Journal of Applied Sciences

ISSN 1812-5654

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Journal of Applied Sciences

ISSN 1812-5654
DOI: 10.3923/jas.2016.230.235

Research Article
Overall Size Distribution of Cities and Towns in India: 2011 Census

[1]A. Subbarayan and [2]G. Kumar

[1]Department of Master of Computer Applications, Sri Ramaswamy Memorial (SRM) University, Kattankulathur, Chennai, 603 203 Tamil Nadu, India
[2]Department of Statistics, Sri Ramaswamy Memorial (SRM) Arts and Science College, Kattankulathur, Chennai, 603 203 Tamil Nadu, India

Abstract

This study analyzes the overall size distribution of cities and towns of India in 2011. The size distribution of urban agglomerations does not follow the Zipf's law. It is shown that lognormal (LN) distribution is fair at best and there are notable deviations, which occur over the entire range of cities and towns. Then it have attempted to fit Double Pareto Lognormal (DPLN) distribution for the entire range of cities and towns. The fit clearly indicates that DPLN fits the data much better than LN based on the computed values of Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC). It has also compared the model performance using LR tests and Bayes factors. The results are consistent with the earlier findings for city sizes in the US and other countries.

Key words: City size distribution, Zipf's Law, Pareto distribution, lognormal distribution, double Pareto lognormal distribution, likelihood ratio test

Received: March 24, 2016 Accepted: April 08, 2016 Published: April 15, 2016

Citation: A. Subbarayan and G. Kumar, 2016. Overall size distribution of cities and towns in India: 2011 census. J. Applied Sci., 16: 230-235.

Corresponding Author: A. Subbarayan, Department of Master of Computer Applications, Sri Ramaswamy Memorial (SRM) University, Kattankulathur, Chennai, 603 203 Tamil Nadu, India Tel: +919444219687

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Systems with measurable entities (which can be defined by their size) are characterized by particular properties of their distribution. The size distribution of cities is the result of the pattern of urbanization, which results in city growth and city creation. Initial attempts to fit Pareto distribution for city size data were made[1,2]. Significant contributions for the study of city size distribution includes the study of many urban researches[3-5]. Pareto distribution breaks down when all cities are taken without size restriction and the lognormal distribution has been considered as the best representation of the city size distribution[6]. For studying the city size distribution of India and China with a threshold level, Pareto and q-exponential distributions were used[7].

Subsequently it has been proved that lognormal distribution fits well only for small and medium sized places. The true parametrization for the overall city size distribution consists of a lognormal, which then switches to Pareto behavior beyond a certain threshold size. Theoretical basis for the functional form for overall city size distribution was not provided[8,9]. The problem has been resolved and parametrization that fits the French overall city size distribution closely is the double Pareto lognormal distribution (DPLN)[10]. It is stated that DPLN has an explicit theoretical foundation and can be rationalized by an economic model that combines scale-independent urban growth with age heterogeneity a cross cities. Four statistical distributions were identified to describe city size distribution and it has been concluded that double Pareto lognormal distribution fits in a best manner[11].

MATERIALS AND METHODS

Data structure on city size distribution in India (1951-2011)

Definition of urban area under Indian census: The census of India defines urban according to several criteria. Firstly, all statutory towns-i.e., all places with a municipal corporation, municipal board cantonment board or notified town area committee etc. are defined as urban. Secondly, all other places which satisfy the following three criteria are regarded as urban: (a) A minimum population as urban, (b) At least 75% of the male working population is engaged in non-agricultural and allied services and (c) A population density of at least 400 km^{-2}. Thirdly, some other places with distinct urban characteristics are also considered as urban, even though they do not satisfy the above criteria.

Urban size class under Indian census: Census of India classifies urban centres into six classes. Urban centre with population of more than one lakh is called a city and less than one lakh is called a town. Cities accommodating population between 1-5 million are called metropolitan cities and more than five million are mega cities.

Urban population by size classification is based on the following:

Class	=	Population
I	=	Greater than 1,00,000
II	=	50,000-1,00,000
III	=	20,000-50,000
IV	=	10,000-20,000
V	=	5,000-10,000
VI	=	Less than 5,000

City size distribution in India (1951-2011): It has been presented the data in Table 1 in respect of the number of cities and towns under six classes noted above.

