Article

Not peer-reviewed version

Generalized Pareto Distribution of Firm Sizes: Evidence from China

Yong Tao * and Ruoxi Liu

Posted Date: 28 April 2024

doi: 10.20944/preprints202404.1840.v1

Keywords: Gibrat's law; generalized Pareto distribution; Zipf distribution; Pareto distribution; exponential distribution

Article

Generalized Pareto Distribution of Firm Sizes: Evidence from China

Yong Tao * and Ruoxi Liu

College of Economics and Management, Southwest University, Chongqing, China
* Correspondence: taoyingyong@163.com

Abstract: It has been empirically observed that the upper tail of the firm size distribution follows either the Pareto distribution or the Zipf distribution, with both patterns being explained by Gibrat's law (Gabaix, 1999). This article analyzed firm revenue data from China, spanning from 2005 to 2013, to examine the whole range of the firm size distribution. Our empirical analysis revealed that firm revenue data over these years is well fitted by a three-parameter generalized Pareto distribution, with the fitted parameters indicating a dichotomy: The size distribution of large-sized firms, namely the upper tail, is asymptotically characterized by a Pareto distribution or a Zipf distribution, whereas the size distribution of smaller and medium-sized firms is approximated by an exponential distribution. This finding suggests that Gibrat's law should be extended to account for the emergence of a generalized Pareto distribution.

Keywords: Gibrat's law; generalized Pareto distribution; Zipf distribution; Pareto distribution; exponential distribution

JEL classification: D39; C13

1. Introduction

It has been empirically observed that the upper tail of the firm size distribution follows either the Pareto distribution or the Zipf distribution (Axtell, 2001; Gabaix, 1999, 2009, 2016). Both laws state that the fraction of firms with a size greater than x is inversely proportional to x itself; that is,

$$P(X > x) = \left(\frac{x}{x_m}\right)^{-a}, \quad (1)$$

where $a \geq 1$ and $x \geq x_m$.

Equation (1) represents the Pareto distribution. In particular, when $a = 1$, it corresponds to the Zipf distribution, which has been proposed to characterize the upper tail of the size distribution of firms and cities (Axtell, 2001; Gabaix, 1999, 2009, 2016; Malevergne et al., 2013). In contrast, the Pareto distribution with $a > 1$ is commonly used to describe the upper tail of the distribution of income and wealth (Nirei and Souma, 2007; Gabaix, 2009; Benhabib et al., 2011; Aoki and Nirei, 2017; Jones and Kim, 2018). To account for the origin of the Pareto distribution (1), Gabaix (1999), applying Gibrat's law, proposed the following random growth model:

$$dX_t = X_t(\mu dt + \sigma dZ_t), \quad (2)$$

where X_t denotes the size of a firm (or city) at the time t, μ and σ are two parameters, and $dZ_t \sim N(0, dt)$ denotes the standard Brownian motion that is independent of the X_t.

The stochastic differential equation (2) is a mathematical representation of Gibrat's law (Gibrat, 1931), which posits that the growth rate a firm's (or city's) size, $r = \left(\frac{dX_t}{dt}\right)/X_t$, is independent of its size X_t. Equation (2) has been widely used to understand the inequality in the upper tail of the distribution (Gabaix, 2009; Benhabib et al., 2011; Aoki and Nirei, 2017; Jones and Kim, 2018). However, the validity of the Pareto distribution (1) in estimating the threshold parameter has been recently questioned (Jenkins, 2017). This challenge may result in a biased estimation of inequality

 doi:10.20944/preprints202404.1840.v1

(Charpentier and Flachaire, 2022). To address this problem, the generalized Pareto distribution (GPD) has been shown to provide more reliable results (Jenkins, 2017; Charpentier and Flachaire, 2022). Therefore, there is a need to consider applying the GPD rather than the Pareto distribution.

Empirical observations have suggested that Gibrat's law (2) does not hold exactly but only asymptotically for large-sized firms (Almus, 2000; Becchetti and Trovato, 2002; Daunfeldt and Elert, 2013). Based on this observation, Tao (2024) proposed an extension of equation (2) by taking smaller-sized firms into account, as follows:

$$dX_t = (1 + \eta X_t)(\mu dt + \sigma dZ_t), \tag{3}$$

where η is a parameter.

Equation (3) asymptotically approaches the functional form of equation (2) as $X_t \to \infty$. Tao (2024) has demonstrated that if firm size x evolves according to equation (3), the resulting size distribution conforms to the GPD. Since equation (3) extends equation (2) by incorporating smaller-sized firms, the GPD is anticipated to encompass the entire range of the firm size distribution. In this paper, we employ firm revenue data from China, ranging from 2005 to 2013, to examine the whole range of the firm size distribution. For this purpose, our dataset includes a large sample from both large-sized firms and small and medium-sized firms.

