View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Repositório Científico do Instituto Politécnico do Porto

Symposium On Fractional Signals and Systems

IPC, Coimbra, Portugal
4 - 5 November 2011

Pareto and Zipf laws for city size distribution

C.M.A. Pinto[1], A.M. Lopes[2], and J.A. Tenreiro Machado[3]

[1]Instituto Superior de Engenharia do Porto
and Centro de Matemática, Universidade do Porto
Rua Dr António Bernardino de Almeida, 431
4200-072 Porto Portugal
cpinto@fc.up.pt, cap@isep.ipp.pt

[2]Faculdade de Engenharia
Universidade do Porto
Rua Dr. Roberto Frias,
4200-465 Porto PORTUGAL
aml@fe.up.pt

[3]Departamento de Engenharia Electrotécnica
Instituto Superior de Engenharia do Porto
Rua Dr António Bernardino de Almeida, 431
4200-072 Porto Portugal
jtm@isep.ipp.pt

Abstract — *Pareto and Zipf distributions have been used in the modeling of distinct phenomena, namely in biology, demography, computer science, economics, amongst others. In this paper, it is presented a short review of applications of these distributions in city sizes.*

1 Introduction

Pareto distribution was introduced in 1896 [17], to describe income distribution. It was shown that the relative number of individuals with an annual income larger than a certain value x was proportional to a power of x. Nowadays, a large amount of studies of income and wealth distribution, and expenditure in distinct economic communities, is available [13, 10, 6]. Zipf [24, 25] applied Pareto laws to words frequencies. The first study concerning city size distribution was due to Auerbach [1], in 1913. He found that the product of the population size of a city by its rank in the distribution appeared to be roughly constant for a given territory. This study brought a new interest to the study of city size distribution [25, 18, 12, 23, 7, 16, 20, 19]. Power laws (PLs) of various forms also appear in computer science [5], in the distribution of biological species [22], war and terrorists attacks [11, 3], books sales [15], music recordings and other commodities [4].

The probability function of a discrete random variable X following a Pareto distribution is given by:

$$P(X = x) = Cx^{-\alpha}$$

where $C > 0$. The corresponding cumulative distribution function is of the form:

$$F(x) = P(X \geq x) = \frac{C}{\alpha - 1} x^{-(\alpha-1)}$$

We let $\tilde{\alpha} = \alpha - 1$. Zipf's law [24, 25] (aka rank-size rule) is a special case of the Pareto law, with $\alpha = 2$. Considering cities ordered by population size, with the one with more population being ranked as 1, then the rank-frequency chart is the plot of $F(x)$ versus the rank r, in logarithmic scales. In log-log scales, the Zipf distribution gives a straight line with slope equal to $\tilde{\alpha} = 1$. More generally, for a random variable following a Pareto distribution the rank-frequency plot, in log-log scales, is asymptotically a straight line, given by:

$$\ln(P[X \geq x]) = \ln C - \ln \tilde{\alpha} - \tilde{\alpha} \ln x$$

2 Pareto and Zipf laws for cities

Cities are complex systems that differ in many distinct ways such as size, shape, and scale. The later have been a major research theme in a variety of scientific areas, from geography up to economy. The one that got more attention was the city size distribution. Pareto distribution, which includes the usual Zipf distribution, as a particular case, has become a major contribution for the distribution of city sizes.

Auerbach [1] (followed by Zipf [25]) proposed that the city size followed a Pareto distribution. In 1980, Rosen and Resnick [18] did a cross-country investigation of city sizes in 44 countries and found that Pareto exponent was in the interval $\tilde{\alpha} \in [0.81, 1.96]$, with sample mean of $\bar{\tilde{\alpha}} = 1.14$ with standard deviation 0.196. These values indicated that population was more evenly distributed than expected by Zipf's law, since the higher the value of the exponent, more even city sizes became. Moreover, Rosen *et al* [18] found that the mean value of α was $\bar{\tilde{\alpha}} = 1.136$. Only in 12 of the 44 countries studied, obtained $\tilde{\alpha} \leq 1$. These authors concluded that cities of wealthier and more populous countries, with a better railway system, grew more evenly, i.e., Pareto coefficient was bigger than for poorest countries.

In 1991, Krugman [12] studied 135 USA metropolitan areas and calculated values of $\tilde{\alpha}$ close to one. Using the same data set, Gabaix [7] derived a statistical explanation of Zipf's law for cities. He showed that if cities followed similar growth processes, i.e., if they evolved randomly with the same growth mean and variance, for a certain range of (normalized) sizes, then, in steady state, the distribution of city sizes would follow a Zipf law. Similar growth processes were often referred as Gibrat's law [8] (aka, Gibrat's law of proportional effects). Thus, Zipf law was considered a steady state distribution arising from Gibrat's law, that is to say, Zipf's law was the limit of a stochastic process.