Urban agglomerations in India (1951-2011)

Definition of urban agglomeration in India: An urban agglomeration may consist of any one of the following three combinations:

- A town and its adjoining urban outgrowths
- Two or more contiguous towns with or without their outgrowths and
- A city and one or more adjoining towns with their outgrowths together forming a contiguous spread

There are 468 urban agglomerations (UAs) in India as per 2011 census, which has a population of 26, 48, 91 and 513.

Methods: The methodological aspects adopted for the study are discussed below. In respect of this it have considered Zipf's law, lognormal distribution and double Pareto lognormal distribution. The definition and estimation of parameters of these distributions and briefly presented and empirical findings are presented in subsequent sections.

Zipf's law: The product of the population size and the rank in the distribution of a city approximate a constant[1]. Rank size relationship for examining a wide variety of issues was studied[2]. The simplest representation of this relationship in Eq. 1 is:

$$a = pr^{-b} \tag{1}$$

Table 1: Size distribution of cities and towns in India (1951-2011)

	Class I	Class II	Class III	Class IV	Class V	Class VI	
Census year	>1,00,000	50,000-1,00,000	20,000-50,000	10,000-20,000	5,000-10,000	<5,000	Total
1951	69	107	363	571	737	372	2219
1961	108	145	478	710	634	254	2329
1971	156	208	609	848	604	237	2662
1981	227	309	797	1046	748	271	3398
1991	326	401	1033	1247	790	189	3986
2001	448	498	1389	1564	1043	235	5177
2011	505	605	1905	2233	2187	498	7933

where, a is a constant, p is the population of a particular city and r is its rank according to population size. When the exponent b equals-1, it is referred to Zipf as the rank size rule. In a natural logarithmic form this relationship can also be expressed as in Eq. 2:

$$\ln p = a+b \ln r \qquad (2)$$

where, a is the estimate of the intercept value, which is also the estimate of the natural logarithm of the population of the largest city and b is the estimate of the slope coefficient of the rank size curve. The b-coefficient is the derivative of the logarithmic function. It evaluates the percentage rate of change in population size associated with the percentage rate of change in rank. On a doubly logarithmic paper, the rank size rule suggested by Zipf's appears as a straight line descending from left to right at an angle of 45, indicating a slope of-1.

Lognormal distribution: The probability density function of the lognormal is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma x}} e^{-(\text{In}x-\mu)^2/2\sigma^2}, x > 0$$

where, μ and σ^2 are the mean and variance of lnx, which in this case denotes the natural logarithm of the population of the cities. The maximum likelihood estimates for μ and σ^2 are given by:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \text{In } \chi_i}{n} \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\text{In } (\chi_i - \hat{\mu}\right]^2$$

DPLN distribution: An extensive study on the rank-size distribution for human settlement was attempted[12]:

$$f(S) = \frac{\alpha\beta}{\alpha+\beta}\left[S^{\beta-1}e\left(\beta\mu_0 + \frac{\beta^2\sigma_0^2}{2}\right)\Phi^c\left(\frac{\log(S)-\mu_0+\beta\sigma_0^2}{\sigma_0}\right) + S^{-\alpha-1}e\left(\alpha\mu_0 + \frac{\alpha^2\sigma_0^2}{2}\right)\Phi\left(\frac{\log(S)-\mu_0-\alpha\sigma_0^2}{\sigma_0}\right)\right]$$

The parameters α and β are the coefficients to regulate the tails. The location and spread of the distribution are determined by μ_0 and σ_0, Φ and Φ^c = 1-Φ represents the normal CDF and complementary CDF, respectively. A special feature of this distribution is that if S is large, than f(s) ~ $s^{-\alpha-1}$ and if S is small, then f(s) ~ $s^{\beta-1}$. The DPLN, therefore incorporates a Pareto distribution in the upper and a reverse Pareto distribution in the lower tail.

Estimation of the parameters of DPLN by the method of maximum likelihood (MLE) and fitting of DPLN: For estimation of parameters of double Pareto lognormal, method of maximum likelihood has been proposed[13].

The steps are indicated below:

- The starting values for α and β are found by regressing (on log-log scales) descending rank vs size for large settlements and ascending rank vs size for small settlements
- Specifically if α and β are the starting values of α and β, starting values of μ and σ are determined as in Eq. 3 and 4:

$$\tilde{\mu} = \bar{y} - \frac{\beta-\alpha}{\alpha\beta} \qquad (3)$$

$$\tilde{\sigma} = \sqrt{s_y^2 + \left(\frac{\beta-\alpha}{\alpha\beta}\right)^2 - 2\left(\frac{\alpha^3+\beta^3}{\alpha^2\beta^2(\alpha+\beta)}\right)} \qquad (4)$$

where, $\bar{y}$ and S_y^2 are the mean and sample variance of the logarithm of observed sizes.