2. The Model

Tao (2024) has shown that, if the dynamics of firm size x obey the stochastic differential equation (3), the resulting density distribution $f(x,t)$ satisfies a Kolmogorov forward equation as follows:

$$\frac{\partial f(x,t)}{\partial t} = -\frac{\partial[\mu(1+\eta x)f(x,t)]}{\partial x} + \frac{1}{2}\frac{\partial^2[\sigma^2(1+\eta x)^2 f(x,t)]}{\partial x^2}. \tag{4}$$

The solution of equation (4) can be written as (Tao, 2024):

$$\begin{cases} f(x) = \left(\frac{\eta+\frac{1}{\theta}}{1+\eta x_0}\right)\left(\frac{1+\eta x}{1+\eta x_0}\right)^{-\frac{1}{\theta\eta}-2} \\ x \geq x_0 \end{cases}, \tag{5}$$

where $\theta = -\frac{\sigma^2}{2\mu}$ with $\mu < 0$.

Here, we use $F(X \leq x) = \int_{x_0}^{x} f(z)dz$ to denote the cumulative distribution. Thus, by equation (5), one has (Tao, 2024):

$$F(X \geq x) = \left(\frac{1+\eta x}{1+\eta x_0}\right)^{-\frac{1}{\theta\eta}-1}, \tag{6}$$

where $x \geq x_0$.

Equation (6) represents a generalized Pareto distribution (GPD) with three parameters. To see this, one can rewrite it in the standard form of the GPD:

$$F(X \geq x) = \left[1 + A\left(\frac{x-x_0}{B}\right)\right]^{-\frac{1}{A}}, \tag{7}$$

where $A = \frac{\theta\eta}{1+\theta\eta}$ and $B = \frac{\theta(1+\eta x_0)}{1+\theta\eta}$.

It is easy to check that, when $\eta > 0$, the GPD (6) has an asymptotic Pareto tail on the right side of the distribution. In particular, when η approaches 0 sufficiently, the GPD (6) can be decomposed into a two-class pattern as follows (Tao, 2024):

$$F(X \geq x) \approx \begin{cases} exp\left(-\frac{x-x_0}{\theta}\right) & x_0 \leq x < x_m \approx \frac{1}{\eta} \\ \left(\frac{x}{x_m}\right)^{-\frac{1}{\theta\eta}-1} & x > x_m \approx \frac{1}{\eta} \end{cases}, \tag{8}$$

 doi:10.20944/preprints202404.1840.v1

where $x_m = (1 + \eta x_0)/\eta$.

3. Data Analysis

We employ firm revenue data from China, which spans from 2005 to 2013, to fit the GPD (6). However, due to the lack of data for the year 2010, this year has been excluded from our analysis. The raw data were collected from the Chinese Industrial Enterprises Database (CIED), which includes a large sample consisting of both state-owned and non-state-owned firms. In existing literature, firm size is typically measured by either the number of employees (Gabaix, 2009) or revenue (Chen et al., 2023). In this paper, we choose to use the firm's revenue as the index for measuring its size.

To mitigate the potential impact of outliers within the dataset, we have chosen to exclude firm samples with negative or zero revenue, as these may indicate instances of bankruptcy. This implies that we are not considering the birth and death processes of firms. In fact, when these processes are taken into account, the resulting distribution of firm sizes follows a generalized double-Pareto distribution (Tao, 2024). The descriptive statistics of the raw data are presented in Table 1, displaying the number of observations, minimum, maximum, average, and median of each year's data. To fit the GPD (6), we organized the firm revenue data into cumulative percentages at various revenue quantiles.

Table 1. Descriptive statistics of data.

Year	Obs	MIN	MAX	AVG	MEDIAN
2005	269553	10	1.49×10^8	92094	19270
2006	299704	10	1.88×10^8	104827	21735
2007	335018	10	1.96×10^8	119572	25580
2008	368529	10	2.28×10^8	121711	26143
2009	335535	10	2.1×10^8	139400	32344
2011	302591	150	2.6×10^8	271600	77667
2012	323960	10	4.13×10^8	277935	79831
2013	344831	145	4.77×10^8	292601	84302

