sigma\sqrt{2}}\right),$$

erf being the error function associated to the normal distribution.

Once the eight parameters of the 3LN model are estimated, we can use the quantile function to predict city sizes according to this distribution as

$$\tilde{x}_{3LN} = F^{-1}_{3LN}(p) = \inf\{x \in (0, \infty) \,|\, F_{3LN}(x) \geq p\}, \quad p \in [0, 1]. \tag{3}$$

This class of log-normal mixtures has been introduced in the study of city size distributions in 2019 by Kwong and Nadarajah and proved to be a better fit to the US 2010 all places' census data and Indian 2011 census data than the PTLN model. A mixture of three log-normal densities can accomodate heavy tails (see, e.g., [23] and [24]), type of tails our data display. By modeling the Romanian data sets by the 3LN probability law we are assuming that the whole population can be grouped into three, in principle different, subpopulations, each following a log-normal distribution. The subpopulations of cities are assumed to have similar growth characteristics [22]. The number of subpopulations can be taken to be also five or seven, for example, but the improvement in the corresponding maximum log-likelihoods is small (for the cases of USA and India, see [22]) and the additional information does not balance the huge increase in the complexity of the model.

In this paper, we show that the PTLN and tdPLN models are statistically equivalent to the 3LN distribution for Romania's census city size. In 2011, Bee et al. [25] gave empirical support that probability laws having log-normal body and Pareto tails can be generated as mixtures of log-normal models. Growiec et al. [26] showed that a log-normal distribution multiplied by a stretching factor leads to a Pareto upper tail.

The Pareto tails log-normal probability law (PTLN) [6] is defined by

$$f_{PTLN}(x; \alpha, \tau_l, \mu, \sigma, \tau_u, \beta) = \begin{cases} dex^{\alpha-1}, & 0 < x \leq \tau_l \\ df_{LN}(x; \mu, \sigma), & \tau_l \leq x \leq \tau_u \\ dcx^{-\beta-1}, & \tau_u \leq x < \infty \end{cases} \tag{4}$$

where the continuity constants are $e = \dfrac{f_{LN}(\tau_l; \mu, \sigma)}{\tau_l^{\alpha-1}}$, $c = \dfrac{f_{LN}(\tau_u; \mu, \sigma)}{\tau_u^{-1-\beta}}$, and the normalization constant d is given by

$$d = \left(f_{LN}(\tau_l; \mu, \sigma)\frac{\tau_l}{\alpha} + \Phi(\tau_u; \mu, \sigma) - \Phi(\tau_l; \mu, \sigma) + f_{LN}(\tau_u; \mu, \sigma)\frac{\tau_u}{\beta} \right)^{-1}.$$

The CDF of the PTLN distribution is defined by

$$F_{PTLN}(x; \alpha, \tau_l, \mu, \sigma, \tau_u, \beta) = \begin{cases} de\dfrac{x^{\alpha}}{\alpha}, & 0 < x \leq \tau_l \\ k_1 + d(\Phi(x; \mu, \sigma) - \Phi(\tau_l; \mu, \sigma)), & \tau_l \leq x \leq \tau_u \\ k_2 + \dfrac{cd}{\beta}(\tau_u^{-\beta} - x^{-\beta}), & \tau_u \leq x < \infty \end{cases} \tag{5}$$

where $k_1 = de \int_0^{\tau_l} x^{\alpha-1}\, dx$ and $k_2 = k_1 + d\int_{\tau_l}^{\tau_u} f_{LN}(x; \mu, \sigma)\, dx$.

Using the estimated parameters and the quantile function, we can predict city sizes according to this distribution as

$$\tilde{x}_{PTLN} = \begin{cases} \left[\dfrac{\alpha F_{PTLN}(x)}{de}\right]^{1/\alpha}, & F_{PTLN}(x) \in [0, k_1) \\ \Phi^{-1}\left[\dfrac{F_{PTLN}(x) - k_1}{d} + \Phi(\tau_l)\right], & F_{PTLN}(x) \in [k_1, k_2] \\ \left[\dfrac{dc}{(\tau_u)^{-\beta} de - \beta\big(F_{PTLN}(x) - k_2\big)}\right]^{1/\beta}, & F_{PTLN}(x) \in (k_2, 1] \end{cases} \tag{6}$$

where $\Phi^{-1}(p; \mu, \sigma) = \inf\{x \in (0, \infty) \,|\, \Phi(x; \mu, \sigma) \geq p\}$ is the quantile function of the log-normal distribution.

One could consider a nested model in the PTLN distribution, denoted by PTLN-diff, in which differentiability of the probability density function f_{PTLN} at the threshold points τ_l, τ_u is required. This means reducing the number of parameters by two. The differentiability conditions boil down to imposing the constraints

$$\alpha = \frac{\mu - \ln(\tau_l)}{\sigma^2} \tag{7}$$

$$\beta = \frac{\ln(\tau_u) - \mu}{\sigma^2} \tag{8}$$

the PTLN-diff distribution having four parameters to be estimated $(\tau_l, \mu, \sigma, \tau_u)$.

