


Piotr GABRIELCZAK\*

 0000-0002-9032-7204

Tomasz SERWACH\*\*

 0000-0002-8253-1029

### Firm-Size Distribution in Poland: Is Power Law Applicable?

---

**Abstract:** This article focuses on the existence of power laws in the firm-size distribution in Poland. Specifically, we empirically test whether the size distribution of companies in Poland has the characteristics of Zipf’s law, a special case of power law observed in many different contexts in empirical economic literature. Our analysis uses 2019 data on the 2,000 largest companies in Poland as ranked by the *Rzeczpospolita* daily newspaper in its “Lista 2000” (Top 2,000 List). We reviewed theoretical mechanisms generating power laws and used several estimators of the power-law exponent in our empirical analysis. Our results confirm statistically significant deviations from Zipf’s law in the firm-size distribution in Poland. We found evidence that the power law cannot satisfactorily approximate the sales-based distribution of firms.

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

**JEL classification codes:** C46, D39, L25

---

Article submitted August 23, 2020, revision received February 13, 2021,  
accepted for publication March 18, 2021.

---

---

\* Department of Macroeconomics, Institute of Economics, University of Lodz, Poland, e-mail: piotr.gabrielczak@uni.lodz.pl

\*\* Department of International Trade, Institute of Economics, University of Lodz, Poland, e-mail: tomasz.serwach@uni.lodz.pl

## Rozkład wielkości firm w Polsce – czy ma zastosowanie prawo potęgowe?

**Streszczenie:** Artykuł koncentruje się na istnieniu praw potęgowych w rozkładzie wielkości firm w Polsce. Przetestowano empirycznie, czy rozkład wielkości firm w Polsce ma cechy prawa Zipfa – szczególnego przypadku prawa potęgowego obserwowanego w wielu różnych kontekstach w literaturze ekonomicznej. W analizie wykorzystano dane z roku 2019, dotyczące 2000 największych przedsiębiorstw w Polsce, notowanych na Liście 2000 „Rzeczpospolitej”. Dokonano przeglądu teoretycznych mechanizmów generujących prawa potęgowe, a w analizie empirycznej zastosowano kilka estymatorów wykładnika potęgi. Uzyskane przez nas wyniki potwierdzają istotne statystycznie odchylenia od prawa Zipfa w przypadku rozkładu wielkości firm w Polsce. Znaleźliśmy dowody na to, że prawo potęgowe nie jest w stanie w zadowalający sposób aproksymować rozkładu firm opartego na sprzedaży.

**Słowa kluczowe:** prawo potęgowe, prawo Zipfa, rozkład wielkości firm, skalowanie

**Kody klasyfikacji JEL:** C46, D39, L25

---

Artykuł złożony 23 sierpnia 2020 r., w wersji poprawionej nadesłany 13 lutego 2021 r.,  
zaakceptowany 18 marca 2021 r.

---

## Introduction

The aim of the article is to investigate the size distribution of Polish companies. We will test a hypothesis that Polish companies are subject to a (weak) power-law distribution. We will also identify some of the possible consequences of such a situation.

Following the seminal papers by Axtel [2001], Gabaix [2009], and di Giovanni and Levchenko [2010], firm size is usually measured with employment or sales, with the latter gaining popularity in the more modern literature. The power law indicates a firm-size distribution with so-called fat tails, which means that, compared to the traditionally assumed normal distribution, the economy has a relatively strong representation of large companies (in terms of sales or employment). Gabaix refers to such a structure as the granular economy. The granular economy has an interesting feature whereby large companies may dominate in a specific sense, which in this case means that idiosyncratic shocks affecting them on a micro level may be translated into macroeconomic conditions. Were firm size distributed normally, we would expect individual shocks to cancel out, thus micro shocks would be irrelevant to the economy on a macro level.

The concept of the granular economy is therefore essential for understanding the economic mechanisms and providing proper policies, e.g. referring to controlling market power concentration or supporting competitiveness. Gabaix [2011] shows that the sum of idiosyncratic shocks to the largest companies, which he defines as the granular residual, is a significant determinant for business cycle fluctuations on a national level. The importance of that aspect, along with the growing awareness of the problem among main-

stream economists in the world, motivated us to raise the question about the distribution of firm size in Poland.

Section 2 includes a literature review. Section 3 incorporates the presentation of the methodology and data. Section 4 focuses on discussing the obtained results. The final section concludes.