Figure 1a shows the fit of the GPD (6) to the data in China from 2005 to 2013. One can observe that the agreement between the GPD (6) and the data is very good. In Table 2, we used nonlinear least squares fitting in MATLAB to estimate three parameters in the GPD (6) and R^2 for each year, with all R^2 values exceeding 0.999. According to Table 2, we find that η is approximately in the order of 10^{-5}. This implies that η sufficiently approaches 0; therefore, by equation (8), the firm size distribution can be decomposed into a two-class pattern, where the size distribution of large-sized firms is asymptotically characterized by the Pareto distribution, while the size distribution of smaller and medium-sized firms is approximated by an exponential distribution. Figure 1b displays the two-class pattern.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2024 doi:10.20944/preprints202404.1840.v1

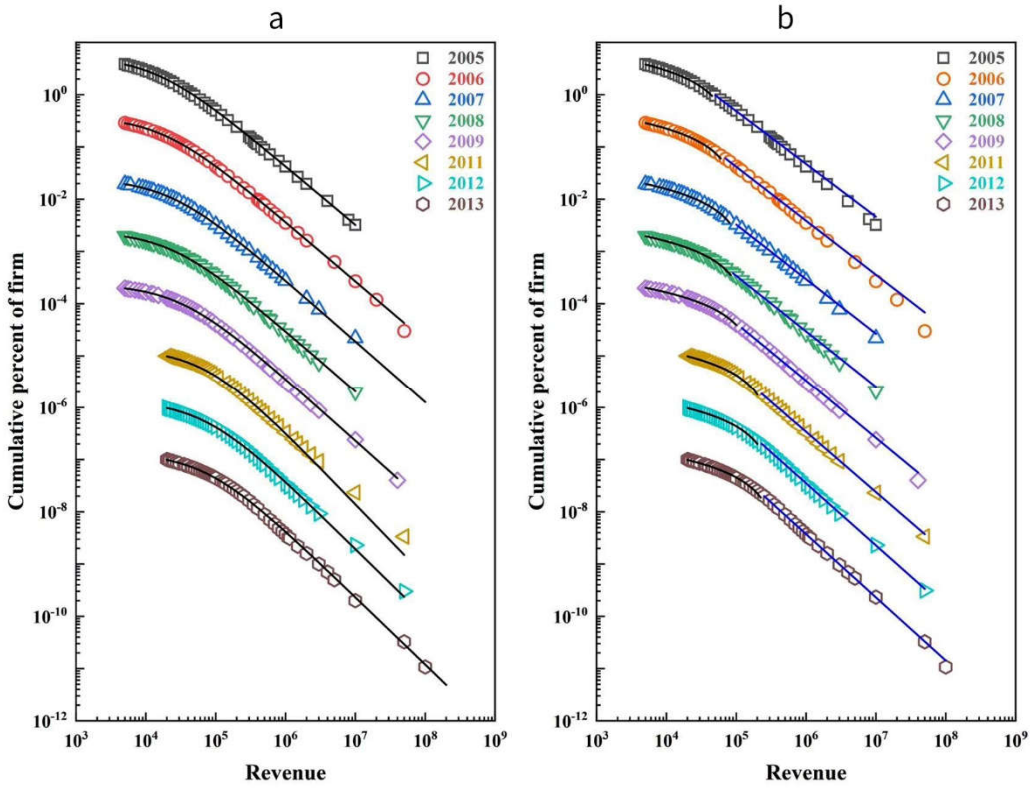

Figure 1. a. Fitting results of generalized Pareto distribution (6). **b.** Fitting results of the two-class distribution (8).

The large-sized firms correspond to the upper tail in equation (8), where the Pareto exponent is denoted by $a = 1 + 1/(\theta\eta)$. We have listed the estimated values of a in Table 2, which are in the vicinity of 1 for all years, roughly agreeing with the Zipf distribution. However, smaller and medium-sized firms follow an exponential distribution rather than a Pareto distribution. This suggests that, unlike large-sized firms, smaller and medium-sized firms face fiercer competition and hence generate significantly less economic profit, aligning with the conditions of competitive market theory, where the exponential distribution is indicative of a spontaneous order in the firm size distribution, as described by Tao (2016).

Table 2. Fitting parameters.