Prior to introducing the tdPGB2 model, let us mention that the Generalized Beta of second kind distribution (GB2) [27, 28, 29, 30] is used often in economics, insurance and income studies, and it has a density function of the form

$$f_{GB2}(x; a, b, p, q) = \frac{a x^{ap-1}}{b^{ap} B(p, q)(1 + (x/b)^a)^{p+q}} \tag{9}$$

where $x > 0$, $a, b, p, q > 0$ and $B(p, q)$ denotes the Beta function.[2]

The CDF corresponding to the GB2 density function is given by

$$F_{GB2}(x; a, b, p, q) = \frac{1}{B(p, q)} B\Big(\frac{(x/b)^a}{1 + (x/b)^a}, p, q\Big) \tag{10}$$

where $B(x, p, q) = \int_0^x t^{p-1}(1 - t)^{q-1}\, dt$, $x \in [0, 1]$ is the incomplete Beta function.

Then, the third statistical model considered in this paper, the tdPGB2 distribution [32], is defined by density function

$$f_{tdPGB2}(x; \alpha^*, \tau_l^*, a, b, p, q, \tau_u^*, \beta^*) = \begin{cases} d^* e^* x^{\alpha^* - 1}, & 0 < x \leq \tau_l^* \\ d^* f_{GB2}(x; a, b, p, q), & \tau_l^* \leq x \leq \tau_u^* \\ d^* c^* x^{-1-\beta^*}, & \tau_u^* \leq x < \infty \end{cases} \tag{11}$$

where the continuity constants are $e^* = \dfrac{f_{GB2}(\tau_l^*; a, b, p, q)}{(\tau_l^*)^{\alpha^*-1}}$, $c^* = \dfrac{f_{GB2}(\tau_u^*; a, b, p, q)}{(\tau_u^*)^{-1-\beta^*}}$, and the normalization constant is given by

$$d^* = \left(e^* \frac{(\tau_l^*)^{\alpha^*}}{\alpha^*} + F_{GB2}(\tau_u^*; a, b, p, q) - F_{GB2}(\tau_l^*; a, b, p, q) + \frac{c^*}{\beta^* (\tau_u^*)^{\beta^*}}\right)^{-1}.$$

The tdPGB2 distribution depends on eight parameters $\alpha^*, \tau_l^*, a, b, p, q, \tau_u^*, \beta^* > 0$, where α^* and β^* are Pareto exponents, τ_l^* being the lower tail switching point and τ_u^* is the upper tail cutoff. Analogously to the fact that the log-normal distribution is a limiting case of the GB2 model, the tdPGB2 has the PTLN distribution as limiting case. For $p = 1$, the tdPGB2 distribution is reduced to tdPSM model [12]. If we take $q = 1$, we obtain a probability law

[2] All the three shape parameters a, p, q control the tail behavior, and large values of the parameter a results in a thinning of the tails [31]. Also, for $p = 1$, the GB2 distribution is reduced to the Singh-Maddala submodel, while for $q = 1$, we get the Dagum submodel. Other submodels include the log-logistic ($p = q = 1$) and Lomax ($a = p = 1$) distributions, while the gamma, Weibull and log-normal models are limiting distributions.

having Pareto tails and Dagum body. Choosing different values for the parameters of the GB2 body, we derive new distributions having Pareto tails and, for example, Lomax or log-logistic body.

The CDF of tdPGB2 distribution is

$$F_{tdPGB2}(x;\alpha^*,\tau_l^*,a,b,p,q,\tau_u^*,\beta^*) = \begin{cases} d^* e^* \dfrac{x^{\alpha^*}}{\alpha^*}, & 0 < x \le \tau_l^* \\ k_1^* + d^*(F_{GB2}(x;a,b,p,q) - F_{GB2}(\tau_l^*;a,b,p,q)), & \tau_l^* \le x \le \tau_u^* \\ k_2^* + \dfrac{d^* c^*}{\beta^*}((\tau_u^*)^{-\beta^*} - x^{-\beta^*}), & \tau_u^* \le x < \infty \end{cases} \tag{12}$$

where $k_1^* = d^* e^* \int_0^{\tau_l^*} x^{\alpha^*-1}\,dx$, and $k_2^* = k_1^* + d^* \int_{\tau_l^*}^{\tau_u^*} f_{GB2}(x;a,b,p,q)\,dx$.

Using the estimated parameters and the quantile function, we can predict city sizes according to the tdPGB2 distribution as

$$\tilde{x}_{tdPGB2} = \begin{cases} \left[\dfrac{\alpha^* F_{tdPGB2}(x)}{d^* e^*}\right]^{1/\alpha^*}, & F_{tdPGB2}(x) \in [0, k_1^*) \\ F_{GB2}^{-1}\left[\dfrac{F_{tdPGB2}(x) - k_1^*}{d^*} + F_{GB2}(\tau_l^*)\right], & F_{tdPGB2}(x) \in [k_1^*, k_2^*] \\ \left[\dfrac{d^* c^*}{(\tau_u^*)^{-\beta^*} d^* e^* - \beta^*\left(F_{tdPGB2}(x) - k_2^*\right)}\right]^{1/\beta^*}, & F_{tdPGB2}(x) \in (k_2^*, 1] \end{cases} \tag{13}$$

where $F_{GB2}^{-1}(p;a,b,p,q) = \inf\{x \in (0,\infty) \mid F_{GB2}(x;a,b,p,q) \ge p\}$ is the quantile function of the GB2 distribution.

