Lodz Economics Working Papers

Firm-Size Distribution in Poland – is Power Law Applicable?

Piotr Gabrielczak,
Tomasz Serwach

3/2019

FACULTY OF ECONOMICS AND SOCIOLOGY
University of LODZ

Firm-size distribution in Poland – is power law applicable?

Piotr Gabrielczak*, Tomasz Serwach**

Abstract

The article focuses on power laws and their growing popularity in science in general and in economics specifically. The theoretical mechanisms responsible for their generating are reviewed. We also empirically test whether firm-size distribution of companies in Poland has the characteristics of the Zipf's law – a special case of a power law. This is confirmed based on an investigation within the sample of 2000 largest companies and a set of alternative estimators of the power law exponent.

Keywords: power law, Zipf's law, firm-size distribution, scaling

JEL: C46, D39, L25

* University of Lodz, Department of Macroeconomics, e-mail: piotr.gabrielczak@uni.lodz.pl
** University of Lodz, Department of International Trade, e-mail: tomasz.serwach@uni.lodz.pl

1. Introduction

Power law (or scaling law) is a relation between two variables X and Y:

(1) $$Y = kX^{\alpha}$$

In this relation α is the so called power law exponent and k is a constant (Gabaix 2008a). In most cases, while the value of k is usually not particularly interesting, the analysis focuses on the values of α, as that parameter has a natural interpretation, e.g. if we multiply X by 2, than the value of Y will be multiplied by 2^{α}. In other words, Y is proportional to X^{α} or X is proportional to $Y^{1/\alpha}$.

Power laws are very popular in recent economic literature, since more and more relations are confirmed to have their features. One of the most popular power laws is the so called Zipf's law, which is a special case empirically derived power law with an exponent $\alpha \cong -1$.

(2) $$Y = kX^{-1} = {}^{k}/_{X}$$

Zipf (1949) gave a few examples of his power law – the most commonly cited and replicated one is associated with city sizes (see Table 1). One can rank cities (e.g. in a certain country) by their population and then compare logged ranks with logged size of population. A linear regression generates a line with a slope of approximately -1. This is equivalent to (2).

$$Y = kX^{-1}$$

$$\log Y = \log k + (-1)\log X$$

$$A \coloneqq \log k$$

(3) $$\log Y = A - \log X$$

This means that the population of a city with rank n is proportional to 1/n, which is the inversed rank[1]. Since ranks are based on the criterion of population, then one can expect that the power law can be transferred to the distribution of the size of population. Indeed, Gabaix (2008b) claims that the probability that the population of a city is greater than X_0 is proportional to $1/X_0$ (or X_0^{-1}). This means that logged rank in the power law can be replaced with a counter-cumulative distribution function without changing the general properties, especially the exponent of the power law.

For convenience, let us now denote size as S. Furthermore, since in case of the Zipf's law the exponent of power law is only approximately -1, let us be more general and denote this exponent as ζ, holding (also for the convenience of interpretation) that ζ is a positive number. Now we can formulate the "distributional" power law as follows:

(4) $$P(S > x) = kx^{-\zeta}$$

The power law (4) becomes a Zipf's power law if its exponent is (close to) 1.

(5) $$\zeta \simeq 1$$

Due to the fact, that we assume that ζ is positive and explicitly add a minus, stressing that the exponent in power law is negative, therefore (4) is sometimes referred to as inverse power

[1] In fact, all of the examples presented by Zipf are based on the inverse relation between absolute and relative measures of frequency or size. Probability, which is e.g. the number of times a word was used on a page, compared to the total sum of words on the page, represents the absolute value of frequency. Rank, which states how many words were more frequent than the considered one, is a relative frequency measure (Kromer 2002).

distribution. However, considering that one of the first researchers to use such a distribution was Vilfredo Pareto and another major contributor to the subject was George Zipf, the inverse power distribution (4) is known as the Pareto distribution, and when condition (5) is included it is often named Pareto-Zipf (power) distribution (Perline 2005).

The aim of the article is to investigate the distribution of size among Polish companies. We shall test the hypothesis that Polish companies are subject to a power law and indicate some of the possible consequences of such a situation.

Section 2 includes a literature review, with a focus not only on economic research, but also on general empirical papers. Section 3 incorporates the presentation of data and methodology. The fourth section concentrates on discussion of the obtained results. Final section concludes.

2. Literature review

2.1 Power law distributions in empirical research

Power laws were widely observed in empirical research starting from Zipf (1949) and throughout the second half of the 20th century and the 21st century. There are a few interesting reviews of empirical research into power laws that are available. One of them, provided by Li (2002), has a revealing title: *Zipf's Law Everywhere*. Table 1 presents some of the key (non-economic) areas in which the power laws have been recently explored.

Table 1. Selected empirical research into power laws

Research problem	References
Frequency of use of words in different languages (both written and spoken)	Zipf 1949; Kucera, Francis 1967; Dahl 1979; Rousseau, Zhang 1992; Egghe 2000; Altmann 2002; Hřebíček 2002; Montemurro, Zanette 2002; Rousseau 2002; Ellis et al. 2015; Mehri, Lashkari 2016
Citation of scientific publications, number of books sold and the frequency of book lending in libraries	Lotka 1926; de Solla Price 1965; Hackett 1967; Fairthorne 1969; Wyllys 1981; White, McCain 1989; Hertzel 1987; Egghe, Rousseau 1990; Egghe 1991; Osareh 1996a; Osareh 1996b; Silagadze 1997; Redner 1998; Clough et al. 2015; Patience et al. 2017
Visits to webpages and Internet traffic	Glassman 1994; Crovella, Bestavros 1997; Barford et al. 1999; Huberman et al. 1998; Barabasi, Albert 1999; Breslau et al. 1999; Adamic, Huberman 2002; Mitzenmacher 2003; Olmedilla et al. 2016; Cao et al. 2017; Bokányi et al. 2019
Population of cities and metropolitan areas	Zipf 1949; Hill 1970; Ijiri, Simon 1977; Rosen, Resnick 1980; Gell-Mann 1994; Gabaix 1999; Knudsen 2001; Brakman et al. 2001; Soo 2005; Gabaix 2008b; Edwards, Batty 2015; Arshad et al. 2018
Natural sciences: geography, biology and physics	Mandelbroot 1982; Schroeder 1991; Bak 1996; Sornette et al. 1996; Serbyn et al. 2016
Applied sciences: medicine and engineering	Sui et al. 2015; Spaide 2016; Biswas et al. 2017a; Biswas et al. 2017b; Wang, Du 2017

