Available online at www.sciencedirect.com
ScienceDirect
Economics Letters 00 (2020) 1–4
Economics Letters

How sensitive is city size distribution to the definition of city? The case of Spain

Miguel Puente-Ajovín[a,*], Arturo Ramos[a,1], Fernando Sanz-Gracia[a,1], Daniel Arribas-Bel[b,1]

[a] *Departamento de Análisis Económico, Universidad de Zaragoza, Gran Vía 2, 50005, Zaragoza, Spain*
[b] *Geographic Data Science Lab, Department of Geography and Planning, University of Liverpool, Roxby Building, 74 Bedford St S, Liverpool, L69 7ZT, United Kingdom*

Abstract

In this paper we want to test whether the choice of different types of urban data for the same country exerts an influence or not on the selection of the best parametric density function (among the Pareto, truncated lognormal, the double Pareto lognormal and mixtures of lognormals) to describe the city size distribution. We have employed four different definitions of city for Spain. We have concluded that the outperforming density is different for each type of data.

Keywords: Pareto distribution, Lognormal distribution, double Pareto lognormal, mixtures, Cities, Spain
JEL Classification Code: C13, C16, R1

1. Introduction

Starting on 2015 in Schmidheiny and Suedekum (2015), there has been an explosion of the study of new methods of delineating and defining urban areas in the years 2019-2020 using building density, machine learning, personal judgement and others (Arribas-Bel et al., 2019; de Bellefon et al., 2020; Ch et al., 2020; Moreno-Monroy et al., 2020; Galdo et al., 2020).

The elucidation of how robust the parametric description of the city size distribution is to different definitions of cities has its antecedents in papers like Ioannides and Skouras (2013) and Bee et al. (2013), although in these references the focus was about whether the Pareto specification is appropriate, a debate that it is still not over.

In this context, in this paper we attempt to answer the following question: when studying the city size distribution of a country (in our case Spain), are the results robust to different definitions of what a city is? To do so, we use four different definitions of 'city' (the new proposed in Arribas-Bel et al. (2019) and other three previously introduced). The answer to the question is straightforward: the best density for each type of data is specific to it; thus city size distribution is sensitive to the definition of city considered.

*Corresponding author: mpajovin@unizar.es
[1] aramos@unizar.es (A. Ramos), fsanz@unizar.es (F. Sanz-Gracia), d.arribas-bel@liverpool.ac.uk (D. Arribas-Bel)

2. Distributions

We let x denote the size (inhabitants) of the urban unit in question.

The first distribution we consider is the power law distribution (or Pareto distribution), with density:

$$f_{\mathrm{P}}(x;\alpha,x_{\min}) = \frac{\alpha}{x_{\min}}\left(\frac{x}{x_{\min}}\right)^{-\alpha-1}$$

where α is the power law exponent and $x_{\min}$ is the minimum value of the data.

However, one issue that has been highlighted in the literature is that it can be difficult to distinguish between a power law distribution on the upper tail from other heavy tailed distributions (Clauset et al., 2009). An implication of this observation is that an empirical analysis should also consider the fits of alternative distributions to the upper tail. Consequently, we also consider a truncated lognormal (LNt)

$$f_{\mathrm{LNt}}(x;\mu,\sigma,x_{\min}) = \frac{f_{\mathrm{LN}}(x;\mu,\sigma)}{1-\mathrm{cdf}_{\mathrm{LN}}(x_{\min};\mu,\sigma)}$$

where $\mathrm{cdf}_{\mathrm{LN}}$ is the cumulative distribution function of the lognormal (LN) function, whose density is given by

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

The fourth distribution in our study is the double Pareto lognormal distribution (dPLN), introduced in Reed (2002, 2003) and used for city size distributions in, e.g., Giesen et al. (2010); González-Val et al. (2015).

Finally, following Kwong and Nadarajah (2019), we consider mixtures of two (2LN) and three (3LN) lognormals.

We estimate all these distributions with maximum likelihood (ML) estimation, with a procedure similar to that of Puente-Ajovín et al. (2020) and the detailed results are available from the authors upon request.

We will assess the goodness-of-fit by means of standard statistical tests as the Kolmogorov–Smirnov (KS), Cramér–von Mises (CM) and Anderson–Darling (AD) tests (see, e.g., Massey (1951); Anderson (1962); Anderson and Darling (1952) and references therein).

We will use three well-known information criteria (IC). They are the AIC (Akaike), BIC (Bayesian) and HQC (Hannan-Quinn, see, e.g.,Burnham and Anderson (2002, 2004) and references therein). The smaller value of each of these criteria, the more preferred the model according to each criterion.