RESULTS AND DISCUSSION

Urban agglomerations in India: Zipf's law: Initially it would like to examine whether Zipf's law holds good for the UAs as per 2011 census data. The data was arranged, so that UAs are labeled with their respective rank. The standard rank-size regression was run as stated in Eq. 1 by simple Ordinary Least

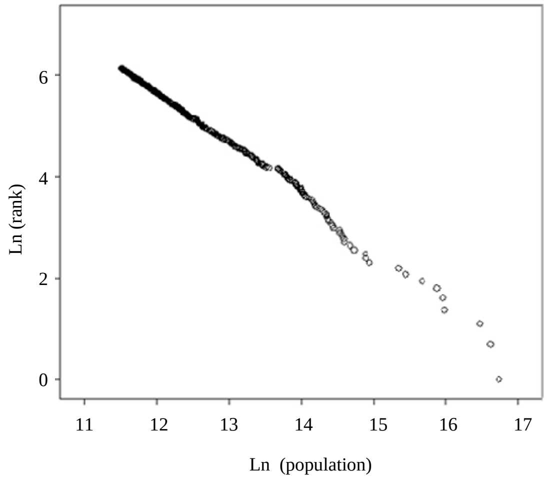

Fig. 1: Results of a Zipf regression, using 468 UAs

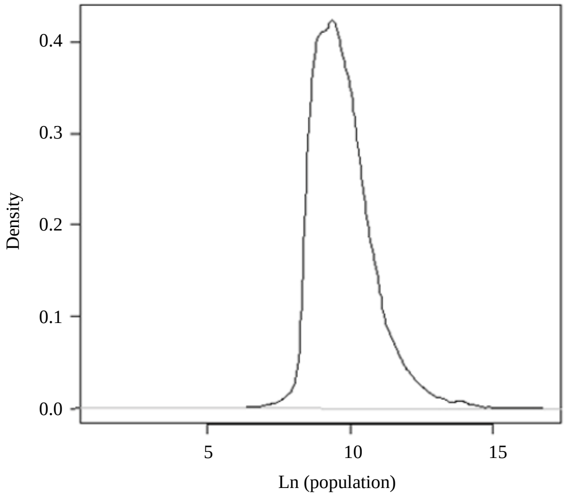

Fig. 2: Kernel density estimate of the overall size distribution

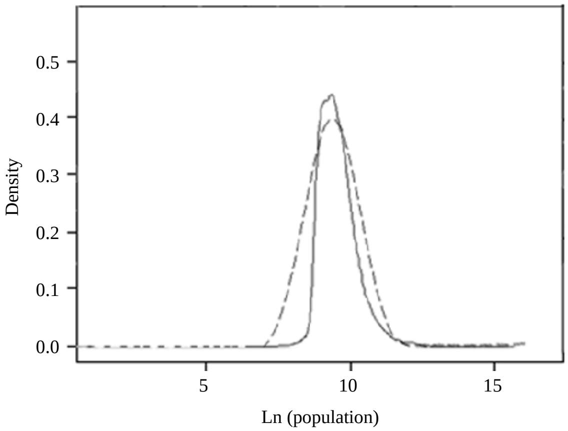

Fig. 3: Kernel density estimate and lognormal distribution

Squares (OLS). The rank size relationship is graphically illustrated in Fig. 1, where the log population size of UAs on the horizontal line and their log rank in the vertical line was depicted. When estimating the rank size regression for these 468 UAs, a slope coefficient of b = 0.969 with a standard error of σ = 0.004 was obtained. The estimated slope coefficient deviates from one and this clearly reveals that Zipf's law does not hold good for UAs in India.

Overall city size distribution 2011 census: Subsequently it has been attempted to study the overall size distributions of cities and towns in India. The analysis is based on the definition of urban as per 2001 census of India. In Fig. 2 a Kernal density estimation of the size distribution across all 7933 cities and towns was depicted, where the population sizes are in logarithmic scales, see the solid curve. It can be very well concluded that a Pareto parameterization cannot possibly fit the overall city size distribution for India. The log settlement sizes rather appear to be close, at least visually to a normal distribution.