Source: own elaboration

Table 1 indicates that power laws are very popular. However, Perline (2005) claims that in some cases researchers use power laws where in fact they should not be applied or are not the best solution for modelling. He proposes dividing power laws into strong power laws, weak power laws and false power laws. The basic problem is what he calls truncation of data. Truncation is a process of selecting a sample only out of the observations with the highest ranks (when referring to rank-size plot based on Zipf's original approach). In some cases, truncation is justified by customs associated with definitions or by data availability. Perline (2005) uses the example of research on the distribution of the size of lakes (e.g. Mandelbroot 1982). Calculations based on a sample of lakes lead to a conclusion that there is a power law in the distribution of the size of lakes. However, Perline stresses that the division between lakes and e.g. ponds is somewhat arbitrary. Perhaps then the research only includes the right tale, while the left tale of distribution is excluded, because we do not consider ponds as small lakes. Similarly, in economics it is much easier to find data and medium and large companies, but not about the micro enterprises. This distinguishes strong power laws, which are fully sure, from weak power laws, in which cases we can only investigate the right tale of the distribution. Disturbances in the left tale, which we cannot detect, could actually prove, that the problem described with a weak power law could be just as well (or even better) modelled with a log-normal distribution or Yule distribution[2].

When Perline (2005) talks about false power laws, he refers to situation in which data is either selected or modified in a way that increases its resemblance to power law, which is not true for raw data. He presents an example of two studies about American and Canadian steel plants (Simon, Bonini 1958; Kendall 1961), which prove that there is a power law in the distribution of their capacity. However, both papers focus only on top 10 steel plants, while Perline (2005) shows that including the other available data changes the conclusions significantly, so this is the case of a false power law. Surprisingly, some of the original examples from Zipf (1949) include either procedures that are now known to be problematic and bias-creating (e.g. not normalizing the intervals of size ranges) or simply seem to be artificial and unjustified (e.g. adding a fixed constant to all the data, which makes the logged rank-size plot more linear, suggesting a power law).

While false power laws are cases of research errors or manipulation, weak power laws are difficult to eliminate. That is why more and more researchers (e.g. Newman 2005; Gabaix 2008b; di Giovanni et al. 2011) explicitly state that their research is focused on the right tale and results may not apply to the left tale of the distribution. Therefore, for scrutiny and safety reasons, they assume that their results are the cases of weak rather than strong power laws.

2.2 Power law distributions in economics

In economics power laws were originally used to describe the distribution of income and wealth (Pareto 1896). This usage for power laws is still very popular, as many papers demonstrate existence of power laws in either of the cases (Atkinson et al. 2011; Benhabib et al. 2011) or, which becomes more and more popular, in both cases jointly (Pickety, Zucman 2014; Gabaix et al. 2016). Considering these issues together is justified, as power law distribution indicates inequalities – and inequalities in income distribution tend to cumulate into even larger inequalities in terms of wealth (Gabaix 2016). Investigation into the distribution of income lead to widening the use of power laws onto areas directly associated with income, such as productivity (Lucas, Moll 2014) or consumption (Toda, Walsh

[2] A variable S has a log-normal distribution when ln(S) has a normal distribution. Log-normal distribution is very close to power law in its right tale. Yule (1925) suggested a distribution similar to a power law. In Yule's case, the counter-cumulative distribution function can be denoted as $P(S > x) = Ax^{-\zeta}B^{-x}$. However, in most cases B is close to 1, so Yule distribution can be easily mistaken for a power law (Simon 1955).

2015). Some focus has been also brought to capital markets, on which one can observe power law distributions of returns, daily numbers of transactions or other parameters associated with stocks (Gopikrishnan et al. 1999; Plerou et al. 2005; Bouchaud et al. 2009; Kyle, Obizhaeva 2019).

Modern economics puts a lot of emphasis on analyses within the framework of imperfect competition (monopolistic competition or oligopoly). In such a case the very distribution of firm size seems to be informative. Many papers focus on testing the power laws for the distribution of firm size (Axtel 2001; Gabaix 2011; di Giovanni et al. 2011). Demonstrating that companies have a power distribution of their size (usually measured with sales or employment) would work as a proof that the market structure is not perfect.

Gabaix (2016) demonstrates a rationale behind the occurrence of multiple power laws among economic phenomena[3]. Power laws appear as a result of the mechanisms arising from the proportional random growth theory. Let us assume that we observe a set of companies, which grow or shrink randomly due to independent shocks, but at the same time they satisfy the Gibrat's law, which states that all the companies have the same expected growth rate (with the same standard deviation). Such a model allows only to draw a conclusion, that in time the distribution of firms should tend towards log-normal with a variance growing over time. There is no guarantee that the observed set of companies (an economy) would obtain a steady state distribution. To assure that, one more condition is needed, that is the assumption of a lower bound of the firm's size. Now this model technically produces a steady state distribution, which is a power law. However, it does not necessarily have to be a Zipf's law, which means that the exponent does not have to be close to 1. However, Gabaix (2008b) demonstrates, that the exponent aims for 1 if we include two very realistic assumptions, namely that the lower bound for size is very low and that the number of companies in the economy is finite.

In other words, Zipf's law is a steady state distribution for companies if we assume the following conditions:

I. The economy has a finite size in terms of number of enterprises
II. There is a lower bound for the size of a company and it is relatively low
III. Companies demonstrate constant economies of scale

The above conditions are in fact very realistic. The first condition is quite obvious from a practical point of view. It would be rather abstract to expect that an economy can produce an infinity of companies. The second condition is also realistic. If we measure the size of a company with its employment, than the lower bound of 1 is the only logical. If we measure the size of a company with its sales, than still we can claim, that because of existing fixed costs, maintaining a company is rational only if it obtains a minimum turnover of a certain level, characteristic for each economy.