Analogously to the case of the PTLN-diff probability law, we can obtain a nested model in the tdPGB2 distribution in which the density function is differentiable at the threshold points τ_l^*, τ_u^*. We denote this statistical model by tdPGB2-diff. The differentiability conditions lead to the following constraints

$$\alpha^* = \frac{a(p - q(\tau_l^*/b)^a)}{1 + (\tau_l^*/b)^a} \tag{14}$$

$$\beta^* = \frac{a(q(\tau_u^*/b)^a - p)}{1 + (\tau_u^*/b)^a}, \tag{15}$$

the reduced set of parameters being $(\tau_l^*, a, b, p, q, \tau_u^*)$.

Using our comparative analysis, we show that the 3LN, PTLN and tdPGB2 models are all very well suited distributions for modeling Romania's cities population; the PTLN and tdPGB2 probability laws being statistically equivalent to the 3LN model by Vuong tests.

3. Empirical analysis

In this Section, we discuss the analysis of cities' size distribution of Romania for 1992-2017 period. As we saw in Table 1, the data sets have similar values for the respective descriptive statistics, so we explicitly show the results obtained for years 1992, 2007 and 2017, and briefly mention that the results for the other years are similar. In order to assess the goodness-of-fit, we perform Kolmogorov-Smirnov (KS), Cramér-von Mises (CM) and Anderson-Darling (AD) tests. The last statistical test is useful when we are interested to see how adequate is the fit of the distribution at the tails [33]. Table 2 reports the statistics and p-values of the mentioned tests. All the three probability laws considered in this paper are clearly non-rejected by the tests. Other criteria used are the Akaike and Bayesian Information Criteria (AIC and BIC). The lower the AIC and the BIC, the better the fit.

3.1. Parameter estimates and discussion

The maximum likelihood estimates of Romania's cities population are displayed in Table 3. It can be observed that all parameter estimates are highly significant as indicated by the low standard errors.

In the case of the 3LN distribution, the estimated parameters represent the means $\hat{\mu}_i$ and standard deviations $\hat{\sigma}_i$ of the log-population of three subgroups of cities, each in proportion $\hat{\pi}_i$, that are assumed to have similar characteristics

Table 2. Statistical tests results of Romania's cities population. Non-rejections at the 5% level are in bold.

Model	Year	Test statistics (p-value)
3LN	1992	KS=**0.007(0.999)**, CM= **0.015 (0.999)** AD=**0.097 (0.999)**
	2007	KS=**0.010 (0.938)**, CM=**0.034 (0.96[illegible])** AD=**0.191 (0.993)**
	2017	KS=**0.008 (0.987)**, CM=**0.021 ([illegible].99[illegible])** AD=**0.131 (0.999)**
PTLN	1992	KS=**0.010 (0.953)**, CM=**0.[illegible]32 (0.[illegible]70)** AD=**0.205 (0.989)**
	2007	KS=**0.011 (0.812)** , CM=**0.0[illegible] ([illegible]87)** AD=**0.291 (0.945)**
	2017	KS=**0.016 (0.406)**, CM=**[illegible].111 (0.532)** AD=**0.579 (0.608)**
tdPGB2	1992	KS=**0.009 (0.978)**, CM=**0[illegible]027 (0.986)** AD= **0.183 (0.99[illegible])**
	2007	KS=**0.009 (0.974)**, CM=**0.040 (0.931)** AD=**0.254 (0.[illegible])**
	2017	KS=**0.009 (0.968)**, CM=**0.019 (0.998)** AD=**0.119 ([illegible])**

with respect to growth [22]. In fact, the means of the log-population $\hat{\mu}_i$, $i = 1, 2, 3$, are in general different from one another, and the standard deviations $\hat{\sigma}_i$, $i = 1, 2, 3$, are distinct by a considerable amount. The weights $\hat{\pi}_i$, $i = 1, 2$, also vary across samples. This may mean that the partition into growth groups may vary along time, since some cities may grow faster than others. However, as [22] also mention, the "actual factors that drive population growth of a city remain unclear", but this 3LN parametrization may lead to a new insight into the problem, to be developed in another paper or papers.

In the case of the model PTLN, the MLE estimate of lower tail switching parameter $\hat{\tau}_l$ is 926 for the year 1992, while for year 2007 is 665 and for year 2017 is 649. The Pareto exponent estimates of the PTLN model for the lower tail fluctuate in time more than Pareto exponent estimates for the upper tail for the 1992-2017 period. This is due in part because in the cases of PTLN and tdPGB2 distributions (to be shown next) there is a small percentages of UATs (under 1.5%) in the lower tails. Comparing with the results obtained for the tdPGB2 model, the Pareto exponent estimates of the PTLN model for the upper tail are higher than the Pareto exponent estimates for the tdPGB2 model for all years except the 2001-2003 period. This means that the results of the fitting of tdPGB2 model report a more unequally population distributed among UATs in the upper tail than what the results of the fitting of the PTLN model report. Some indications on why there is an unequally population distributed among UATs in the upper tail may be the urban-urban migration flow and the degree of economic development of urban areas. The upper tails of both PTLN and tdPGB2 models consist of only urban areas.