### Literature review

Power laws have been widely observed in empirical research, starting from Zipf [1949] and throughout the second half of the 20<sup>th</sup> century and the first two decades of the 21<sup>st</sup> century. In economics, power laws were originally used to describe the distribution of income and wealth [Pareto, 1896]. This usage for power laws is still popular as many papers demonstrate the existence of power laws in either of cases [Atkinson et al., 2011; Benhabib et al., 2011] or increasingly in both cases jointly [Pickety, Zucman, 2014; Gabaix et al., 2016]. Considering these issues together is justified, as the power law distribution indicates inequalities – and inequalities in income distribution tend to cumulate into even larger inequalities in terms of wealth [Gabaix, 2016]. An investigation into the distribution of income led to widening the use of power laws to include areas directly associated with income, such as productivity [Lucas, Moll, 2014] and consumption [Toda, Walsh, 2015]. Some focus has also been brought to capital markets, where it is possible to observe power-law distributions of returns, daily numbers of transactions and 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 distribution of firm size seems to be informative. Many papers focus on testing the power laws for the distribution of firm size. For example, Axtel [2001] was one of the first to focus on the distribution of firm size in the context of power laws, specifically Zipf's law. He calculated firm-size distributions for American companies based on the size of their employment, considering data for all the companies in the COMPUSTAT base (roughly about 5 million firms in each of the analysed years: 1988–1997). He managed to show that, even though the number of companies and their average size grew, the distribution was well approximated by a power law with an exponent close to 1 – Zipf's law. Gabaix [2009] first replicated the results by Axtel [2001] to show that the American economy is granular and subject to Zipf's law – only this time it was supported by data on both employment and sales. Later Gabaix [2011] took a slightly different approach and proved that the sum of productivity shocks of the 100 largest American companies weighted with their sales-to-GDP ratio (the aforementioned granular residual) is a significant independent variable for a factor model predicting the GDP dynamics in the United States. Di Giovanni et al. [2011] also replicate

the power law estimation, but this time proving that the firm-size distribution in the case of French companies can also be approximated by Zipf's law. What's more, they show that the power-law exponent for exporters is lower in absolute terms than for non-exporters, which indicates that the distribution of exporting firms is systematically more fat-tailed, thus granularity among exporters is even stronger.

Gabaix [2016] demonstrates the rationale behind the occurrence of multiple power laws among economic phenomena. Power laws appear as a result of mechanisms arising from the proportional random growth theory. Let us assume that we observe a set of companies that grow or shrink randomly due to independent shocks, but at the same time they satisfy Gibrat's law, which states that all companies have the same expected growth rate (with the same standard deviation). Such a model only makes it possible to draw a conclusion that in time the distribution of firms should tend toward a 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 ensure that, one more condition is needed: the assumption of a lower bound of the firm's size. Now this model technically produces a steady-state distribution, which has the form of a power law. However, it does not necessarily have to be Zipf's law, which means that the exponent does not have to be close to 1. However, Gabaix [2009] 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, in light of the proportional random growth theory and Gibrat's law, Zipf's law is a steady-state distribution for companies if we assume the following conditions:

- (1) The economy has a finite size in terms of the number of enterprises,
- (2) There is a lower bound for the size of a company and it is relatively low,
- (3) 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 include an infinity of companies. The second condition is also realistic. If we measure the size of a company with its employment then the lower bound of 1 is the only logical. If we measure the size of a company with its sales then we can still claim that, because of existing fixed, but not sunk, costs<sup>1</sup>, maintaining a company is only rational if it obtains a minimum turnover at 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 the average company, and this condition is obvious in such models. However, Gabaix [2016] notes that there are many microeconomic models that assume the occurrence of economies of scale. Therefore a bigger company should be more cost-effi-

---

<sup>1</sup> We mean fixed costs that could be recovered should the company exit the market.

cient and able to grow even faster than a smaller company. In such a case, 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 institutional factors such as stricter tax regulations and less support for bigger companies. Another possibility is simply that economies of scale might exist, but they are not very strong. That would be enough for many processes to be estimated as power laws with exponents close to 1. Therefore, based on the insight from Gabaix [2016], this last condition can be significantly relaxed:

(3') Companies either demonstrate constant economies of scale or weak economies of scale or the economy incorporates exogeneous (e.g. institutional) mechanisms that mitigate economies of scale.

In fact, intuition suggests that if an economy is granular, this should be considered as proof that the market structure is far from the model of perfect competition. Since, in a model free-market economy (without institutional or informational frictions), monopolisation processes may arise only due to economies of scale, then the relaxed condition 3' should be considered more realistic than its original formulation. The existence of economies of scale should be expected at least in the case of the largest companies.

Surprisingly, the power-law distribution of company sizes may affect the 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 & Landier [2008] focused on top managers in their analyses, but in fact their conclusions could be drawn to apply to top artists or athletes. They proved that if companies have a power-law distribution and there is no upper bound for the size (or resourcefulness) of the company, then even small differences in the skilfulness or talent of a scarce group of workers will translate into 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. Using 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 laws as a rule for size distribution makes sense in such a case, it is worth stressing that estimating Zipf's law for the largest companies may serve a different purpose. The existence of Zipf's law for a reasonable sample of the largest companies indicates a fat right tail 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].