The third condition is a bit more problematic. Some economists assume that constant returns to scale are a typical situation for an average company and in such models this condition is obvious. However, Gabaix (2016) notices, that there are many microeconomic models which assume occurrence of the economies of scale, therefore a bigger company should be more cost-efficient and able to grow even faster than a smaller company. In such a case, the Gibrat's law should not stand, but Gabaix (2016) claims that empirically it usually does, which is due to non-economic counter-factors that balance positive economies of scale in real life. These could be e.g. institutional factors,

[3] Gabaix (2016) provides a rather non-technical narration based on economic theories and rational explanations. Gabaix (1999; 2008b) supplies a more formalized explanations of the mathematical side of the conduct.

such as more strict tax regulations and less support for bigger companies. Another possibility is simply that economies (or diseconomies) of scale might exist, but are not very strong. That would be enough for many processes to be estimated as power laws with exponents close to 1.

Surprisingly, power law distribution of company sizes may affect power law distribution of incomes, as suggested by Rosen (1981) and developed by Gabaix and Landier (2008). One of the mechanisms connecting those areas may be through competition over most skilled workers – both Rosen (1981) and Gabaix and Landier (2008) focused on top managers in their analyses, but in fact their conclusions could be drawn even to top artists or sportsmen. They proved that if the companies have a power law distribution and there is no upper bound for the size (or resourcefulness) of the company, than even small differences in skilfulness or talent of a scarce group of workers will translate to their earnings having a power law distribution without an upper bound. Gabaix and Landier (2008) called that mechanism of generating income distribution a double power law.

One important remark needs to be made in the context of strong, weak and false power laws in economics. It seems that in many cases economic power laws are weak at best. In fact, di Giovanni and Levchenko (2013) claim that Zipf's law of firm sizes can only be estimated for the largest companies up to a certain cut-off. Utilizing a sample including smaller companies, from below that threshold, quickly biases the exponent of the power law towards 0. While one may question if estimating power law as a rule for size distribution makes sense in such a case, it is worth stressing, that estimating Zipf's law for top largest companies may serve a different purpose. Existence of ZIpf's law for a reasonable sample of largest companies indicates a fat right tale of the distribution, which means that there is a significant amount of relatively large companies which can generate idiosyncratic shocks that could not be cancelled out and would eventually affect the situation of an entire economy (Gabaix 2011).

3. Data and methodology

In our study we applied data provided by "Rzeczpospolita". That popular and at the same time renowned newspaper presents lists of biggest firms operating in Poland – these lists are labelled "Lista 500" and "Lista 2000", indicating the number of firms analysed. Instead of being a mere rankings of firms, "Lista 500" and "Lista 2000" present various firm-level data – for example, total sales, employment, total assets, equity etc.

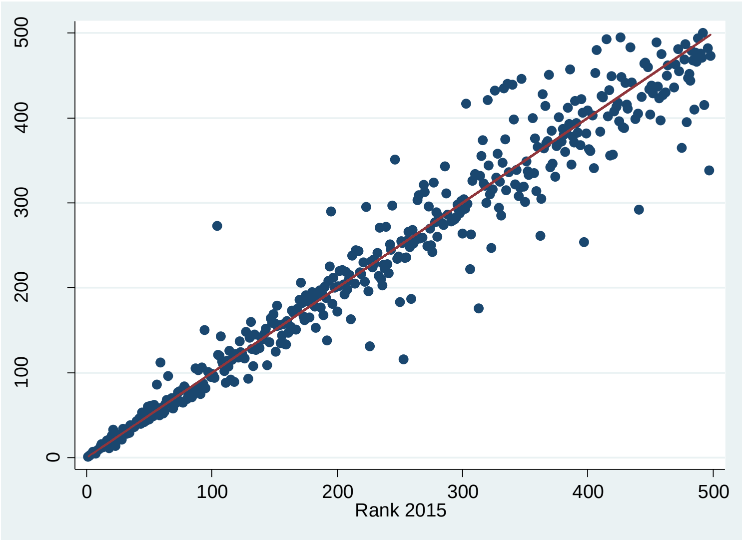
We used the latest "Lista 2000" ranking available online – the one for 2016 with data for fiscal year 2015. We preferred "Lista 2000" to "Lista 500", since we wanted to investigate how our estimates depend on the number of firms analysed. As described above, di Giovanni and Levchenko (2013) stated that power law may be observed only for firms with size bigger than a certain threshold. The use of "Lista 2000" allowed us to check what happens when we restrict our sample to 1500, 1000 and 500 firms, instead of relying on the whole ranking.

The indicator of firm size used in our study is total sales. Other possible indicators, like employment or total assets, have some missing values. "Rzeczpospolita" builds its ranking on the basis of sales, hence that particular indicator was complete for the whole dataset. According to 2015 data, the biggest firm in Poland was an oil refiner and petrol retailer with sales of around 88 billion PLN (roughly 4.9% of Polish GDP). The second biggest firm was foreign-owned firm that operates in food distribution and specialized retail with sales of 39 billion PLN (2.2% GDP). The third one was an oil and gas company with sales of 36 billion PLN (2% GDP). The last company on the list (ranked at 2000th position) was a producer of herbs, aromatic teas, full-flavoured syrups etc. Its sales were around 124 million PLN (0.01% GDP). The average sales calculated for the whole set of firms was slightly more than 876 million PLN (0.05% GDP), while the median was almost 303 million PLN (0.02%

GDP). Huge difference between the average and the median indicates significant skewness of the firm-size distribution in Poland.

The rank of firms is remarkably stable. Figure 1 illustrates the relationship between rankings of firms for 2015 and 2016. We used only "Lista 500" and skipped these firms that were not placed on the 2015 list. Eventually, we were left with 421 companies. The correlation coefficient calculated for these variables is 0.97. If one runs the regression of 2016 ranking on the one from 2015, almost one-to-one relationship may be found, as indicated in Table 1.

Figure 1. The relationship between 2016 and 2016 rankings of firms



Source: authors' own calculations.