According to data provided by the National Institute of Statistics (INS), in the year 1992, on average, 5.7 out of 1,000 of urban residents changed their residential status to other urban areas, while in the year 2017, the average annual flow of internal migration from one urban area to another urban area was 8.9 people out of 1,000 inhabitants. In the year 2007, on average, 7.4 out of 1,000 of urban residents changed their residential status to other urban areas.

The upper cutoff MLE estimates of the PTLN probability law are 11,618, 10,125 and 9,486, respectively for all years considered. Most places for the PTLN distribution are estimated to be in the log-normal body ($\simeq$ 92%), while the lower tail has a low percentage of places (<1%). The dispersion estimates $\hat{\sigma}$ of the PTLN distribution are 0.516, 0.558 and 0.592, respectively for all years considered. This means that in 2017 there was a more unequally population distributed among the UATs in the log-normal body compared to the year 1992.

In the case of tdPGB2 distribution, for the year 1992, the MLE estimate of the lower Pareto exponent $\hat{\alpha}^*$ is 3.214

Table 3. Parameter estimates of Romania's cities population

Parameter estimators and criteria information	1992 (standard errors)	2007	2017
3LN distribution			
$\hat{\mu}_1$	9.308 (0.104)	9.224 (0.105)	9.230 (0.114)
$\hat{\sigma}_1$	1.549 (0.068)	1.582 (0.069)	1.6[illegible] (0.075)
$\hat{\mu}_2$	8.253 (0.054)	8.029 (0.011)	7.863 (0.026)
$\hat{\sigma}_2$	0.365 (0.044)	0.530 (0.009)	0.4[illegible]5 (0.021)
$\hat{\mu}_3$	8.189 (0.013)	9.259 (0.105)	8.103 (0.019)
$\hat{\sigma}_3$	0.521 (0.010)	0.253 (0.107)	0.662 (0.015)
$\hat{\pi}_1$	0.107 (0.008)	0.102 (0.00[illegible])	0.096 (0.009)
$\hat{\pi}_2$	0.136 (0.004)	0.882 (0.005)	0.317 (0.031)
log-likelihood	-27,415	-29,306	-29,441
AIC	54,847	58,628	58,897
BIC	54,895	58,677	58,946
PTLN distribution			
$\hat{\alpha}$	2.263 (0.309)	2.237 (0.[illegible]54)	2.537 (0.340)
$\hat{\tau}_l$	926 (46)	6[illegible]5 (43)	649 (47)
$\hat{\mu}$	8.207 (0.009)	8.050 (0.009)	8.007 (0.010)
$\hat{\sigma}$	0.516 (0.006)	0.558 (0.007)	0.592 (0.007)
$\hat{\tau}_u$	11,618 (257)	10,125 (251)	9,486 (293)
$\hat{\beta}$	0.962 (0.051)	1.039 (0.050)	1.136 (0.050)
log-likelihood	-27,420	-29,312	-29,449
AIC	54,851	58,637	58,910
BIC	54,887	58,673	58,946
tdPGB2 distribution			
$\hat{\alpha}^*$	3.214 (0.156)	2.168 (0.384)	2.669 (0.299)
$\hat{\tau}_l{}^*$	1,725 (171)	608 (52)	731 (107)
$\hat{a}$	1.845 (0.047)	2.020 (0.026)	1.844 (0.024)
$\hat{b}$	3,0[illegible]6.9[illegible] (37.126)	2,343.883 (23.974)	1,996.868 (21.810)
$\hat{p}$	1.163 (0.027)	2.347 (0.037)	2.686 (0.043)
$\hat{q}$	1.149 (0.019)	1.422 (0.020)	1.400 (0.020)
$\hat{\tau}_u{}^*$	13,941 (445)	14,178 (551)	14,520 (656)
$\hat{\beta}^*$	0.904 (0.054)	0.931 (0.055)	0.997 (0.058)
log-likelihood	-27,420	-29,309	-29,441
AIC	54,855	58,633	58,899
BIC	54,902	58,682	58,947

which changes to 2.669 for the year 2017, having a value of 2.168 corresponding to 2007. On the other hand, the upper Pareto exponent estimates are 0.904, 0.931 and 0.997, respectively. The Pareto exponents for the lower tail of the tdPGB2 model fluctuate in time more than Pareto exponents for the upper tail for the 1992-2017 period. This means that the population distribution among the UATs in the upper tail is less likely to change than the population distribution among UATs in the lower tail. As an observation, all the locations in the upper tail are from the urban area and hold approximative 50% of the total population for each year resulting in a small percent of places being in the lower tail, between 0.09% (2017) and 1.30% (1992). The number of places in the upper tail has increased from 146 places in 1992 to 151 places in 2017, but the percentage has decreased slightly from 4.95% to 4.74%.