Power laws are also commonly used in non-economic empirical research. They are popular in linguistics [e.g. Kucera, Francis, 1967; Altmann, 2002; Ellis et al., 2015; Mehri, Lashkari, 2016]; bibliometrics [e.g. Lotka, 1926; Wyllys, 1981; White, McCain, 1989; Clough et al., 2015; Patience et al., 2017]; and analysis of internet traffic [e.g. Mitzenmacher, 2003; Olmedilla et al., 2016; Bokányi et al., 2019]. Power laws are also strongly present in urban analyses and geography [e.g. Hill, 1970; Gabaix, 1999; Brakman et al., 2001; Soo, 2005; Edwards, Batty, 2015]. To a lesser extent they have been utilised in natural sciences [e.g. Mandelbrot, 1982; Schroeder, 1991; Bak, 1996; Sornette et al., 1996; Serbyn et al., 2016] and in applied sciences, such as medicine [e.g. Spaide, 2016] and engineering [e.g. Sui et al., 2015; Biswas et al., 2017, Wang, Du, 2017].

### Methodology and data

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

$$Y = kX^\alpha \quad (1)$$

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

Power laws have been popular in recent economic literature, with a growing number of relations confirmed to have their features. One of the most popular power laws is Zipf's law, which is a special case of empirically derived power law with exponent  $\alpha \cong -1$ .

$$Y = kX^{-1} = \frac{k}{X} \quad (2)$$

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 logarithmic ranks with the logarithmic size of the population. A linear regression generates a line with a slope of approximately  $-1$ . This is equivalent to (2).

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

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

$$A := \ln k$$

$$\ln Y = A - \ln X \quad (3)$$

This means that the population of a city with rank  $n$  is proportional to  $1/n$ , which is the inverse rank<sup>2</sup>. 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 population size. Indeed, Gabaix [2009] 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 the logarithmic 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 the case of Zipf's law the exponent of the power law is only approximately  $-1$ , let us be more general and denote this exponent as  $\zeta$ , holding (also for the convenience of interpretation) that  $\zeta$  is a positive number. Now we can formulate the "distributional" power law as follows:

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

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

$$\zeta \approx 1 \quad (5)$$

Because we assume that  $\zeta$  is positive and explicitly add a minus, stressing that the exponent of the power law is negative, therefore (4) is sometimes referred to as the inverse power 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 called the Pareto-Zipf (power) distribution [Perline, 2005].

Power laws are popular, but Perline [2005] claims that in some cases researchers use power laws where in fact these 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 truncation of data. Truncation is a process of selecting a sample only out of observations with the highest ranks (when referring to a 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 lakes by size [e.g. Mandelbrot, 1982]. Calculations based on a sample of lakes lead to the conclusion that there is a power law in the distribution of lakes by size. However, Perline stresses that the division between lakes and ponds, for example, is some-

---

<sup>2</sup> 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].



what arbitrary. Perhaps then the research only includes the right tail, while the left tail of the distribution is excluded because we do not consider ponds as small lakes.

Similarly, in economics it is much easier to find data about medium-sized and large companies, but not about micro enterprises. This distinguishes strong power laws, which are fully certain, from weak power laws, in which case we can only investigate the right tail of the distribution. Disturbances in the left tail, 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<sup>3</sup>.

When Perline [2005] talks about false power laws, he refers to a 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 the top 10 steel plants, while Perline [2005] shows that including 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 logarithmic 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 a growing number of researchers [e.g. Newman, 2005; Gabaix, 2009; di Giovanni et al., 2011] explicitly state that their research is focused on the right tail and the results may not apply to the left tail of the distribution. Therefore, for scrutiny and safety reasons, they assume that their results are a case of weak rather than strong power laws<sup>4</sup>.

In our study, we applied data provided by the *Rzeczpospolita* daily. That popular and renowned newspaper compiles lists of the largest firms operating in Poland. These lists are labelled "Lista 500" (Top 500) and "Lista 2000" (Top 2,000), indicating the number of firms analysed. Instead of being mere rankings of firms, "Lista 500" and "Lista 2000" present various firm-level data, including total sales, employment, total assets, and equity.

---

<sup>3</sup> Variable  $S$  has a log-normal distribution when  $\ln(S)$  has a normal distribution. The log-normal distribution is very close to a power law in its right tail. Yule [1925] suggested a distribution similar to a power law, with the counter-cumulative distribution function being  $P(S > x) = Ax^{-\zeta}B^{-x}$ . However, in most cases  $B$  is close to 1, so the Yule distribution can be easily mistaken for a power law [Simon, 1955].

<sup>4</sup> A similar reservation could be made with regard to our research, which is specifically focused on the right-hand tail of the distribution. However, it is that part of the distribution that has a potentially significant impact on the economy and economic policies. That is why we find such a reservation irrelevant from a practical point of view.



We used the latest “Lista 2000” ranking available online – the one for 2020 with data for the fiscal year 2019 [Lista 2000, 2020]. 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 a power law may only be observed for firms with a size bigger than a certain threshold. The use of “Lista 2000” allowed us to check what happens when we restrict our sample to 1,500, 1,000 or 500 firms instead of relying on the whole ranking.