Table 1. The relationship between 2016 and 2016 rankings of firms – OLS estimation results

Source	SS	df	MS
Model	8025002.04	1	8025002.04
Residual	495205.936	419	1181.87574
Total	8520207.97	420	20286.2095

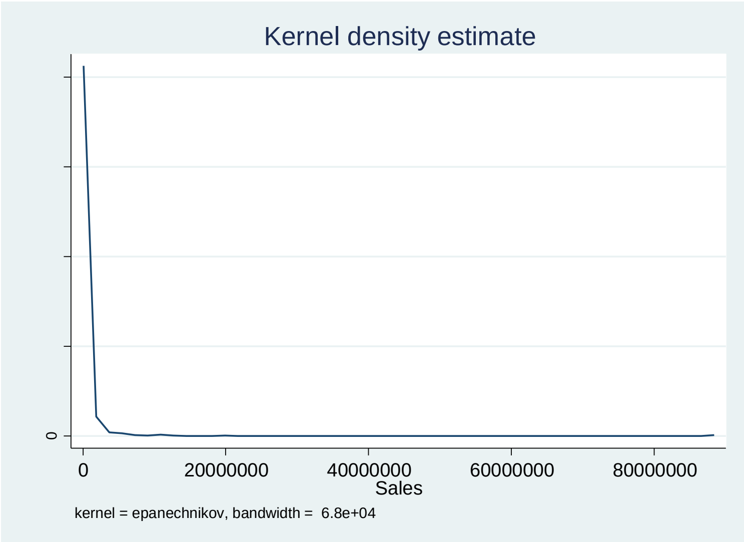
Number of obs	= 421
F(1, 419)	= 6790.06
Prob > F	= 0.0000
R-squared	= 0.9419
Adj R-squared	= 0.9417
Root MSE	= 34.378

Poz2016	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Poz2015	.9624176	.0116796	82.40	0.000	.9394598	.9853755
_cons	6.955255	3.270014	2.13	0.034	.5275788	13.38293

Source: authors' own calculations.

Figure 2 presents the kernel density estimate of sales. One may argue that it illustrates Pareto distribution of firm sales in Poland.

Figure 2. Kernel density estimates



Source: authors' own calculations.

We utilised several methods of estimating the exponent of the power law. In the first one, modelled after Zipf's original approach, we regressed log rank on log sales, as presented below:

(6) $$\ln(Rank_i) = \ln(k) - \zeta \ln(S_i)$$

The second one was based on the definition of the power law stated in Eq. (4).

(7) $$\ln\big(P(S_i > x)\big) = \ln(k) - \zeta \ln(s)$$

The left-hand side is the number of firms in the sample with sales higher than s divided by the total number of firms. Then we regress the natural log of this probability on log sales.

Finally we used the Gabaix-Ibragimov (2011) estimator, in which case one is to regress the natural log of $Rank_i - 1/2$ of each firm in the sales distribution on its logged sales.

(8) $$\ln(Rank_i - 1/2) = \ln(k) - \zeta \ln(S_i)$$

As presented in the next section, these methods lead to similar results.

4. Results

We present the results below in the form of groups of tables – what we call "table" are three panels with results obtained with the use of three methods described in the previous section. The following results (Table 2) are based on the whole set (2000) of firms.

Table 2. Estimation results – 2000 firms

a) log rank regressed on log sales

Source	SS	df	MS		
Model	1934.01974	1	1934.01974	Number of obs =	2000
Residual	30.5452264	1998	.015287901	F(1, 1998) =	.
				Prob > F =	0.0000
				R-squared =	0.9845
				Adj R-squared =	0.9844
Total	1964.56497	1999	.982773871	Root MSE =	.12364

lnRank	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnS	-.9558177	.0026873	-355.68	0.000	-.9610879	-.9505474
_cons	18.93317	.034776	544.43	0.000	18.86497	19.00137

b) log probability regressed on log sales

Source	SS	df	MS		
Model	1927.49596	1	1927.49596	Number of obs =	1999
Residual	36.0732466	1997	.018063719	F(1, 1997) =	.
				Prob > F =	0.0000
				R-squared =	0.9816
				Adj R-squared =	0.9816
Total	1963.5692	1998	.982767369	Root MSE =	.1344

lnP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnS	-.9608402	.0029414	-326.66	0.000	-.9666088	-.9550716
_cons	11.39396	.0380549	299.41	0.000	11.31933	11.46859

c) Gabaix-Ibragimov

Source	SS	df	MS		
Model	1957.7169	1	1957.7169	Number of obs =	2000
Residual	36.193457	1998	.018114843	F(1, 1998) =	.
				Prob > F =	0.0000
				R-squared =	0.9818
				Adj R-squared =	0.9818
Total	1993.91036	1999	.997453905	Root MSE =	.13459

GabaixIbra~v	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnS	-.9616556	.0029252	-328.74	0.000	-.9673924	-.9559187
_cons	19.00629	.037855	502.08	0.000	18.93205	19.08053

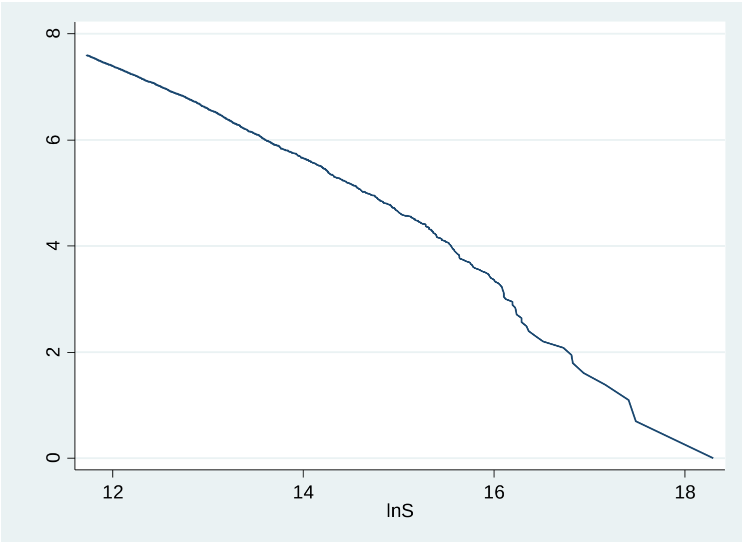
Source: authors' own calculations.

As one may observe, the results obtained with the use of different methods are quantitatively similar. In each case the estimated absolute value of the power law exponent is close to 1. It indicates the existence of Zipf's law regarding firm-size distribution in Poland. These results are also quite similar to these obtained by di Giovanni and Levchenko (2013). They utilized ORBIS database and estimated power law exponent for 50 countries. For each country they used the year with the largest number of firms – in practice, it was a year from the 2006 to 2008 range. Their result for Poland was slightly below 1 in absolute terms.