The estimates of lower tail switching point $\hat{\tau}_l{}^*$ are 1,725, 608 and 731, respectively, for all years 1992, 2007 and

2017, while those of the upper cutoff are 13,941, 14,178 and 14,520, respectively. Comparison of our upper tail cutoff point and lower tail switching estimates for 1992 data to that for 2017 data reveals that a smaller portion of cities and population was in the GB2 body (2,573 places, 87.33% of all cities, or 48.85% percent of the population) most of whom lived in the rural area (86.66% percent of the population) for the 1992 data relative to the 2017 data (2,990 places, 94% of all places, or 49.61% of the population out of which 86.51% lived in the rural area).

Let us remark that the PTLN and tdPGB2 models have in general different bodies and the fit of the body is not equal in general. Since the specifications are continuous everywhere, the overall fit of the distributions may imply different values of the cutoff or threshold values by this mathematical requirement of continuity, and by the same reason the values of the Pareto exponents may change as well among these two distributions.

Table 4. Zipf's test results

	t-statistic (p-value) PTLN	tdPGB2
1992	**-0.745 (0.456)**	**-1.778 (0.075)**
2007	**0.780 (0.435)**	**-1.255 (0.210)**
2017	2.72 (0.007)	**-0.052 (0.959)**

Table 5. Vuong test results

	Vuong statistic (p-value) 3LN vs PTLN	3LN vs tdPGB2
1992	1.248 (0.212)	1.221 (0.222)
2007	1.503 (0.1[illegible])	0.894 (0.372)
2017	1.875 (0.061)	0.315 (0.753)

Since the fulfillment of Zipf's law is an issue of importance in the literature, that is, that the Pareto exponent for the upper tail is equal to one, let us perform a simple t-statistic test to assess whether for the estimated cases of the PTLN and tdPGB2 distributions we can reject that the upper tail Pareto exponents are one. The results are shown in Table 4. In short, the null hypothesis of upper tail Pareto exponent equal to one is rejected at the 5% level of significance only for the PTLN model in 2017. By contrast, the tdPGB2 model in the same year shows a clear non-rejection of the null. In all other cases, the null is also non-rejected, so the proposed PTLN and tdPGB2 models are capable of reproducing the Zipf's law regularity for Romanian data to a great extent.

Looking at the information criteria given by the AIC and BIC, we notice that for years 1992 and 2007 the PTLN model has the lowest BIC, thus making it the more appropriate distribution among the three considered in this paper, to fit the data. The 3LN model has the lowest AIC for all years, and the lowest BIC for year 2017 which is equal to the BIC value of PTLN model. The Vuong tests' results are displayed in Table 5 for 3LN model against PTLN and tdPGB2 probability laws. Vuong's closeness test for all three years yields that the 3LN model cannot be rejected to be statistically equivalent to the PTLN and to the tdPGB2 models. The Bayes factor which can be approximated by $BF \simeq \exp\left(\frac{1}{2}\left(BIC^u - BIC^r\right)\right)$ can be interpreted using Jeffrey's scale [34]. If $BF < 0.1$, then we have strong support for model u, if $0.1 < BF < 1/3$ then the support is moderate, while a Bayes factor greater than $1/3$ suggests a weak support for the model chosen. The results of the Bayes factor tests are displayed in Table 6. There is strong support for the PTLN model for years 1992 and 2007, while for year 2017 there is weak support for either PTLN or 3LN models. The latter suggests that for this year, all three models can be considered as suitable fits for the data, the differences between the BIC values being small. However, the 3LN model has the lowest AIC. Also, there is a moderate support for PTLN probability law against 3LN model for year 2007, while for 3LN model there is a strong support for years 1992 and 2007 against tdPGB2 model. The analysis so performed shows that the 3LN, PTLN and tdPGB2 models fit very appropriately the Romanian city size distribution. The final preference for one over another may depend on the desire for accuracy in the results versus the simplicity of the model. We have shown a slight statistical preference of the PTLN distribution over the other two, but if one prefers a model with the greatest simplicity of specification, estimation and computation [22] one might choose the 3LN model instead, but then the modeling of the tails as Pareto is lost.

Table 6. Bayes factor

	1992	2007	2017
PTLN vs tdPGB2	<0.001	0.01	0.60
PTLN vs 3LN	0.018	0.13	1
3LN vs tdPGB2	0.03	0.08	0.60
3LN vs PTLN	-	-	1

For the sake of comparison, let us analyze in brief the results for the alternative PTLN-diff and tdPGB2-diff models where differentiability at the threshold points is imposed by means of the constraints (7), (8) and (14), (15), respectively. In Table 7 we show the estimated parameter values, the maximum log-likelihoods and the AIC and BIC information criteria. We show only the quantities for the years 2007 and 2017 because we have not been able to estimate the PTLN-diff model for the year 1992. As expected, since the differentiable models are nested into the non-differentiable ones, the values of the maximum log-likelihoods are lower (or at most, equal) than for the non-differentiable models. In Table 8 we show the results of the corresponding KS, CM and AD tests. There are more rejections of the differentiable models than the non-differentiable ones, so the goodness-of-fit is in general worse for the former ones. Nevertheless, we can observe that the goodness-of-fit of the differentiable models improves with later samples of Romanian data. We have performed as well standard log-likelihood ratio tests between PTLN-diff and PTLN distributions on the one hand and tdPGB2-diff and tdPGB2 distributions on the other hand to see if they are statistically equivalent (that being the null hypothesis) or if the more complex models (the non-differentiable ones at the threshold values) are favored. The results are shown in Table 9. The rejection of the null is clear always and the non-differentiable models are significantly selected.