“Lista 2000” and similar rankings may be criticised for a potential selection bias. However, they are still commonly used in empirical economic literature that touches on topics related to enterprises. Examples include Doryń and Stachera [2008], and Jaworek et al. [2018]. In fact, *Rzeczpospolita*’s “Lista 2000” is compiled with significant attention paid to preventing a selection bias<sup>5</sup>. Once potential candidates for the list are selected based on data from previous years, they are asked to complete a survey focusing on their current data. If a company does not respond, *Rzeczpospolita* tries to gather information about it from Bisnode, a contracted data collection agency, which also provides information for the initial selection. If this fails, then the data is supplemented with official statistics requested from the government. Only if all these efforts prove ineffective, the company is dropped from the sample. Therefore one could argue that “Lista 2000” is a reliable source of information on the largest companies in Poland. Furthermore, alternative commercial databases suffer from similar limitations. For instance, as shown by Kalemli-Özcan et al. [2019], in 2012 firms from the ORBIS database covered only 59% of Poland’s gross output (the average for the 1999–2012 period was 54%). The indicator of firm size used in our study is total sales. Other possible indicators, such as employment and total assets, have some missing values. *Rzeczpospolita* builds its league table on the basis of sales, hence that particular indicator was complete for the entire dataset. According to our data, the biggest firm in Poland was an oil refiner and petrol retailer, with sales of around PLN 110 billion (roughly 5.2% of Poland’s GDP<sup>6</sup>). The second-biggest firm was a foreign-owned company active in food distribution and specialized retail, with sales of PLN 51 billion (2.4% of GDP). In third place was an oil and gas company, with sales of PLN 41 billion (1.9% of GDP). The last company on the list (ranked 2,000th) was a designer, manufacturer and service provider active in the field of energy and optimization solutions for aerospace and industrial markets.

<sup>5</sup> We would like to express our gratitude to the editorial staff of *Rzeczpospolita* for sharing the methodological details of the process.

<sup>6</sup> One may question the comparison between sales (based on total) and GDP (based on value added). However, in the literature, the ratio of these variables serves as a good approximation of the significance of an individual firm in a given economy. See, for instance, Gabaix [2011]. Moreover, according to Hulten’s Theorem, the impact of a microeconomic (idiosyncratic) productivity shock on aggregate TFP growth depends on a firm’s sales’ share in GDP. Specifically, aggregate productivity growth is given by the sum of firm-level productivity shocks multiplied by their sales-to-GDP ratios [see Hulten, 1978].

Its sales were around PLN 222 million (0.01% of GDP). The average sales calculated for the whole set of firms were just over PLN 1.1 billion (0.05% of GDP), while the median was almost PLN 448 million (0.02% of GDP). The huge difference between the average and the median indicates a significant skewness of the firm-size distribution in Poland.

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

$$\ln(\text{Rank}_i) = \ln(k) - \zeta \ln(S_i) \quad (6)$$

$\text{Rank}_i$  is the rank of the  $i$ -th company,  $S_i$  is the sales of the  $i$ -th company,  $k$  is a technical parameter that does not have an interpretation, while  $\zeta$  is the estimated power-law exponent.

The second approach was based on the definition of the power law as in (4).

$$\ln(P(S > S_i)) = \ln(k) - \zeta \ln(S_i) \quad (7)$$

The left-hand side is the number of firms in the sample with sales higher than  $S_i$  divided by the total number of firms. Then we regress the natural log of this probability on log sales and the remaining notation is as in (6).

Finally, we used the Gabaix-Ibragimov [2011] estimator, in which case it is necessary to regress the natural log ( $\text{Rank}_i - 0.5$ ) of each firm in the sales distribution on its logarithmic sales. The traditional OLS estimation as in (6) may be biased if the sample is too small. According to Gabaix and Ibragimov [2011], a small correction to the formula significantly reduces the potential small-sample bias.

$$\ln(\text{Rank}_i - 0.5) = \ln(k) - \zeta \ln(S_i) \quad (8)$$

Regardless of which estimation technique is selected, we expect  $\zeta$  to be close to 1. However, validating Zipf's law requires us to determine if it is statistically significantly different from 1. The standard t-test procedure provided by most statistical packages tests the parameters against 0. That is why we suggest two alternative approaches to verification against 1. One of them is based on altered regression formulas, where we manipulated with the equations so that the parameter standing next to the logarithm of sales is expected to be 0. In all three estimation techniques, it seems enough to add the logarithm of sales to both sides of the equation. Thus we achieve verification equations for the parameters obtained from our three estimators respectively:

$$\ln(\text{Rank}_i \cdot S_i) = \ln(k) + (1 - \zeta) \ln(S_i) \quad (9)$$

$$\ln((P(S > S_i)) \cdot S_i) = \ln(k) + (1 - \zeta) \ln(S_i) \quad (10)$$

$$\ln((Rank_i - 0.5) \cdot S_i) = \ln(k) + (1 - \zeta) \ln(S_i) \quad (11)$$

Equations (9), (10) and (11) are not very practical for interpretation. They are also subject to a serious endogeneity problem. Their only useful feature is they make it possible to verify if  $(1 - \zeta)$  is indeed different from 0 or not, which automatically determines if  $\zeta$  is different from 1 or not.