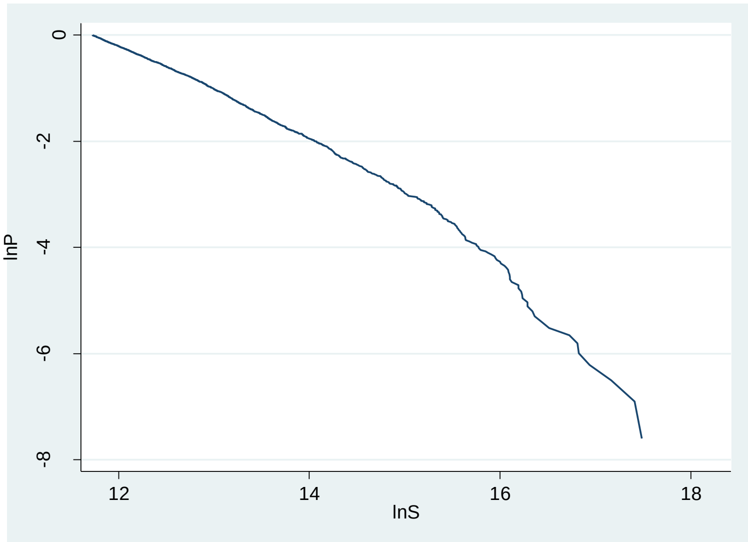
Figure 3 plots a given dependent variable on log sales. As our quantitative results indicate, the relationship between variables in question resembles a straight line with the coefficient of around -1.

Figure 3. The relationships illustrating the power law – 2000 firms

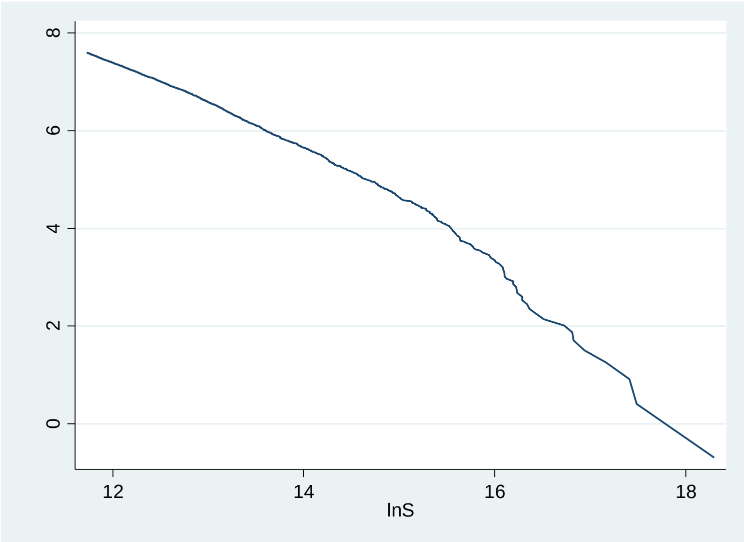
a) log rank and log sales



b) log probability and log sales



c) Gabaix-Ibragimov estimator



Source: authors' own calculations.

Table 3 summarizes the results for 1500 firms. The estimated coefficients are even closer to -1. The difference between these results and those for 2000 firms are rather small.

Table 3. Estimation results – 1500 firms

a) log rank regressed on log sales

Source	SS	df	MS		
				Number of obs	= 1500
				F(1, 1498)	=96267.99
Model	1444.47146	1	1444.47146	Prob > F	= 0.0000
Residual	22.4770284	1498	.015004692	R-squared	= 0.9847
				Adj R-squared	= 0.9847
Total	1466.94849	1499	.978618072	Root MSE	= .12249

lnRank	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnS	-.9974023	.0032146	-310.27	0.000	-1.003708	-.9910967
_cons	19.51315	.0426509	457.51	0.000	19.42949	19.59682

b) log probability regressed on log sales

Source	SS	df	MS		
				Number of obs	= 1499
				F(1, 1497)	=78248.29
Model	1438.43465	1	1438.43465	Prob > F	= 0.0000
Residual	27.5192792	1497	.018382952	R-squared	= 0.9812
				Adj R-squared	= 0.9812
Total	1465.95393	1498	.978607428	Root MSE	= .13558

lnP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnS	-1.004234	.00359	-279.73	0.000	-1.011276	-.9971923
_cons	12.28627	.0476174	258.02	0.000	12.19287	12.37968

c) Gabaix-Ibragimov

Source	SS	df	MS
Model	1466.69601	1	1466.69601
Residual	27.4137153	1498	.01830021
Total	1494.10973	1499	.996737644

Number of obs	= 1500
F(1, 1498)	=80146.40
Prob > F	= 0.0000
R-squared	= 0.9817
Adj R-squared	= 0.9816
Root MSE	= .13528

GabaixIbra~v	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnS	-1.005046	.0035501	-283.10	0.000	-1.01201	-.9980822
_cons	19.61147	.0471023	416.36	0.000	19.51908	19.70386

Source: authors' own calculations.

The next results are based on the estimation of power law coefficient with the use of data for 1000 firms only. Again, the estimation proves the existence of Zipf's law.

Table 4. Estimation results – 1000 firms

a) log rank regressed on log sales

Source	SS	df	MS
Model	954.911755	1	954.911755
Residual	15.2544119	998	.015284982
Total	970.166167	999	.971137305

Number of obs	= 1000
F(1, 998)	=62473.86
Prob > F	= 0.0000
R-squared	= 0.9843
Adj R-squared	= 0.9843
Root MSE	= .12363

lnRank	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnS	-1.055207	.0042217	-249.95	0.000	-1.063492	-1.046923
_cons	20.33908	.0578521	351.57	0.000	20.22555	20.4526

b) log probability regressed on log sales

Source	SS	df	MS
Model	949.494833	1	949.494833
Residual	19.6790753	997	.01973829
Total	969.173908	998	.97111614

Number of obs	= 999
F(1, 997)	=48104.21
Prob > F	= 0.0000
R-squared	= 0.9797
Adj R-squared	= 0.9797
Root MSE	= .14049

lnP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnS	-1.065594	.0048585	-219.33	0.000	-1.075128	-1.05606
_cons	13.5674	.0665519	203.86	0.000	13.4368	13.698

c) Gabaix-Ibragimov

Source	SS	df	MS
Model	975.063068	1	975.063068
Residual	19.3276387	998	.019366371
Total	994.390707	999	.995386093

Number of obs	= 1000
F(1, 998)	=50348.26
Prob > F	= 0.0000
R-squared	= 0.9806
Adj R-squared	= 0.9805
Root MSE	= .13916

GabaixIbra~v	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnS	-1.066283	.004752	-224.38	0.000	-1.075608	-1.056958
_cons	20.48648	.0651195	314.60	0.000	20.35869	20.61427

Source: authors' own calculations.