3.2. *Graphical analyses*

Figs. 3, 4 and 5 graph the rank-size plots for ascending and descending city sizes in log-log scale. The solid green line represents the empirical city sizes, while the red, blue and purple lines depict the predicted city sizes using Eqs. (3), (6) and (13), and the parameter estimates given in Table 3 for the 3LN, PTLN and tdPGB2 distributions, respectively. These graphs show that all three models predict accurately the city sizes for the upper and lower tails.

4. Conclusion

Romania's cities population is very well modeled by the 3LN, PTLN and tdPGB2 distributions. The statistical tests KS, CM and AD provide substantial evidence that the Romanian's cities size can be easily predicted by these models. The Vuong tests prove that we cannot reject that the PTLN and tdPGB2 models are statistically equivalent to the 3LN probability law for all years. In conclusion, there are models that are clearly not rejected for the same samples, and only some of them have Pareto tails. Thus the question of having Pareto tails or not is quite interesting, since both possibilities may occur at the same time [35]. For 1992 and 2007, the tests applied provide support for the PTLN distribution. As for 2017, opting for one model or other is not quite easy as all of them provide similar performances. If one selects the simpler model in terms of specification and computation [22] one might favour the 3LN distribution, but then the modeling of the tails as Pareto is lost.

Acknowledgements

The authors would like to thank the Editor and the two referees for careful reading and comments which greatly improved the paper. The work of Miguel Puente-Ajovín and Arturo Ramos has been supported by the Spanish *Ministerio de Economía y Competitividad* (ECO2017-82246-P) and by Aragon Government (ADETRE Reference Group).

Table 7. Parameter estimates of Romania's cities population for years 2007 and 2017

Parameter estimators and criteria information	2007 (standard errors)	2017
PTLN-diff distribution		
$\hat{\tau}_l$	1,291 (58)	1,167 (57)
$\hat{\mu}$	7.999 (0.006)	7.945 (0.00[illegible])
$\hat{\sigma}$	0.506 (0.005)	0.537 (0.005)
$\hat{\tau}_u$	4,424 (42)	4,3[illegible] (44)
log-likelihood	-29,347.7	-29,464.6
AIC	58,703.3	58,937.2
BIC	58,727.6	58,961.5
tdPGB2-diff distribution		
$\hat{\tau}_l^*$	1,209 (76)	1,[illegible]7 (102)
$\hat{a}$	0.800 (0.006)	[illegible].123 (0.021)
$\hat{b}$	6,614 (37)	[illegible],408 (18)
$\hat{p}$	9.950 (0.04[illegible])	1.945 (0.025)
$\hat{q}$	18.845 (0.078)	1.147 (0.016)
$\hat{\tau}_u^*$	4,282 ([illegible])	4,470 (66)
log-likelihood	-29,347.5	-29,464.5
AIC	58,707	58,941
BIC	58,743.4	58,977.4

References

[1] J. Luckstead, S. Devadoss, A comparison of city size distributions for China and India from 1950 to 2010, Economics Letters, 124 (2) (2014) 290-295.

[2] J. Luckstead and S. Devadoss, Do the World's Largest Cities Follow Zipf's Law and Gibrat's Law?, Economics Letters, 125(2) (2014) 182-186.

[3] J. Luckstead and S. Devadoss, A Nonparametric Analysis of the Growth Process of Indian Cities, Economics Letters, 124(3) (2014) 516-519.

[4] R. González-Val, A. Ramos, F. Sanz-Gracia, M. Vera-Cabello, Size distribution for all cities: Which one is best?, Papers in Regional Science, 94 (1) (2015) 177-197.

[5] Y. Ioannides, S. Skouras, US city size distribution: Robustly pareto, but only in the tail, Journal of Urban Economics, 73 (1) (2013) 18-29.

[6] J. Luckstead, S. Devadoss, Pareto tails and log-normal body of US cities size distribution, Physica A: Statistical Mechanics and its Applications, 465 (2017) 573-578.

[7] J. Luckstead, S. Devadoss, and D. Danforth, The Size Distributions of All Indian Cities, Physica A: Statistical Mechanics and its Applications, 474 (2017) 237-249.

Table 8. Statistical tests results of Romania's cities population. Non-rejections at the 5% level are in bold.