Another approach we use is based on a variation of the t-Student test, which was introduced by Welch [1947] in order to compare estimations of identically specified equations<sup>7</sup>. In our case, we do not have two estimations, but we could compare the actual results to a hypothetical estimation with an identical number of observations and standard error but specific  $\zeta = 1$ .

## Results

The results are divided into two categories. The core findings refer to research into the entire sample of companies or sub-samples of the full sample based on cut-offs limiting sample size under the assumption of a power-law distribution. Our additional results look at alternative distributions.

### Power-law firm-size distribution

Our results are summarized in tables that include estimations based on different methods described in section 3. The following results, presented in Table 1, are based on the whole set of firms (2,000 in total). The methods are listed using their shortened names, where *lnRank* stands for an estimation based on the rank of logarithmic sales, as in (6); *lnP\_ccdf* stands for an estimation based on the logarithmic values of the counter cumulative distribution function<sup>8</sup>, as in (7); and *Gab\_Ibrag* stands for Gabaix-Ibragimov estimator, as specified in (8). In all the specifications  $\zeta$  denoted the absolute value of the parameter; however, it was negative, thus the coefficients with *lnSales* are in fact negative.

<sup>7</sup> The Welch [1947] procedure assumes that we compare estimations of identically specified equations  $i$  and  $j$ , which were originally estimated on two different samples. Two equivalent parameters  $\zeta_i$  and  $\zeta_j$  and their standard errors  $SE_i$  and  $SE_j$  can be used to calculate the test statistic:

$$t_{ij} = \frac{|\zeta_i - \zeta_j|}{\sqrt{SE_i^2 + SE_j^2}}$$

Statistic  $t_{ij}$  has to be compared with the proper critical values from the t-Student distribution or, if the number of observations is sufficiently high, from the normal distribution. When  $t_{ij}$  is higher than the proper critical value, we reject the null hypothesis that both equivalent parameters are in fact the same in favour of an alternative hypothesis that they are different.

<sup>8</sup> The number of observations in this method is reduced by 1. This is because for the top company  $P(S_i > x) = 0$  and the logarithms' domain is positive real numbers.

**Table 1. Estimation results – 2,000 firms**

	lnRank	lnP_ccdf	Gab_Ibrag
lnSales	-1.108*** (0.002)	-1.116*** (0.002)	-1.115*** (0.002)
Cons.	21.298*** (0.023)	13.802*** (0.028)	21.392*** (0.027)
R-squared	0.995	0.993	0.993
N	2000	1999	2000

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Source: authors' own elaboration.

As one may observe, the results obtained with the use of different methods are quantitatively similar. In all cases they are above 1 in absolute terms, which might indicate a pattern.

Due to the existence of the so-called weak power-law distributions, which are well approximated by a power law only at the very right-hand end of the distribution, we found it interesting to see how the estimation might change if we limit the number of observations to a smaller number of the largest (by sales) companies. Table 2 summarises the results for 1,500 firms.

**Table 2. Estimation results – 1,500 firms**

	lnRank	lnP_ccdf	Gab_Ibrag
lnSales	-1.130*** (0.002)	-1.140*** (0.003)	-1.139*** (0.003)
Cons.	21.610*** (0.029)	14.435*** (0.036)	21.733*** (0.034)
R-squared	0.995	0.992	0.993
N	1500	1499	1500

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Source: authors' own elaboration.

**Table 3. Estimation results – 1,000 firms**

	lnRank	lnP_ccdf	Gab_Ibrag
lnSales	-1.154*** (0.003)	-1.169*** (0.004)	-1.167*** (0.004)
Cons.	21.964*** (0.042)	15.252*** (0.053)	22.137*** (0.050)
R-squared	0.993	0.990	0.991
N	1000	999	1000

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Source: authors' own elaboration.

The results for 1,500 companies seem to be even further from 1 compared to the results for the entire sample of 2,000 firms. We decided to test if a further reduction of the sample size would lead to systematic changes in the estimated parameter. Tables 3 presents the results for estimations based on a subsample of 1,000 companies.

Table 4 contains results obtained with the use of data for only 500 companies with the largest sales. In the case of results for both 1,000 and 500 companies, one can observe that the estimated power-law exponents tend to deviate further and further from 1.

**Table 4. Estimation results – 500 firms**

	lnRank	lnP_ccdf	Gab_Ibrag
lnSales	-1.206*** (0.006)	-1.231*** (0.007)	-1.228*** (0.007)
Cons.	22.747*** (0.081)	16.893*** (0.105)	23.062*** (0.097)
R-squared	0.989	0.983	0.986
N	500	499	500

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Source: authors' own elaboration.

Table 5 provides the results of regression (11), which is a modified version of the Gabaix-Ibragimov estimator<sup>9</sup> in order to test the power-law exponent against 1. If parameter  $(1 - \zeta)$  is not statistically significant, then the estimation may be treated as proof for Zipf's law. Otherwise, the obtained power laws are significantly different from it.