3. Data

There are several ways of defining a city or an urban area. Because of that, the choice of the specific definition can affect the outcome of the analyzed city size distribution. In our analysis we use four different definitions to do so.

The first is that of Spanish Municipalities: This database is based on the Spain's Census of 2011. It covers 100% of the population, distributed along 8,115 municipalities, which is officially the smallest administrative division.

The second is BMLA: By this new acronym we mean "Buildings Machine Learning Algorithm", and is the database of city centers of Arribas-Bel et al. (2019), which uses a new methodology that delineates 717 urban areas using a machine learning algorithm that groups buildings existing in a space of a sufficient high population density, using the geolocation of all of 12 million buildings in Spain. In this case, the minimum amount of population for an area to be considered is 1,000 inhabitants, covering 74.8% of all the population of Spain.

The third is AUDES: This specification identifies the Urban Areas of Spain following a particular method based on Mendelson and Lefebvre (2003)[2]. The key is to look for urban cores with more than 10,000 residents, formed by adjacent urban entities. Urban Areas are then defined as one or more municipalities that are situated about that urban core and have more than 20,000 residents. Because of this cut-off condition, the database covers 77.4% of the population, being 263 the number of Urban Areas.

The fourth is that of Functional Urban Areas (FUA OECD): Based on the requirements of Dijkstra et al. (2019), a functional urban area is the combination of the city with its commuting zone. They have a greater cut-off than that of BMLA, and thus this database covers 69.7% of all of Spain's population in 80 Functional Urban Areas.

[2]The full description of the method can be found in: https://alarcos.esi.uclm.es/per/fruiz/audes/modelo.htm

4. Results

For each type of data six densities are estimated: Pareto, LNt, LN, dPLN, 2LN and 3LN. The Pareto and LNt are not estimated for Municipalities and BMLA because they are clearly non-truncated data sets. The dPLN cannot be estimated for the AUDES and FUA OECD datasets because the ML estimators there seem not to exist in those cases.

	Pareto			LNt		
	KS	CM	AD	KS	CM	AD
Municipalities	–	–	–	–	–	–
BMLA	–	–	–	–	–	–
AUDES	**0.771 (0.041)**	**0.671 (0.084)**	**0.670 (0.576)**	**0.615 (0.047)**	**0.569 (0.103)**	**0.571 (0.686)**
FUA OECD	**0.197 (0.120)**	**0.123 (0.315)**	**0.132 (1.717)**	**0.992 (0.049)**	**0.975 (0.030)**	**0.938 (0.300)**
	LN			dPLN		
	KS	CM	AD	KS	CM	AD
Municipalities	0 (0.050)	0 (5.767)	0 (35.403)	0 (0.032)	0 (2.531)	0 (16.161)
BMLA	0 (0.097)	0 (2.561)	0 (14.564)	**0.810 (0.024)**	**0.846 (0.055)**	**0.892 (0.355)**
AUDES	0 (0.166)	0 (1.896)	0 (5.535)	–	–	–
FUA OECD	**0.326 (0.106)**	**0.192 (0.248)**	**0.135 (1.703)**	–	–	–
	2LN			3LN		
	KS	CM	AD	KS	CM	AD
Municipalities	**0.495 (0.010)**	**0.587 (0.099)**	**0.450 (0.844)**	**0.618 (0.009)**	**0.941 (0.038)**	**0.847 (0.402)**
BMLA	**0.589 (0.029)**	**0.584 (0.100)**	**0.667 (0.580)**	**0.999 (0.014)**	**0.999 (0.011)**	**0.999 (0.099)**
AUDES	**0.282 (0.061)**	**0.301 (0.184)**	**0.156 (1.592)**	**0.319 (0.059)**	**0.567 (0.104)**	**0.371 (0.974)**
FUA OECD	**0.991 (0.049)**	**0.919 (0.043)**	**0.878 (0.370)**	**0.988 (0.050)**	**0.997 (0.020)**	**0.999 (0.150)**

Table 1: Outcomes of the statistical tests. The format is p-value (statistic). Non-rejections at the 5% level are marked in bold

Table 1 shows the results of the KS, CM and AD tests for the six density functions and four definitions of city. As can be seen, the AUDES and FUA OECD datasets are fitted well by either Pareto or LNt, the LN is almost always rejected (except for the FUA OECD specification) and the 2LN and 3LN are never rejected for all kinds of considered data.