Model	Year	Test statistics (p-value)	
PTLN-diff	2007	KS=0.029 (0.011) , AD=2.876 (0.032)	CM=**0.314 (0.124)**
	2017	KS=**0.020 (0.186)**, AD=**1.561 (0.162)**	CM=**0.154 (0.379)**
tdPGB2-diff	2007	KS=0.028 (0.016), AD=2.605 (0.044)	CM=**0.276 (0.158)**
	2017	KS=**0.017 (0.330)**, AD=**1.319 (0.226)**	CM=**0.122 (0.488)**

Table 9. Log-likelihood ratio test results.

	llr test statistic (p-value) PTLN-diff vs PTLN	tdPGB2-diff vs tdPGB2
2007	70.542 (0.000)	77.754 (0.000)
2017	31.195 (0.000)	46.302 (0.000)

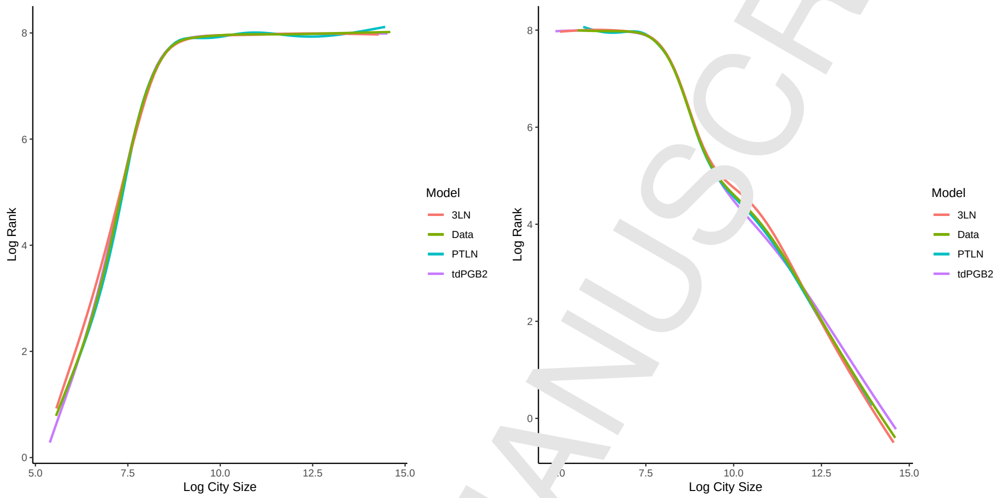

Figure 3. Rank-size plot for ascending and descending city sizes for year 1992

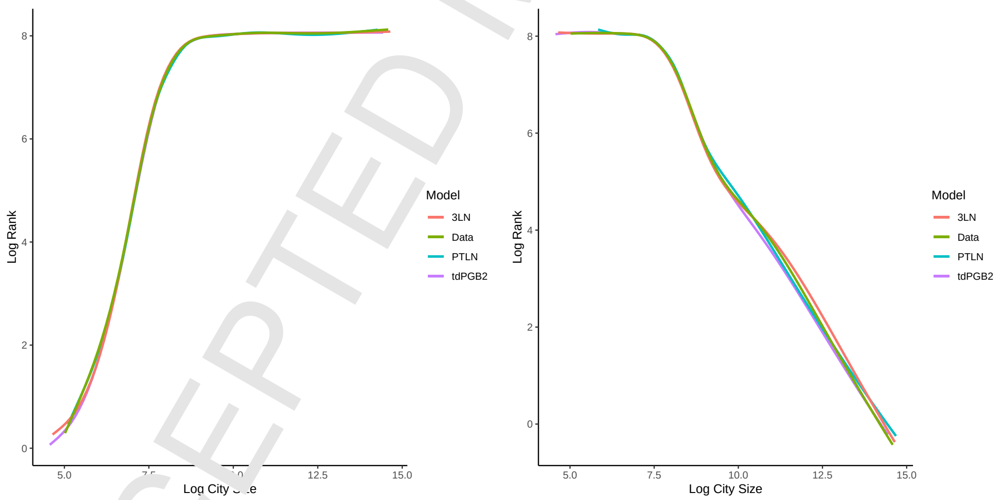

Figure 4. Rank-size plot for ascending and descending city sizes for year 2007

[8] S. Devadoss, J. Luckstead, Size distribution of US lower tail cities, Physica A: Statistical Mechanics and its Applications, 444 (2016) 158-162.

[9] L.C. Malacarne, R.S. Mendes, E.K. Lenzi, q-Exponential distribution in urban agglomeration, Physical Review E, 65 (1) (2001) 017106.

[10] K. Gangopadhyay, B. Basu, City size distributions for India and China, Physica A: Statistical Mechanics and its Applications, 388 (13) (2009) 2682-2688.

[11] I. Băncescu, q-Log-distributions. Log-concavity and Log-convexity, The European Physical Journal Plus, 133 (2018) 1-18.

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

[13] A. Ramos, F. Sanz-Gracia, R. González-Val, On the parametric description of US city size distribution: new empirical evidence, (2014)