**Table 5. Testing  $\zeta$  against 1, results for Gabaix-Ibragimov estimator, regression-based**

	N = 2000	N = 1500	N = 1000	N = 500
$(1 - \zeta)$	-0.115*** (0.002)	-0.139*** (0.003)	-0.167*** (0.004)	-0.228*** (0.007)

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Source: authors' own elaboration.

Table 6 displays the results of a test based on the Welch [1947] procedure, where we checked if the obtained parameters were different from a hypothetically identical estimation with an exponent of exactly 1.

<sup>9</sup> For the sake of brevity, apart from the main regression results, we only report results based on the Gabaix-Ibragimov estimators, which we consider to be the most advanced method for investigating power laws. The results based on two other approaches are not substantially different and are available upon request.

**Table 6. Testing  $\zeta$  against 1, results for Gabaix-Ibragimov estimator, based on Welch [1947]**

	N = 2000	N = 1500	N = 1000	N = 500
tij > t	40.66***	32.76***	29.52***	23.03***

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Source: authors' own elaboration.

The results in Table 5 and Table 6 show that none of the previous estimations resulted in identifying a case of Zipf's law. Though the exponents were close to 1, they were still substantially different. In fact, it is possible to observe that when the sample size is smaller,  $\zeta$  becomes higher in terms of absolute values (further from 1). Since smaller absolute values of the power-law exponent indicate a fatter right-hand tail, our results show that the firm-size distribution in Poland has a significantly slimmer tail than Zipf's law and that once we focus on the largest firms (in terms of sales), the tail seems to be relatively thinner.

When comparing our results with the existing literature, one must be aware that research on Polish firms is rare and based on different data sources and periods. In the only study that presents the results of similar estimations, di Giovanni and Levchenko [2013] used a broader dataset (ORBIS) for the 2006–2008 period. They obtained an estimate of a power law coefficient equal to 1.086. Our result based on the full sample was similar by an order of magnitude, but distinctively different in terms of interpretation as it cannot allow us to indicate the existence of Zipf's law. This is because, contrary to the findings of di Giovanni and Levchenko [2013], our estimated coefficient turned out to be significantly different from 1. Moreover, we observed a clear pattern that whenever we analysed more firms, the absolute value of the coefficient was closer to 1 and at each step the results were statistically significant. The estimates suggest that there is a striking underrepresentation of the largest firms in Poland (compared to the distribution based on Zipf's law) and that the gap becomes wider whenever we restrict our research to a smaller sample. Such a change in the estimated parameter led us to conclude that (3) may be incorrect and that non-linearity could be at work.

### Alternative firm-size distributions

Perline [2005] suggested that power laws are sometimes selected for their usefulness and simplicity in modelling, while in fact they could be easily mistaken for other distributions, similar in terms of some of their traits. The two natural candidates for alternative distributions are a log-normal distribution and the Yule-Simon distribution [see section 3].

The logarithms of sales do not have a normal distribution in our sample. A basic skewness and kurtosis test or the Shapiro-Wilk test, which is recommended for non-aggregated data, provide evidence for that. Therefore, a log-normal distribution can be easily ruled out.

However, the Yule-Simon distribution<sup>10</sup> seems to be a potentially fitting choice. In fact, a comparison of the counter-cumulative distribution functions of both distributions shows that a power law is a specific case of the Yule-Simon distribution, while the latter allows for the logarithmic firm-size distribution to deviate from a linear form. It depends on whether the B parameter differs from 1 or not. Assuming that A, B and  $\zeta$  are positive parameters, the Yule-Simon counter-cumulative distribution function can be expressed as:

$$P(S > S_i) = AS_i^{-\zeta} B^{-S_i} \quad (12)$$

We linearise (12) in order to estimate the parameters of the Yule-Simon distribution.

$$\ln(P(S > S_i)) = \ln(A) - \zeta \ln(S_i) - \ln(B)S_i \quad (13)$$

Table 7 presents the results of the estimation. Please note that the signs which were directly presented in the specification of (13) are now incorporated into the estimated parameter values.

**Table 7. Estimation results – Yule-Simon distribution**

	lnP_ccdf
Sales	-0.000*** (0.000)
lnSales	-1.022*** (0.001)
Cons.	12.610*** (0.013)
R-squared	0.999
N	1999

\* p < 0.1, \*\* p < 0.05, \*\*\* p < 0.01

Source: authors' own elaboration.

The negative signs of the coefficients of sales and the logarithm of sales are consistent with expectations. Our results clearly demonstrate that the firm-size distribution in Poland is well approximated by the Yule-Simon distribution. To be exact, the coefficient of sales was estimated as 0.0000000452, thus the B parameter is just above 1. Its deviation is small, yet significant. However, if we neglect this difference, the estimated distribution becomes close to that of Zipf's law. Although, in reality the parameter with the logarithm of sales is also significantly different from 1. Nevertheless, this alleged similarity, along with the relative popularity of power laws in the literature, explains why the firm-size distribution creates the impression of being a case of Zipf's law.