Zanette *et al* [23] developed an intermittency model to large-scale city size distributions. The model predicted a PL distribution with a coefficient $\tilde{\alpha}$ that was close to 1, the coefficient for large city sizes known in the literature.

Overman and Ioannides [16], implemented a test to check validity of Zipf's law for cities. Their results confirmed the validity of Zipf's law for the US cities, though it was observed a variation in the estimates of the Zipf coefficient $\tilde{\alpha}$ across city sizes. They also proved that Gibrat's law hold generically for city growth processes.

Soo [20] tested the validity of Zipf's law for cities of 73 countries using two estimation methods, ordinary least squares (OLS) and Hill estimator. He found that Zipf's law was rejected in 73% of the countries using OLS estimator, and 41% of the countries using Hill estimator. Soo updated $\tilde{\alpha}$ values for the interval $[0.73, 1.72]$. Soo also obtained a Pareto value for urban agglomerations less than 1, contradicting Rosen and Resnick [18], that stated that Zipfs law hold for urban agglomerations. Soo also tried to explain variations in the Pareto exponent. Unlike Rosen and Resnick [18], he argued that political economy might be the main factor influencing city size distribution.

Moura *et al* [14] studied Brazilian cities with more than 30,000 inhabitants. They showed that Pareto distribution was not valid for smaller cities. For these cases, the city size cumulative distribution function did not follow a power-law behavior. The coefficient $\tilde{\alpha}$ values were calculated with three methods: maximum likelihood estimator, least squares fitting and average parameter estimator, and were in the interval $[1.3, 1.4]$.

Other distributions to model city sizes have been proposed in the literature. In 2001, Bi *et al* [2], proposed the Discrete Gaussian Exponential (DGX) distribution for mining, massive skewed data. The Pareto law was included as a special case of the DGX. These authors applied DGX to distinct data sets, a text from the Bible, clickstream data, sales data and telecommunication data. In all cases, the DGX fitted well the data. Sarabia *et al* [19] introduced the Pareto-positive stable (PPS) distribution as a new model for city size distribution. Pareto and Zipf distributions were included as particular cases. PPS distribution could be obtained by a monotonic transformation of the classical Weibull distribution or by mixing the shape parameter $\tilde{\alpha}$ of the classical Pareto distribution with a positive stable law. Sarabia and co-workers [19] compared their results on data from population of Spanish cities from 1998 up to 2007, with Pareto, lognormal and Tsallis [21] distributions, and PPS provided better fits than the three previous distributions.

Giesen, Zimmermann, and Suedekum [9] introduced the Double Pareto lognormal distribution (DPLN) to model city sizes. They used data from eight countries and their results showed that DPLN was a good fit for cities of all sizes. It was compatible with Zipf's law among large cities.

2.1 A case study

We collected data from the website of Thomas Brinkhoff: City Population, `http://www.citypopulation.de`. This website supplies data about city populations for all countries and, for each country, there is data from several censuses of the last 40 years. In this study, we adopt the concept of administratively defined cities and consider, due to the limited number of pages, only the case of turkish cities.

The data used here corresponds to the last available census of city size population in Turkey. We have built a rank-frequency plot. The procedure followed consisted in collecting, sorting and ranking the data. Before plotting the results, a normalization of the values was carried out. Namely, the data (x-axis) was divided by the value corresponding to the population of the largest city, and the rank (y-axis) was divided by the rank of the

smallest city. Finally, the data was fitted by a PL, using a least squares algorithm (see Fig 1). We obtained a PL with parameters $\tilde{\alpha} = 1.021$, yielding a squared correlation coefficient of $R^2 = 0.9955$. The value of $\tilde{\alpha}$ close to 1 suggested that city size distribution in Turkey followed a Zipf law.

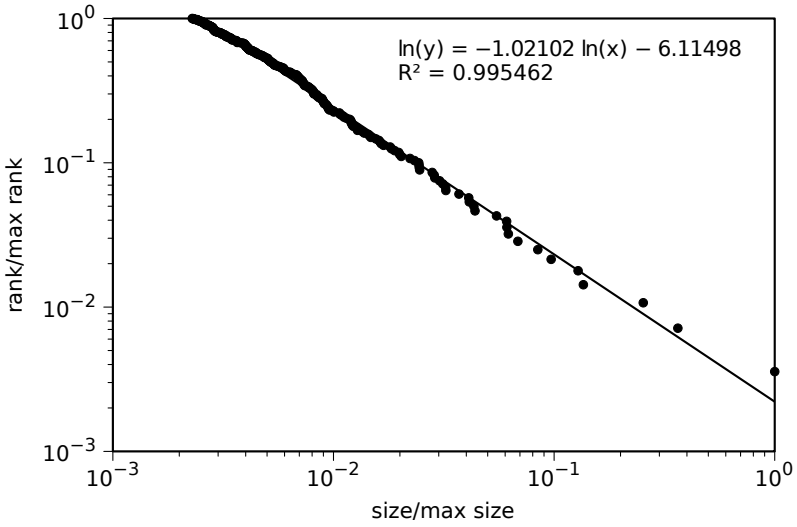

Figure 1: Rank-frequency of the size of populations of the Turkish cities.