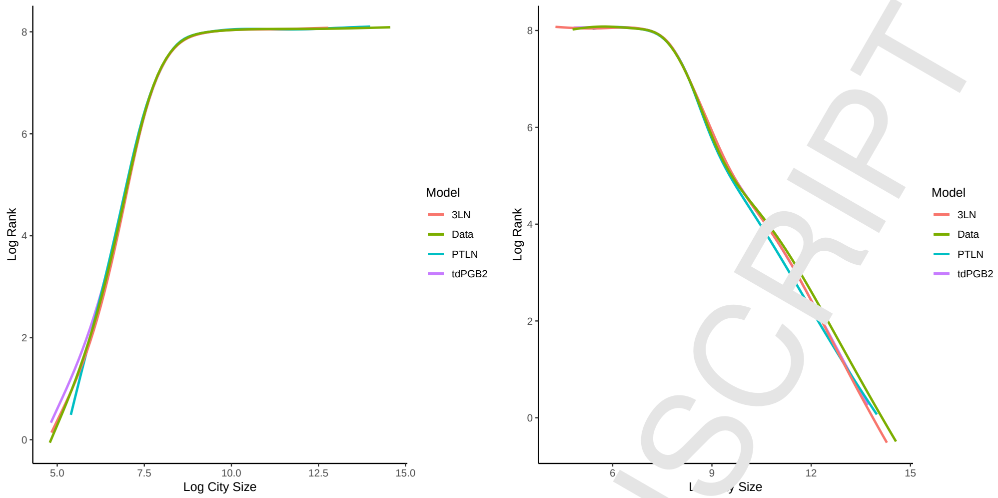

Figure 5. Rank-size plot for ascending and descending city sizes for year 2017

Working paper. Available at Munich RePEC. http://mpra.ub.uni-muenchen.de/57645/
[14] D. Bunea, Modern Gravity Models of Internal Migration. The Case of Romania, Theoretical and Applied Economics, 4(4) (2012) 127-144.
[15] I. Alexe, I. Horváth, R. Noica, M. Radu, Alexe, I., Social impact of emigration and rural-urban migration in Central and Eastern Europe: final country report; Romania. EC DG Employment, (2012) Social Affairs and Inclusion. http://ec.europa.eu/social/main.jsp?langId=en&catId=89&newsId=1778&furtherNews=yes
[16] M. Kupiszewski, D. Berinde, V. Teodorescu, H. Durham and P. Rees, Internal migration and regional population dynamics in Europe: Romanian case study, (1997) Working Paper, School of Geography, University of Leeds.
[17] V. Gheţău, Our demographic distress. Romania's population according to the October 2011 census, Compania Publishing House, Bucharest, 2(3) (2012) 234-238.
[18] I. Ianos, Causal Relationships Between Economic Dynamics and Migration: Romania as Case Study, Springer Science+Business Media Singapore J. Domínguez-Mujica (ed.), Global Change and Human Mobility, Advances in Geographical and Environmental Sciences, (2016) DOI 10.1007/978-981-10-0050-8_16
[19] L. Chivu, C. Ciutacu, Romania and the Four Economic Freedoms: From Theory to Practice, in "Economic Dynamics and Sustainable Development - Resources, Factors, Structures and Policies" (Proceedings ESPERA 2015, Part 1), ISBN: 9783631696644, Peter Lang Academic Publishing Group, Frankfurt, (2016).
[20] D. Sandu, C. Radu, M. Constantinescu, and O. Ciobanu, A country report on Romanian migration abroad: stocks and flows after 1989, Multicultural Center Prague, Ruth Ferrero Turrión (ed.), November, 2004).
[21] D. Sandu, Destination Selection Among Romanian Migrants in Times of Crisis: an Origin Integrated Approach, Romanian Journal of Population Studies, 11(2) (2017) 145-1[illegible]
[22] H.S. Kwong, S. Nadarajah, A note on "Pareto tails and log-normal body of US cities size distribution", Physica A: Statistical Mechanics and its Applications, 513 (2019) 55-62.
[23] D.M. Titterington, Some problems with data from finite mixture distributions, Technical Summary Report #2369. University of Wisconsin-Madison (1982).
[24] G. McLachlan and D. Peel, Finite Mixture Models, John Wiley & Sons, New York, (2000).
[25] M., Bee, M. Riccaboni, S. Schiavo. Pareto versus log-normal: A maximum entropy test, Physical Review E, 84(2) (2011) 026104.
[26] J. Growiec, F. Pammolli, M. Riccaboni, H.E. Stanley, On the size distribution of business firms, Economics Letters, 98(2) (2008) 207-212.
[27] J.B. McDonald, Y.J. Xu, A generalization of the beta distribution with applications, Journal of Econometrics, 66 (1995) 133-152.
[28] J.B. McDonald, Some generalized functions for the size distribution of income, Econometrica, 52 (1984) 647-663.
[29] J.B. McDonald, A. Mantrala, Apples, oranges and distribution trees. Discussion paper (1993), Brigham Young University, Provo, Utah.
[30] J.B. McDonald, A. Mantrala, The distribution of income, revisited, Journal of Applied Econometrics, 10 (1995) 201-204.
[31] C. Kleiber, S. Kotz, Statistical size distributions in economics and actuarial sciences (2003), Wiley, London.
[32] A. Ramos, F. Sanz-Gracia, US city size distribution revisited: Theory and empirical evidence, (2015) Working paper. Available at Munich RePEC. https://mpra.ub.uni-muenchen.de/71928/
[33] P. Cirillo, Are your data really Pareto distributed?, Physica A: Statistical Mechanics and its Applications, 392 (2013) 5947-5962.
[34] R.E. Kass, A.E. Raftery, Bayes factors, Journal of the American Statistical Association, 90 (430) (1995) 773-795.
[35] M. Bee, M. Riccaboni, S. Schiavo, The size distribution of US cities: Not Pareto, even in the tail, Economics Letters, 120 (2013) 232-237.