<sup>10</sup> Introduced by Yule [1925] and developed by Simon [1955].



Our results, contrary to those of di Giovanni and Levchenko [2013], suggest that the distribution of Polish companies is a false power law. This observation has consequences for the characteristics of the Polish economy. The Yule-Simon distribution has a slimmer right-hand tail and could be associated with lower granularity. This, on the other hand, should result in lower susceptibility to idiosyncratic shocks on a macro level. However, for this conjecture to be validated, more detailed research is needed into the impact of granular residuals on macroeconomic volatility.

## Conclusions

The latest commonly available data on the sales of Poland's 2,000 largest companies proves that the distribution of firms across the country by size is not well approximated by a power law. We found strong evidence against the existence of Zipf's law in Poland, since the power-law exponent deviates from 1 in absolute terms. It seems that the right-hand tail of the firm-size distribution in Poland may be slimmer than in the case of Zipf's law. Moreover, since the estimation of the power-law exponent is sensitive to the extent of concentration of the largest companies only, it is also possible that the logarithmic firm-size distribution in Poland is characterized by non-linearity and that the existing studies err by missing that point. It is possible that the economic literature identifies only a weak power law at best.

Our research proves that an alternative known as the Yule-Simon distribution fits the data on Polish companies much better. It is characterised by a distribution function very similar to a power law, but it allows non-linearity of the relationship between probability and the logarithm of sales. In fact, this framework is more general and the power law can be treated as a special case of the Yule-Simon distribution. Furthermore, compared to Zipf's law, the identified distribution has a slimmer right-hand tail. This indicates lower granularity and a lower risk of idiosyncratic shocks transferred to the economy. However, the consequences of such an alternative firm-size distribution require further research.

All things considered, our results stand in opposition to widely recognised empirical research into the firm-size distribution in Poland so far. While di Giovanni and Levchenko [2013] identified Zipf's law, we believe that it is a false power law. So are power laws even applicable when analysing the firm-size distribution in Poland? They may be, but only as a far-reaching simplification, e.g. as an element of auxiliary analyses.

## References

- Altmann G. [2002], Zipfian linguistics, *Glottometrics*, 3: 19–26.
- Atkinson A.B., Piketty T., Saez E. [2011], Top Incomes in the Long Run of History, *Journal of Economic Literature*, 49(1): 3–71.

- Axtell R.L. [2001], Zipf Distribution of US Firm Sizes, *Science*, 293(5536): 1818–1820.
- Bak P. [1996], *How Nature Works*, Copernicus, New York.
- Benhabib J., Bisin A., Zhu S. [2011], The Distribution of Wealth and Fiscal Policy in Economies with Finitely Lived Agents, *Econometrica*, 79(1): 123–157.
- Biswas A., Triki H., Zhou Q., Moshokoa S.P., Ullah M.Z., Belic M. [2017], Cubic – quartic optical solitons in Kerr and power law media, *Optik*, 144: 357–362.
- Bokányi E., Kondor D., Vattay G. [2019], Scaling in words on Twitter, *Royal Society Open Science*, 6(2): 1–11.
- Bouchaud J.-P., Farmer J.D., Lillo F. [2009], How Markets Slowly Digest Changes in Supply and Demand, in: K.R. Schenk-Hoppe, T. Hens, (eds.), *Handbook of Financial Markets: Dynamics and Evolution*: 57–160, North-Holland, Amsterdam.
- Brakman S., Garretsen H., van Marrewijk C. [2001], *An Introduction to Geographical Economics*, Cambridge University Press, Cambridge.
- Clough J.R., Gollings J., Loach T.V., Evans T.S. [2015], Transitive reduction of citation networks, *Journal of Complex Networks*, 3(2): 189–203.
- Di Giovanni J., Levchenko A.A. [2013], Firm entry, trade, and welfare in Zipf’s world, *Journal of International Economics*, 89(2): 283–296.
- Di Giovanni J., Levchenko A.A., Rancièrè R. [2011], Power laws in firm size and openness to trade: Measurement and implications, *Journal of International Economics*, 85(1): 45–52.
- Doryń W., Stachera D. [2008], Wpływ internacjonalizacji na wyniki ekonomiczne największych polskich przedsiębiorstw przemysłowych, *Gospodarka Narodowa*, 11–12: 95–114.
- Edwards R., Batty M. [2015], City size: Spatial dynamics as temporal flows, *Environment and Planning A*, 48(6): 1–3.
- Ellis N.C., O’Donnell M.B., Römer U. [2015], Usage-Based Language Learning, in: B. MacWhinney, W. O’Grady (eds.), *The Handbook of Language Emergence*: 69–87, John Wiley & Sons, Chichester.
- Gabaix X. [1999], Zipf’s law for cities: an explanation, *Quarterly Journal of Economics*, 114(3): 739–767.
- Gabaix X. [2008], Power Laws, in: S.N. Durlauf, L.E. Blume (eds), *The New Palgrave Dictionary of Economics*, Palgrave Macmillan, London.
- Gabaix X. [2009], Power Laws in Economics and Finance, *Annual Review of Economics*, 1(1): 256–293.
- Gabaix X. [2011], The Granular Origins of Aggregate Fluctuations, *Econometrica*, 79(3): 733–772.
- Gabaix X. [2016], Power Laws in Economics: An Introduction, *Journal of Economic Perspectives*, 30(1): 185–205.
- 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*, 29(1): 24–39.
- Gabaix X., Landier A. [2008], Why Has CEO Pay Increased So Much?, *Quarterly Journal of Economics*, 123(1): 49–100.
- Gabaix X., Lasry J.-M., Lions P.-L., Moll B. [2016], The Dynamics of Inequality, *Econometrica*, 84(6): 2071–2111.
- 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*, 60(5): 305–316.

