Features and Determinants of the Firm Size Distribution: An Empirical Analysis with Brazilian Data*

Miguel Nathan Foguel† Eduardo Pontual Ribeiro‡

Abstract

The Pareto distribution has been used to describe firm sizes in many theoretical models for its convenience and empirical validity. We provide estimates of the Pareto parameters across industries and investigate the determinants of the shape of the firm size distribution in Brazil. The Pareto tail distribution is not rejected for about 70% of the industries, with the Zipf tail distribution (scale coefficient equal to 1) accepted for about 50% of the industries. The size distributions in manufacturing and all industries are affected by the human capital intensity and may be affected by industry uncertainty and instability, in line with the literature.

Resumo

A distribuição Pareto tem sido usada para descrever o tamanho de firmas em modelos teóricos pela sua conveniência e validade empírica. Trazemos estimativas do parâmetro da Pareta em todos os setores da economia brasileira e investigamos os determinantes do formato da distribuição de tamanho no Brasil. A distribuição Pareto para as firmas maiores não é rejeitada em 70% dos casos, sendo a hipótese de Zipf (parâmetro igual a 1) aceita em metade dos setores. A distribuição de tamanho da indústria e entre todos os setores é afetada pela intensidade de capital humano e pode ser influenciada pela incerteza e instabilidade setorial, o que está em linha com a literatura.

Keywords: Firm size distribution; Pareto distribution; Human capital.
Palavras-chave: Distribuição de tamanho; Distribuição de Pareto; Capital Humano
JEL: C12, D30, L10

1 Introduction

The firm size distribution in a country can influence an ample set of economic variables. This includes the level and quality of employment, earnings inequality, the degree of competition, the innovation rate, and the productivity of the economy. Investigating the features and the determinants of this distribution is thus key not only to understand some of the underlying factors that affect the functioning of the economy but also to shed some light onto which areas policy makers should focus their attention.

Given its importance, it is not surprising that the firm size distribution has received a lot attention in the academic literature. There is a large set of studies that focus on the statistical properties of this distribution (e.g., Axtell, 2001; Fujiwara et al., 2004; Growiec et al., 2008;

---

*We thank the financial support of CNPq and superb research assitance by Glaucia Ferreira.
†Researcher at IPEA. E-mail: miguel.foguel@ipea.gov.br
‡Professor at UFRJ. E-mail: eribeiro@ie.ufrj.br

Gabbaix e Ibragimov, 2011; Halvarsson, 2013), and there is wide consensus that the tails of the firm size distribution in many countries is consistent with a Power Law (or Pareto distribution).[1] Additionally, the literature has considered a special case of the Pareto distribution, namely the Zipf distribution, for which the shape parameter is equal to unity. The empirical regularity on the Pareto distribution and its special Zipf case implies that the frequency of firms' sizes is inversely proportional to their sizes, at least above a minimum threshold.

A growing theoretical literature investigates the factors that shape the firm size distribution. Lucas Jr. (1978) relates the size distribution with entrepreneurial talent and Jovanovic (1982) argues that firms' survival is related to the selection process on the learning of firms' true productivity after entry in the market. Ericson and Pakes (1995) propose that the selection process is determined by the sequence of productivity shocks that affect firms' decisions and Luttmer (2007) includes the size of entry costs and imitation capacity of firms as important factors, in addition to the productivity shocks. In a different strand, Cabral and Mata (2003) argue that the existence of financial frictions hinder the growth of small firms, leading to firm size distributions with thinner tails (i.e., with relatively fewer large firms and a large shape parameter of the Pareto distribution). Rossi-Hansberg and Wright (2007) put forward a model based on efficient allocation of production factors and industry-specific characteristics that predicts that industries in which industry-specific human capital is more intense tend to have size distributions with thicker tails.

This theoretical literature provides a conceptual reference that can be used to empirically investigate the importance of a set of determinants of the firm size distribution. There are only a few studies that pursue this line of investigation, namely Henrekson e Johansson (1999) e Halvarsson (2013), both of which based on data from Sweden. This paper seeks to not only fill part of this gap but also characterize the firm size distribution and its determinants for a developing country such as Brazil. To the best of our knowledge, this is the first study that does that for a developing country.[2]

To accomplish that we employ the two-stage methodology of Halvarsson (2013), who adapted the empirical strategy used by Ioannides et al. (2008), Rosen and Resnick (1980), and Soo (2005) in the context of city size distributions. In the first stage, the shape parameter of the Pareto distribution is estimated for different 3-digit industries over the years. Given that the Pareto distribution is more appropriate for firms' sizes above a certain (unknown) level, we follow Halvarsson (2013) and employ the procedure proposed by Clauset et al. (2009) to determine this level. The Clauset et al. (2009) method is based on the Kolgomorov-Smirnov (KS) test. However, as Goerlich (2013) points outs, this method has some limitations and thus we also employ the Lagrange Multiplier (LM) test proposed by Goerlich (2013) to pinpoint the minimum size level. Using the minimum threshold sizes determined by the aforementioned methods, we estimate the shape parameter using the rank-size log-linear regression proposed by Gabaix and Ibragimov (2011). The second stage uses the estimates of the shape parameters from the first stage as the dependent variable in pooled and panel regressions that are run against a set of explanatory variables that intend to measure relevant determinants of the shape of tail of the firm size distribution.