Does the lognormal distribution fit the overall city size distribution: Lognormal distribution has been proposed for studying the settlement populations[14]. The theory relating to the convergence of overall city size distribution of a country to lognormal distribution has been established[6]. It has been depicted that a Kernel density estimation of the size distribution across all 7933 cities and towns, see the broken grey-line, which represents the fitted lognormal distribution. This is illustrated in Fig. 3. The lognormal does not feature a power law in the upper tail and hence, it is not compatible with the Zipf. It can be then investigated that, whether the suggested lognormal parameterization fits the Indian city size data. Using maximum likelihood estimation, this study find that the best fit of a lognormal parameterization to the empirical size distribution is achieved with parameters μ = 9.705 and σ = 1.076.

It is to be noted that a pure visual inspection will reveal that the overall fit of the lognormal is fair at best. It is observe that there are notable deviations, which occur over the entire range of size of cities and towns. The data also exhibits a slight skew to the left. This is a distributional feature that by the lognormal, which is symmetrical in logarithmic scales.

Results based on DPLN and a comparison with lognormal distribution: The fit for the size distribution of cities and towns in India is achieved with parameters:

$$\alpha = 1.735, \quad \beta = 1.901, \quad \mu_0 = 9.705 \text{ and } \sigma_0 = 0.741$$

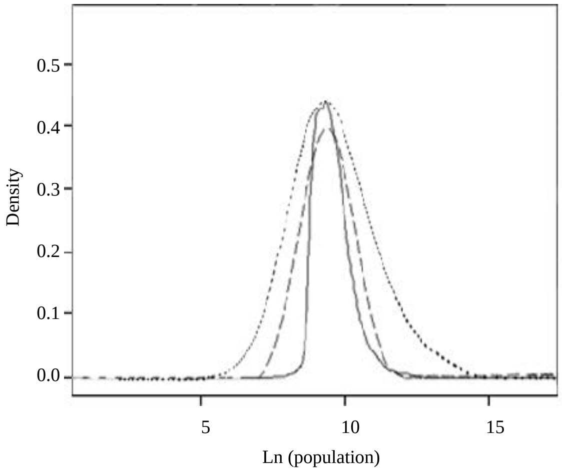

Fig. 4: Kernel density estimate, lognormal and DPLN distribution

Table 2: Estimated parameters and formal selection test

Size distribution of cities and towns in India 2011

	DPLN	LN
α	1.735	-
β	1.901	-
μ	9.705	9.755
σ	0.741	1.076
AIC	175606.002	178454.558
BIC	175633.917	178468.516
ln (L_j)	-87799.001	-89225.229
LR (p-value)	2852.556 (0.0001)	
Bayes factor	<0.0001	
Jeffrey's scale	Strong for DPLN	
N	7933	
Minimum	5	
Maximum	12442373	

In Fig. 4, the dotted black line represents the fitted DPLN distribution. Already visually it is clear that DPLN fits the data much better than LN. The DPLN is almost everywhere closely in line with empirical city size distribution, while this is not in the case for the lognormal.

The DPLN fits the data better than the lognormal almost throughout the entire range of distribution. It has been computed that the log likelihood values for the lognormal distribution and double Pareto lognormal distribution. Based on these values it have been calculated the value for Akaike's Information Criterion (AIC) and Bayes Information Criterion (BIC). The results are presented in Table 2. The DPLN with lower numerical value of AIC is favored as the best one from a statistical point of view. For comparing model performance it can use either likelihood ratio test or Bayes factors. In respect of using the likelihood ratio test, the test statistic is given by LR = 2 [ln (L_{DPLN})–ln (L_{ln})]. It has been shown that Bayes factors can be approximated by using Schwarz criterion (BIC)[15]. For comparing DPLN and LN, it can be computed that Bayes factors as B ~ exp (V), where, V = 1/2[BIC_{DPLN}-BIC_{lN}][10]. The value of B isinterpreted by using Jeffrey's scale. The result clearly indicates that there is strong evidence in favor of DPLN.

The earlier studies that analyzed size distribution of cities have mainly focused only on the upper tail (largest metropolitan areas) because of the high concentration of people in the large cities and data for large cities were readily available. Many studies have shown that Zipf's law for upper tail cities is a regularly observed phenomenon[16-19]. During the last decade urban researchers have made attempts to study in detail the overall size distribution of cities and towns in many countries of the world because of the availability of data over the entire range of cities and towns in a country[10-13]. Keeping these points in view an attempt has been made to examine the suitability of two statistical distributions viz., lognormal and double Pareto lognormal distribution for the size distribution of cities and towns in India.