The last results are generated with the use of data for 500 firms.

Table 5. Estimation results – 500 firms

a) log rank regressed on log sales

Source	SS	df	MS
Model	464.903417	1	464.903417
Residual	10.3782586	498	.020839877
Total	475.281675	499	.952468287

Number of obs	= 500
F(1, 498)	=22308.36
Prob > F	= 0.0000
R-squared	= 0.9782
Adj R-squared	= 0.9781
Root MSE	= .14436

lnRank	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnS	-1.143378	.0076552	-149.36	0.000	-1.158418	-1.128338
_cons	21.64669	.1101522	196.52	0.000	21.43027	21.86311

b) log probability regressed on log sales

Source	SS	df	MS
Model	460.25481	1	460.25481
Residual	14.0409341	497	.028251377
Total	474.295744	498	.952401093

Number of obs	= 499
F(1, 497)	=16291.41
Prob > F	= 0.0000
R-squared	= 0.9704
Adj R-squared	= 0.9703
Root MSE	= .16808

lnP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnS	-1.163269	.0091138	-127.64	0.000	-1.181176	-1.145363
_cons	15.70665	.1310597	119.84	0.000	15.44915	15.96415

c) Gabaix-Ibragimov

Source	SS	df	MS
Model	481.458129	1	481.458129
Residual	13.4128271	498	.026933388
Total	494.870957	499	.991725364

Number of obs	=		500
F(1,	498)	=	17875.88
Prob > F	=		0.0000
R-squared	=		0.9729
Adj R-squared	=		0.9728
Root MSE	=		.16411

GabaixIbra~v	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnS	-1.163557	.0087027	-133.70	0.000	-1.180656	-1.146459
_cons	21.92919	.125225	175.12	0.000	21.68315	22.17522

Source: authors' own calculations.

The general finding is that power law exponent ζ is close to one, no matter the method applied or the sample size. At the same time, one may observe that when sample size is smaller, ζ becomes higher. Reduction in sample size by 500 firms means typically that the exponent's value grows by around 0.06. Therefore, the difference between the results obtained for the biggest and the smallest samples (2000 and 500 firms, respectively) is around 0.2.

5. Conclusions

The latest commonly available data on the sales of top 2000 companies in Poland proves that the firm-size distribution in Poland is well approximated by the Zipf's law, a power law with exponent close to 1. Taking into account that we have used three estimation strategies and four restrictions on the sample size, none of which substantially affected the obtained results, our findings should be considered robust.

One important observation is that in the case of our dataset, regardless of the restrictions limiting our sample to 2000, 1500, 1000 or 500 companies, or of the utilised estimation method, we obtain exponents close to 1. However, for more limited samples, excluding the smallest companies, the estimated power law exponent has a greater absolute value, while widening the sample pulls that result towards zero. This is consistent with di Giovanni and Levchenko (2013) and may suggest that while power law might as well in general be a good fit for the firm-size distribution as such (we are unable to test that due to lack of data about smaller companies), the Zipf's law could be a weak case, appropriate only for the right tale of the firm-size distribution.

Reference list

Adamic L. A., Huberman B. A. (2002), *Zipf's law and the internet*, "Glottometrics", No. 3.

Altmann G. (2002), *Zipfian linguistics*, "Glottometrics", No. 3.

Arshad S., Hu S., Ashraf B. N. (2018), *Zipf's law and city size distribution: A survey of the literature and future research agenda*, "Physica A: Statistical Mechanics and its Applications", Vol 492, DOI 10.1016/j.physa.2017.10.005.

Atkinson A. B., Piketty T., Saez E. (2011), *Top Incomes in the Long Run of History*, "Journal of Economic Literature", Vol. 49, No. 1, DOI 10.1257/jel.49.1.3.

Axtell R. L. (2001), *Zipf Distribution of US Firm Sizes*, "Science", Vol. 293, No. 5536, DOI 10.1126/science.1062081.

Bak P. (1996), *How Nature Works*, Copernicus, New York.

Barabasi A. L., Albert R. (1999), *Emergence of scaling in random networks*, "Science", Vol. 286, No. 5439, DOI 10.1126/science.286.5439.509.

Barford P., Bestavros A., Bradley A., Crovella M. (1999), *Changes in web client access patterns: characteristics and caching implications*, "World Wide Web", Vol. 2, No. 1-2.

Benhabib J., Bisin A., Zhu S. (2011), *The Distribution of Wealth and Fiscal Policy in Economies with Finitely Lived Agents*, "Econometrica" Vol. 79, No. 1, DOI 10.3982/ECTA8416.

Biswas A., Kara A. H., Ullah M. Z., Zhou Q., Moshokoa S. P., Belic M. (2017b), *Conservation laws for cubic–quartic optical solitons in Kerr and power law media*, "Optik", Vol. 145, DOI 10.1016/j.ijleo.2017.08.047.

Biswas A., Triki H., Zhou Q., Moshokoa S. P., Ullah M. Z., Belic M. (2017a), *Cubic–quartic optical solitons in Kerr and power law media*, "Optik", Vol. 144, DOI 10.1016/j.ijleo.2017.07.008.

Bokányi E, Kondor D., Vattay G. (2019), *Scaling in Words on Twitter*, mimeo, https://arxiv.org/pdf/1903.04329.pdf (28.08.2019).

Bouchaud J.-P., Farmer J. D., Lillo F. (2009), *How Markets Slowly Digest Changes in Supply and Demand*, in: Schenk-Hoppe K. R., Hens T. (eds.), *Handbook of Financial Markets: Dynamics and Evolution*, North-Holland, Amsterdam.

Brakman, S., Garretsen, H., van Marrewijk, C. (2001), *An Introduction to Geographical Economics*, Cambridge University Press, Cambridge.