Table 2 shows, for each type of data, the selected density according to the standard information criteria AIC, BIC and HQC. The three criteria are coincidental for each type of data, except BIC for municipalities, but according to the other two the 3LN is the best function and so, this is our choice for this type of data.

	Pareto				LNt			
	log-likelihood	AIC	BIC	HQC	log-likelihood	AIC	BIC	HQC
Municipalities	–	–	–	–	–	–	–	–
BMLA	–	–	–	–	–	–	–	–
AUDES	-3146	**6295**	**6298**	**6296**	-3147	6298	6305	6301
FUA OECD	-1074	2151	2153	2152	-1072	**2148**	**2152**	**2150**
	LN				dPLN			
	log-likelihood	AIC	BIC	HQC	log-likelihood	AIC	BIC	HQC
Municipalities	-69851	139705	139719	139710	-69737	139481	139509	139509
BMLA	-7986	15976	15985	15979	-7905	**15819**	**15837**	**15826**
AUDES	-3250	6505	6512	6508	–	–	–	–
FUA OECD	-1087	2179	2184	2181	–	–	–	–
	2LN				3LN			
	log-likelihood	AIC	BIC	HQC	log-likelihood	AIC	BIC	HQC
Municipalities	-69598	139207	**139242**	139219	-69587	**139191**	139247	**139210**
BMLA	-7908	15827	15849	15835	-7902	15821	15857	15835
AUDES	-3181	6372	6390	6379	-3169	6354	6383	6366
FUA OECD	-1078	2166	2178	2171	-1074	2165	2184	2172

Table 2: Outcomes of the maximum log-likelihoods and information criteria. Of these last ones, the lowest value in each case is marked in bold

5. Conclusions

We have analyzed which is the best density to describe four different types of Spanish urban nuclei data. And this density is different for each type. For municipalities the chosen function is the three-lognormal (3LN), for the cities defined in Arribas-Bel et al. (2019) the chosen function is the double Pareto-lognormal (dPLN), for AUDES (*Áreas Urbanas de España*) data the best density is the Pareto one and for FUA OECD (Functional Urban Areas) data the outperforming function is the truncated lognormal; in each case, the selected distributions are never rejected according to the information in Table 1. This result is related to the main conclusion derived from Puente-Ajovín et al. (2020), that is to say, different countries are best described (in terms of information criteria) by different densities and not by a single dominating one, although there maybe densities that are not rejected (almost) always by standard statistical tests.

References

Anderson, T. W. (1962). On the distribution of the two sample Cramér–Von Mises criterion. *The Annals of Mathematical Statistics*, 33:1148–1159.

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23:193–212.

Arribas-Bel, D., Garcia-López, M.-À., and Viladecans-Marsal, E. (2019). Building(s and) cities: Delineating urban areas with a machine learning algorithm. *Journal of Urban Economics*, page 103217.

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

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

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

Ch, R., Martin, D. A., and Vargas, J. F. (2020). Measuring the size and growth of cities using nighttime light. *Journal of Urban Economics*.

Clauset, A., Shalizi, C. R., and Newman, E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

de Bellefon, M.-P., Combes, P.-P., Duranton, G., Gobillon, L., and Gorin, C. (2020). Delineating urban areas using building density. *Journal of Urban Economics*.

Dijkstra, L., Poelman, H., and Veneri, P. (2019). The eu-oecd definition of a functional urban area. *oecd-ilibrary.org*.

Galdo, V., Li, Y., and Rama, M. (2020). Identifying urban areas by combining human judgment and machine learning: An application to india. *Journal of Urban Economics*.

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., 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.

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

Kwong, H. S. and Nadarajah, S. (2019). A note on "Pareto tails and lognormal body of US cities size distribution". *Physica A: Statistical Mechanics and its Applications*, 513(C):55–62.

Massey, F. J. (1951). The Kolmogorov–Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46:68–78.

Mendelson, R. and Lefebvre, J. (2003). *Reviewing census metropolitan areas (CMA) and census agglomerations (CA) in Canada according to metropolitan functionality*. Statistics Canada.

Moreno-Monroy, A. I., Schiavina, M., and Veneri, P. (2020). Metropolitan areas in the world. delineation and population trends. *Journal of Urban Economics*.

Puente-Ajovín, M., Ramos, A., and Sanz-Gracia, F. (2020). Is there a universal parametric city size distribution? empirical evidence for 70 countries. *The Annals of Regional Science*.

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

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

Schmidheiny, K. and Suedekum, J. (2015). The pan-european population distribution across consistently defined functional urban areas. *Economics Letters*, 133:10–13.