- Hill B.M. [1970], Zipf's law and prior distributions for the composition of a population, *Journal of the American Statistical Association*, 65(331): 1220–1232.
- Hulten C. [1978], Growth Accounting with Intermediary Inputs, *Review of Economic Studies*, 45: 511–518.
- Jaworek M., Karaszewski W., Kuczmarska M. [2018], Przedsiębiorstwa z udziałem kapitału zagranicznego na tle ogółu przedsiębiorstw w Polsce w okresie 1994–2017, *Przegląd Organizacji*, 9: 6–14.
- Kalemli-Özcan S., Sørensen B.E., Villegas-Sanchez C., Volosovych V., Yeşiltaş S. [2019], *How to Construct Nationally Representative Firm Level Data from the Orbis Global Database: New Facts and Aggregate Implications*, Tinbergen Institute Discussion Paper No. TI 2015–110/IV.
- Kromer V. [2002], Zipf's law and its modification possibilities, *Glottometrics*, 5: 1–13.
- 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](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2023776) (23 August 2020).
- Li W. [2002], Zipf's Law Everywhere, *Glottometrics*, 5: 14–21.
- Lista 2000 [2020], *Rzeczpospolita*, <https://rankingi.rp.pl/lista2000/2020> (accessed: 31 May 2021).
- Lotka A.J. [1926], The frequency distribution of scientific productivity, *Journal of the Washington Academy of Sciences*, 16(12): 317–323.
- Lucas Jr.R.E., Moll B. [2014], Knowledge Growth and the Allocation of Time, *Journal of Political Economy*, 122(1): 1–51.
- 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*, 89(241): 1–6.
- Mitzenmacher M. [2003], A Brief History of Generative Models for Power Law and Lognormal Distributions, *Internet Mathematics*, 1(2): 226–251.
- Newman M.E.J. [2005], Power laws, Pareto distributions and Zipf's law, *Contemporary Physics*, 46(5): 323–351.
- 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*, 28(5): 601–613.
- 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*, 3(5): 1–24.
- Perline R. [2005], Strong, Weak and False Inverse Power Laws, *Statistical Science*, 20(1): 68–88.
- Piketty T., Zucman G. [2014], Capital is Back: Wealth-Income Ratios in Rich Countries, 1700–2010, *Quarterly Journal of Economics*, 129(3): 1255–1310.
- Plerou V., Gopikrishnan P., Stanley H.E. [2005], Quantifying Fluctuations in Market Liquidity: Analysis of the Bid-Ask Spread, *Physical Review E*, 71(4): 1–8.
- Rosen S. [1981], The Economics of Superstars, *American Economic Review*, 71(5): 845–858.
- 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*, 117: 1–6.
- Simon H. [1955], On a Class of Skew Distribution Functions, *Biometrika*, 42 (3–4): 425–440.
- Soo K.T. [2005], Zipf’s Law for cities: a cross-country investigation, *Regional Science and Urban Economics*, 35(3): 239–263.
- 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*, 101 (B6): 13883–13893.
- Spaide R.F. [2016], Choriocapillaris Flow Features Follow a Power Law Distribution: Implications for Characterization and Mechanisms of Disease Progression, *American Journal of Ophthalmology*, 170: 58–67.
- 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*, 85: 1023–1033.
- Toda A.A., Walsh K. [2015], The Double Power Law in Consumption and Implications for Testing Euler Equations, *Journal of Political Economy*, 123(5): 1177–1200.
- Wang L., Du J. [2017], The diffusion of charged particles in the weakly ionized plasma with power-law kappa-distributions, *Physics of Plasmas*, 24(10): 1–4.
- Welch B.L. [1947], The Generalization of ‘Student’s’ Problem when Several Different Population Variances are Involved, *Biometrika*, 34 (1/2): 28–35.
- White H., McCain K.W. [1989], Bibliometrics, *Annual Review of Information Science Technology*, 24: 119–186.
- Wyllys R.E. [1981], Empirical and theoretical bases of Zipf’s law, *Library Trends*, 30(1): 53–64.
- 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*, 213: 21–87.
- Zipf G.K. [1949], *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Addison-Wesley Press Inc., Cambridge (MA).