3 Conclusions

In this paper we reviewed some applications of Pareto and Zipf laws to city size distribution. We considered city size distribution of turkish cities as a case study. We found a value close to 1 to the $\tilde{\alpha}$ coefficient, suggesting that city size distribution in Turkey followed a Zipf law.

Acknowledgments

CP was supported by the European Regional Development Fund through the programme COMPETE and by the Portuguese Government through the FCT – Fundação para a Ciência e a Tecnologia under the project PEst-C/MAT/UI0144/2011./

References

[1] F. Auerbach. Das Gesetz der Belvolkerungskoncentration. *Petermanns Geographische Mitteilungen*, **59** (1913) 74–76.

[2] Z. Bi, C. Faloutsos, and F. Korn. The "DGX" Distribution for Mining Massive, Skewed Data. In *Proceedings of the Seventh ACMSIGKDD International Conference on Knowledge Discovery and Data mining*, San Francisco, California, August (2001).

[3] J. C. Bohorquez, S. Gourley, A. R. Dixon, M. Spagat, N. F. Johnson. Common Ecology Quantifies Human Insurgency. *Nature* **462** No. 7275 (2009) 911–914.

[4] R. A. K. Cox, J. M. Felton, and K. C. Chung. The concentration of commercial success in popular music: an analysis of the distribution of gold records. *Journal of Cultural Economics* **19** (1995) 333-340.

[5] M. E. Crovella and A. Bestavros, Self-similarity in World Wide Web traffic: Evidence and possible causes. In B. E. Gaither and D. A. Reed (eds.), *Proceedings of the 1996 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, 148-159, Association of Computing Machinery, New York (1996).

[6] F.C. Figueira, N.J. Moura, M.B. Ribeiro. The Gompertz-Pareto income distribution. *Physica A* **390** (2011) 689–698.

[7] X. Gabaix. Zipf's Law and the Growth of Cities. *American Economic Review Papers and Proceedings* **LXXXIX** (1999) 129–132.

[8] R. Gibrat. Les inégalités économiques. Paris, France, Librairie du Recueil Sirey (1931).

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

[10] P. Gopikrishnan, V. Plerou, Y. Liu, L.A.N. Amaral, X. Gabaix, H.E. Stanley. Scaling and correlation in financial time series. *Physica A* **287** (3,4) (2000) 362373.

[11] B. Gutenberg, R. F. Richter. Frequency of earthquakes in california. *Bulletin of the Seismological Society of America* **34** (1944) 185–188.

[12] P. Krugman. *The Self-Organizing Economy*, Cambridge, MA: Blackwell (1996).

[13] B. Mandelbrot. The Pareto-Levy Law and the Distribution of Income. *International Economic Review* **I** (1960) 79–106.

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

[15] M.E.J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* **46** (2005) 323–351.

[16] H.G. Overman, Y. Ioannides. Zipf Law for Cities: an Empirical Examination. *Centre for Economic Performance*, London School of Economics and Political Science, London.

[17] V. Pareto. *Cours d'Economie Politique*, Geneva, Switzerland: Droz (1896).

[18] K. Rosen, M. Resnick. The Size Distribution of Cities: An Examination of the Pareto Law and Primacy. *Journal of Urban Economics* **8** (1980) 165–186.

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

[20] K.T. Soo. Zipfs law for cities: A cross-country investigation. *Regional Science and Urban Economics* **35** (3) (2005) 239–263.

[21] C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics* **52** (1988) 479–487.

[22] J. C. Willis and G. U. Yule. Some statistics of evolution and geographical distribution in plants and animals, and their significance. *Nature* **109** (1922) 177-179.

[23] D.H. Zanette, S.C. Manrubia. Role of intermittency in urban development: A model of large-scale city formation. *Physics Review Letters* (1997) 523–526.

[24] G. Zipf. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press, cambridge, MA (1932).

[25] G. Zipf. *Human Behavior and the Priciple of Least Effort*. Addison-Wesley, Cambridge, MA (1949).