The double Pareto lognormal provides an excellent date fit to the overall size distribution of cities and towns in India for the census year 2011. It is a four parameter distribution (∝, β, μ and σ) featuring a lognormal shape in the body and power law in the tails. The parameters ∝ and β are the slope parameters of the Pareto tails, where the parameters μ and σ pertain to the location and scale of normal body. In logarithmic scale, the DPLN can be skewed and its kurtosis can have positive or negative excess i.e., it can be more peaked (lepto kurtic) or more flat (platy kurtic) than the lognormal.

CONCLUSION

In this study, it has been shown that double Pareto lognormal provides an excellent fit to over a size distribution of cities and towns in India for the census year 2011. The distribution features a lognormal shape in the body and power law in the tails. The findings may reconcile the debate about city size distributions between the urban researchers and thereby also build a bridge to the older Zipf's literature.

- Systematic empirical research on the age profile of cities within a country is still a largely neglected topic in urban economics and there is little empirical study on the evolution of the number of cities in a country, when it can be considered that the small settlements are to be included in the analysis
- The fitting of double Pareto lognormal for regional city size distribution may be attempted for further understanding the regional structure of the population in a country for planning purposes

- From the point of view of urban economics the study of city size distribution has deep economic implications and studies in this respect may be attempted

ACKNOWLEDGMENTS

The authors are grateful to the authorities of SRM University, Kattankulathur-603 203, India for providing the excellent facilities for carrying out this study. The authors also would like to thank Dr.J. Ranganathan, Former Professor, Department of Statistics, University of Madras, Chennai for helpful discussions during the preparation of this study. The views, findings and interpretations presented in the study are sole responsibility of the authors.

REFERENCES

1. Auerbach, F., 1913. Das gesetz der bevolkerungskonzentration. Petermann's Geographische Mitteilungen, 59: 74-76.
2. Zipf, G.K., 1949. Human Behavior and the Principle of Leeast Effort. Addison-Wesley, Cambridge.
3. Rosen, K.T. and M. Resnick, 1980. The size distribution of cities: An examination of the Pareto law and primacy. J. Urban Econ., 8: 165-186.
4. Black, D. and V. Henderson, 2003. Urban evolution in the USA. J. Econ. Geogr., 3: 343-372.
5. Anderson, G. and Y. Ge, 2005. The size distribution of Chinese cities. Region. Sci. Urban Econ., 35: 756-776.
6. Eeckhout, J., 2004. Gibrat's law for (all) cities. Am. Econ. Rev., 94: 1429-1451.
7. Basu, B. and S. Bandyapadhyay, 2009. Zipf's law and distribution of population in Indian cities. Indian J. Phys., 83: 1575-1582.
8. Ioannides, Y.M. and S. Skouras, 2009. Gibrat's law for (all) cities: A rejoinder. Department of Economics, Discussion Paper, Tufts University. http://papers.ssrn.com /sol3/papers.cfm?abstract_id=1481254
9. Malevergne, Y., V. Pisarenko and D. Sornette, 2011. Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities. Phys. Rev. E, Vol. 83.
10. Giesen, K. and J. Suedekum, 2012. The French overall city size distribution. Region Dev., 36: 107-126.
11. Gonzalez Val, R., A. Ramos, F. Sanz Gracia and M. Vera Cabello, 2015. Size distributions for all cities: Which one is best? Papers Reg. Sci., 94: 177-196.
12. Reed, W.J., 2002. On the rank size distribution for human settlements. J. Regional Sci., 42: 1-17.
13. Reed, W.J. and M. Jorgensen, 2004. The double Pareto lognormal distribution: A new parametric model for size distributions. Commun. Stat. Theory Meth., 33: 1733-1753.
14. Parr, J.B. and K. Suzuki, 1973. Settlement populations and the lognormal distribution. Urban Stud., 10: 335-352.
15. Kass, R.E. and A.E. Raftery, 1995. Bayes factors. J. Am. Stat. Assoc., 90: 773-795.
16. Gabaix, X., 1999. Zipf’s law and the growth of cities. Am. Econ. Rev., 89: 129-132.
17. Gabaix, X., 1999. Zipf's law for cities: An explanation. Quart. J. Econ., 114: 739-767.
18. Ioannides, Y.M. and H.G. Overman, 2003. Zipf's law for cities: An empirical examination. Reg. Sci. Urban Econ., 33: 127-137.
19. Soo, K.T., 2005. Zipf's law for cities: A cross-country investigation. Regional Sci. Urban Econ., 35: 239-263.