Breslau L., Cao P., Fan L., Philips G., Shenker S. (1999), *Web caching and Zipf-like distributions: evidence and implications*, Proceedings of IEEE Infocom'99, DOI 10.1109/INFCOM.1999.749260.

Cao Z., He Z., Johnson N. (2017), *Asymmetric access to information impacts the power-law exponent in networks*, mimeo, https://arxiv.org/pdf/1712.06169.pdf (28.08.2019).

Clough J. R., Gollings J., Loach T. V., Evans T. S. (2015), *Transitive reduction of citation networks*, "Journal of Complex Networks", Vol. 3, No. 2, DOI 10.1093/comnet/cnu039.

Crovella M. E., Bestavros A. (1997), *Self-similarity in world wide web traffic: evidence and possible causes*, "IEEE/ACM Transactions on Networking", Vol. 5, No. 6, DOI 10.1109/90.650143.

Dahl G. (1979), *Word Frequencies of Spoken American English*, CT Verbatim, Essex.

de Solla Price D. J. (1965), *Networks of Scientific Papers*, "Science", Vol. 149, No. 3683, DOI 10.1126/science.149.3683.510.

di Giovanni J., Levchenko A. A. (2013), *Firm entry, trade, and welfare in Zipf's world*, "Journal of International Economics", Vol. 89, No. 2, DOI 10.1016/j.jinteco.2012.08.002.

di Giovanni J., Levchenko A. A., Rancière R. (2011), *Power laws in firm size and openness to trade: Measurement and implications*, "Journal of International Economics", Vol. 85, No. 1, DOI 10.1016/j.jinteco.2011.05.003.

Edwards R., Batty M. (2015), *City size: Spatial dynamics as temporal flows*, "Environment and Planning A", Vol. 48, No.6, DOI 10.1177/0308518X15621634.

Egghe L. (1991), *The exact place of Zipf's and Pareto's law amongst the classical informetric laws*, "Scientometrics", Vol. 20, No. 1.

Egghe L. (2000), *The distribution of N-grams*, "Scientometrics" 47, No. 2.

Egghe L., Rousseau R. (1990), *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science*, Elsevier, Amsterdam-New York.

Ellis N. C., O'Donnell M. B., Römer U. (2015), *Usage-Based Language Learning*, in: MacWhinney B., O'Grady W. (eds.), *The Handbook of Language Emergence*, John Wiley & Sons, Chichester.

Fairthorne R. A. (1969), *Empirical hyperbolic distributions (Bradford Zipf Mandelbrot) for bibliometric description and prediction*, "Journal of Documentation", Vol. 25, No. 4, DOI 10.1108/eb026481.

Gabaix X. (1999), *Zipf's law for cities: an explanation*, "Quarterly Journal of Economics", Vol. 114, No.3.

Gabaix X. (2008a), *Power Laws*, in: Durlauf S. N., Blume L. E. (eds), *The New Palgrave Dictionary of Economics*, Palgrave Macmillan, London.

Gabaix X. (2008b), *Power Laws in Economics and Finance*, NBER Working Paper No. 14299, Cambridge (MA).

Gabaix X. (2011), *The Granular Origins of Aggregate Fluctuations*, "Econometrica", Vol. 79, No. 3, DOI 10.3982/ECTA8769.

Gabaix X. (2016), *Power Laws in Economics: An Introduction*, "Journal of Economic Perspectives", Vol. 30, No. 1, DOI 10.1257/jep.30.1.185.

Gabaix X., Ibragimov R. (2011), Rank-1/2: a simple way to improve the OLS estimation of tail exponents, "Journal of Business and Economic Statistics", Vol. 29, No. 1, 24–39, DOI: 10.1198/jbes.2009.06157

Gabaix X., Landier A. (2008), *Why Has CEO Pay Increased So Much?*, "Quarterly Journal of Economics", Vol. 123, No. 1, 10.1162/qjec.2008.123.1.49.

Gabaix X., Lasry J.-M., Lions P.-L., Moll B. (2016), *The Dynamics of Inequality*, "Econometrica", Vol. 84, No. 6, DOI 10.3982/ECTA13569.

Gell-Mann M. (1994), *The Quark and the Jaguar*, Freeman, New York.

Glassman S. (1994), *A caching relay for the world wide web*, “Computer Networks and ISDN Systems”, Vol. 27, No.2, DOI 10.1016/0169-7552(94)90130-9.

Gopikrishnan P., Plerou V., Nunes Amaral L. A., Meyer M., Stanley H. E. (1999), *Scaling of the Distribution of Fluctuations of Financial Market Indices*, “Physical Review E”, Vol. 60, No. 5, DOI 10.1103/physreve.60.5305.

Hackett A. P. (1967), *70 Years of Best Sellers, 1895-1965*, R.R. Bowker Company, New York.

Hertzel D. H. (1987), *Bibliometrics, history of the development of ideas*, In: Kent A. (ed.), *Encyclopedia of Library and Information Science*, Vol. 42, Suppl. 7, Marcel Dekker Inc., Basel-New York.

Hill B. M. (1970), *Zipf’s law and prior distributions for the composition of a population*, “Journal of the American Statistical Association”, Vol. 65, No. 331, DOI 10.2307/2284288.

Hřebíček L. (2002), *Zipf’s law and text*, “Glottometrics”, No. 3.

Huberman B. H., Pirollo P. L. T., Pitkow J. E., Lukose R. M. (1998), *Strong regularities in world wide web surfing*, “Science”, Vol. 280, No. 5360, DOI 10.1126/science.280.5360.95.

Ijiri Y., Simon H. A. (1977), *Skew Distributions and the Sizes of Firms*, North-Holland, Amsterdam.

Knudsen T. (2001), *Zipf’s Law for Cities and Beyond: The Case of Denmark*, “American Journal of Economics and Sociology”, Vol. 60, No. 1.

Kromer V. (2002), *Zipf´s law and its modification possibilities*, “Glottometrics”, No. 5.

Kucera H., Francis W. N. (1967), *Computational Analysis of Present-Day American English*, Brown University Press, Providence.

Kyle A. S., Obizhaeva A. A. (2019), *Large Bets and Stock Market Crashes*, mimeo, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2023776 (1.09.2019), SSRN, DOI 10.2139/ssrn.2023776.

Li W. (2002), *Zipf’s Law Everywhere*, “Glottometrics”, No. 5.