Year	η	x_0	θ	a	R^2
2005	7.34×10^{-5}	4.07×10^{3}	1.10×10^{5}	1.12	0.999
2006	5.89×10^{-5}	4.07×10^{3}	1.22×10^{5}	1.14	0.999
2007	4.41×10^{-5}	4.34×10^{3}	1.14×10^{5}	1.20	0.999
2008	4.47×10^{-5}	4.26×10^{3}	1.40×10^{5}	1.16	0.999
2009	3.27×10^{-5}	4.52×10^{3}	1.56×10^{5}	1.20	0.999
2011	1.43×10^{-5}	1.90×10^{4}	1.79×10^{5}	1.39	0.999
2012	1.47×10^{-5}	1.84×10^{4}	2.17×10^{5}	1.31	0.999

 doi:10.20944/preprints202404.1840.v1

2013	1.58×10^{-5}	1.93×10^{4}	2.94×10^{5}	1.22	0.999

4. Conclusion

Existing research has shown that the upper tail of the firm size distribution follows either the Pareto distribution or the Zipf distribution, with both patterns being explained by Gibrat's Law in the form of equation (2). Using firm revenue data from China (2005-2013), our empirical analysis suggests that the firm size distribution is adequately represented by a three-parameter generalized Pareto distribution, which is explained by an extension of Gibrat's Law presented by equation (3). The fitted parameters reveal a dichotomy: The size distribution of large-sized firms, namely the upper tail, is asymptotically characterized by a Pareto distribution or a Zipf distribution, and the size distribution of smaller and medium-sized firms is approximated by an exponential distribution. Our findings suggest extending the version of Gibrat's law given in equation (2) to the form in equation (3) to accommodate the emergence of a generalized Pareto distribution.

Author Contributions: Yong Tao designed research; Ruoxi Liu organized raw data for the research; Ruoxi Liu analyzed data; Yong Tao wrote the paper.

Data Availability Statement: This study analyzed publicly available datasets from the Chinese Industrial Enterprises Database.

Acknowledgement: This work was supported by the Social Science Planning Project of Chongqing (Grant No. 2019PY40) and the Research project on education and teaching reform in Southwest University (Grant No. 2021JY045)

Author Information: The authors declare no competing interests. Correspondence and requests for materials should be addressed to Yong Tao (taoyingyong@swu.edu.cn) or Ruoxi Liu (liuruoxi64@163.com).

References

Almus, M. (2000): Testing "Gibrat's Law" for Young Firms – Empirical Results for West Germany. *Small Business Economics* **15**, 1-12 (2000)

Aoki, S. and Nirei, M. (2017): Zipf's Law, Pareto's Law, and the Evolution of Top Incomes in the United States. *American Economic Journal: Macroeconomics* **9**, 36-71

Axtell, R. (2001): Zipf Distribution of U.S. Firm Sizes. *Science* **293**, 1818-1820

Becchetti, L. and Trovato, G. (2002): The Determinants of Growth for Small and Medium Sized Firms. The Role of the Availability of External Finance. *Small Business Economics* **19**, 291–306

Benhabib, J., Bisin, A. and Zhu, S. (2011): The Distribution of Wealth and Fiscal Policy in Economies With Finitely Lived Agents. *Econometrica*, **79**, 123-157

Charpentier, A. and Flachaire, E. (2022): Pareto models for top incomes and wealth. *Journal of Economic Inequality* **20**, 1-25

Chen, Y., Hsu, W., and Peng, S. (2023): Innovation, firm size distribution, and gains from trade. *Theoretical Economics* **18**, 341–380

Daunfeldt, S. and Elert, N. (2013): When is Gibrat's law a law? *Small Business Economics* **41**, 133-147

Dragulescu, A. and Yakovenko, V. M. (2001): Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. *Physica A* **299**, 213-221

Gabaix, X. (1999): Zipf's Law for Cities: An Explanation. *Quarterly Journal of Economics* **114**, 739-767

Gabaix, X. (2009): Power Laws in Economics and Finance. *Annual Review of Economics* **1**, 255-294

Gabaix, X. (2016): Power Laws in Economics: An Introduction. *Journal of Economic Perspective* **30**, 185-206

Gibrat, R. (1931): *Les inegalites economiques*. Paris: Librairie du Receuil Sirey.

Jenkins, S. P. (2017): Pareto Models, Top Incomes and Recent Trends in UK. *Economica* **84**, 261-289

Jones, C. I. and Kim, J. (2018): A Schumpeterian Model of Top Income Inequality. *Journal of Political Economy* **126,** 1785-1826

Malevergne, Y., Saichev A., and Sornette, D. (2013): Zipf's law and maximum sustainable growth. *Journal of Economic Dynamics and Control* **37**, 1195-1212

 doi:10.20944/preprints202404.1840.v1

Nirei, M. and Souma, W. (2007): A Two Factor Model of Income Distribution Dynamics. *Review of Income and Wealth* **53**, 440-459

Tao, Y. (2016): Spontaneous economic order. *Journal of Evolutionary Economics* **26**, 467-500

Tao, Y. (2024): Generalized Pareto Distribution and Income Inequality: An extension of Gibrat's law. *AIMS Mathematics*. Forthcoming