Lotka A. J. (1926), *The frequency distribution of scientific productivity*, “Journal of the Washington Academy of Sciences”, Vol. 16, No. 12.

Lucas Jr. R.E., Moll B. (2014), *Knowledge Growth and the Allocation of Time*, “Journal of Political Economy” , Vol. 122, No. 1.

Mandelbrot B. (1982), *The Fractal Geometry of Nature*, W. H. Freeman, San Francisco.

Mehri A., Lashkari S. M. (2016), *Power-law regularities in human language*, “The European Physical Journal B”, Vol 89, No. 241, DOI 10.1140/epjb/e2016-70423-9.

Mitzenmacher M. (2003), *A Brief History of Generative Models for Power Law and Lognormal Distributions*, “Internet Mathematics”, Vol. 1, No. 2.

Montemurro M. A., Zanette D. H. (2002), *New perspectives on Zipf’s law in linguistics: from single texts to large corpora*, “Glottometrics”, No. 4.

Newman M. E. J. (2005), *Power laws, Pareto distributions and Zipf's law*, “Contemporary Physics”, Vol. 46, No. 5, DOI 10.1080/00107510500052444.

Olmedilla M., Martinez-Torres M. R., Toral S. L. (2016), *Examining the power-law distribution among eWOM communities: a characterisation approach of the Long Tail*, “Technology Analysis & Strategic Management”, Vol. 28, No. 5, DOI 10.1080/09537325.2015.1122187.

Osareh F. (1996a), *Bibliometrics, citation analysis and co-citation analysis: a review of literature I*, “Libri”, Vol. 46, No 3, DOI 10.1515/libr.1996.46.3.149.

Osareh F. (1996b), *Bibliometrics, citation analysis and co-citation analysis: a review of literature II*, “Libri”, Vol. 46, No. 4, DOI 10.1515/libr.1996.46.4.217.

Pareto V. (1896), *Cours d’économie politique*, Librairie Droz, Lausanne.

Patience G. S., Patience C. A., Blais B., Bertrand F. (2017), *Citation analysis of scientific categories*, “Heliyon”, Vol. 3, No. 5. DOI 10.1016/j.heliyon.2017.e00300.

Perline R. (2005), *Strong, Weak and False Inverse Power Laws*, “Statistical Science”, Vol. 20, No. 1, DOI 10.1214/088342304000000215.

Piketty T., Zucman G. (2014), *Capital is Back: Wealth-Income Ratios in Rich Countries, 1700–2010*, “Quarterly Journal of Economics”, Vol. 129, No.3, DOI 10.1093/qje/qju018.

Plerou V., Gopikrishnan P., Stanley H. E. (2005), *Quantifying Fluctuations in Market Liquidity: Analysis of the Bid-Ask Spread*, “Physical Review E”, Vol. 71, No.4, DOI 10.1103/PhysRevE.71.046131.

Redner S. (1998), *How popular is your paper? An empirical study of the citation distribution*, “European Physical Journal B”, Vol. 4, No. 2.

Rosen K. T., Resnick M. (1980), *The size distribution of cities: an examination of the Pareto law and primacy*, “Journal of Urban Economics”, Vol. 8, No. 2, DOI 10.1016/0094-1190(80)90043-1.

Rosen S. (1981), *The Economics of Superstars*, “American Economic Review”, Vol. 71, No. 5.

Rousseau R. (2002), *George Kingsley Zipf: life, ideas, his law and informetrics*, “Glottometrics”, Np. 3.

Rousseau R., Zhang Q. (1992), *Zipf’s data on the frequency of Chinese words revisited*, “Scientometrics”, Vol. 24, No.2.

Schroeder M. (1991), *Fractals, Chaos, Power Laws*, Freeman, New York.

Serbyn M., Michailidis A. A., Abanin D. A., Papić Z. (2016), *Power-Law Entanglement Spectrum in Many-Body Localized Phases*, “Physical Review Letters”, No. 117, DOI 10.1103/PhysRevLett.117.160601.

Silagadze Z. K. (1997), *Citations and the Zipf-Mandelbrot’s law*, “Complex Systems”, Vol. 11, No. 6.

Simon H. (1955), *On a Class of Skew Distribution Functions*, “Biometrika”, Vol. 42, No. 3-4, DOI 10.1093/biomet/42.3-4.425.

Soo K. T. (2005), *Zipf’s Law for cities: a cross country investigation*, “Regional Science and Urban Economics”, Vol. 35, No. 3, DOI 10.1016/j.regsciurbeco.2004.04.004.

Sornette D., Knopoff L., Kagan Y. Y., Vanneste C. (1996), *Rank-ordering statistics of extreme events: application to the distribution of large earthquakes*, “Journal of Geophysical Research”, Vol. 101, No. B6, DOI 10.1029/96JB00177.

Spaide R. F. (2016), *Choriocapillaris Flow Features Follow a Power Law Distribution: Implications for Characterization and Mechanisms of Disease Progression*, "American Journal of Ophthalmology", Vol. 170, DOI 10.1016/j.ajo.2016.07.023.

Sui J., Zheng L., Zhang X., Chen G. (2015), *Mixed convection heat transfer in power law fluids over a moving conveyor along an inclined plate*, "International Journal of Heat and Mass Transfer", Vol. 85, DOI 10.1016/j.ijheatmasstransfer.2015.02.014.

Toda A. A., Walsh K. (2015), *The Double Power Law in Consumption and Implications for Testing Euler Equations*, "Journal of Political Economy", Vol. 123, No. 5, DOI 10.1086/682729.

Wang L., Du J. (2017), *The diffusion of charged particles in the weakly ionized plasma with power-law kappa-distributions*, "Physics of Plasmas", Vol. 24, No. 10, DOI 10.1063/1.4996775.

White H., McCain K. W. (1989), *Bibliometrics*, "Annual Review of Information Science Technology", Vol. 24.

Wyllys R. E. (1981), *Empirical and theoretical bases of Zipf's law*, "Library Trends", Vol. 30, No. 1.

Yule G. U. (1925), *A Mathematical Theory of Evolution. Based on the Conclusions of Dr J. C. Willis, F.R.S.*, "Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character", Vol. 213.

Zipf G. K. (1949), *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Addison-Wesley Press Inc., Cambridge (MA).