The main source of information used in the study is the *Relação Anual de Informações Sociais (RAIS - Annual Roll of Social Information)*, a very large administrative dataset collected

[1]The Pareto distribution has also been used to model many economic phenomena including the distribution of wealth, income, financial variables, and city size; see Gabaix (2009) for a survey.

[2]There is some literature that discusses whether the firm size distribution in developing countries exhibit the so-called "missing middle", that is, a relative scarcity of firms of medium size in the distribution (see, e.g., Tybout, 2000; Hsieh and Olken, 2014; Coelho et al., 2017). Our empirical analysis is based on the characterization of the firm size distribution above minimum size thresholds, which are empirically determined. Our results can thus be useful for this literature, for instance by helping demarcating the upper tail of the size distribution.

by the Ministry of Labor that contains information on every labor contract on the formal sector in Brazil. Apart from data on the number of firms' employees during the year, the available information includes the industry of firms. This allows us to construct industry-level firm size distribution for every year between 2007 and 2019. We also use RAIS to construct some of the explanatory variables used in the second stage regression. Industry-level information for the manufacturing sector is also used for the second stage.

Apart from this Introduction, the paper is organized in another five sections. In section 2, we discuss the theoretical literature that guides our choice of the determinants of the firm size distribution at the industry level. Section 3 presents the data in more detail and how we construct the explanatory variables that are used in the second stage. In section 4, we lay out the empirical methods that are employed both to estimate the shape parameters of the firm size distribution (first stage) and the regression models used in the second stage. Section 5 presents the results for the first and second stages and section 6 concludes.

2 Theoretical literature and economic determinants of the firm size distribution

Modelling the firm size distribution in a concise manner based on a known density distribution has attracted the attention of economists for decades (e.g., Quandt, 1966). An intimately related literature is the one on firm growth and the discussion whether firms' growth follow the Gibrat's Law, which states that growth is independent of size, with log-Normal shocks to size growth (see e.g. Ribeiro, 2007 and references therein). These two literatures integrate, as Axtell (2001) and Gabaix (2009) show that if firm growth follows Gibrat's Law with frictions, the cross-sectional distribution of firms would converge to the Power Law, where the Pareto distribution is a major example.

Recently, there has been interest in learning whether the Power Law, or more precisely the Zipf's law, would model firm sizes across industries as this would give rise to granularity (Gabaix, 2011; Giovanni and Levchenko, 2012; Gaubert and Itskhoki, 2021). In broad terms, this is the property that aggregate dynamics of an industry or an economy can be determined by shocks to a small number of large firms, when the underlying firm size distribution follows a Power Law with shape parameter close to unity, i.e. the Zipf's law.

While it may be too simple to consider that the actual firm size distribution follows a Pareto distribution, cross-country studies have suggested that the tail behavior of firm sizes follows such distribution. Recent analysis of modelling the full size distribution, including very small firms, suggests that the actual distribution is a mixture of Log-Normal or Exponential and, for the right tail, the Pareto distribution (e.g., Kondo et al., 2021 for the U.S.; Na et al., 2017 for Korea; and Resende and Cardoso, 2022 for Brazil). Given the aforementioned granularity effects, this suggests that focusing on the right tail of the size distribution is of its own interest.

Once the firm size distribution (or its upper tail) has been modelled by the Pareto distribution, one can associate cross-industry differences in the distributions (or the shape parameter which characterize that distribution) and industry characteristics. This can contribute to learn about how to make an economy more resilient to small-localized shocks with great magnification if industries could be made to have a relatively smaller proportion of very large firms, i.e., thinning the firm size distribution away from the Zipf's law and observe larger Pareto coefficients.

As mentioned in the Introduction, there are various theories that explains firm dynamics and size distributions that are based on selection mechanisms (e.g., Lucas Jr., 1978; Jovanovic, 1982; Hopenhayn, 1992; Ericson and Pakes, 1995; Luttmer, 2007). Based on alternative mechanisms, there are basically two main theories of firm size distribution that can be higlighted.[3] Rossi-Hansberg and Wright (2007) model firm growth based on a framework of efficient allocation of production factors and industry-specific differences in production technologies. Conditional on survival, industries with more industry-specific human capital would allow firms to grow more and reach larger sizes, leading to size distributions with higher frequency of larger firms. On the other hand, industries where economies of scale are significant and intensive in physical capital, the proportion of large firms will be smaller. In the realm of Pareto distributions, this would entail that industries that are intensive in human (physical) capital would have a thicker (thinner) firm size distribution, implying a smaller (larger) Pareto distribution parameter.

Cabral and Mata (2003) provide an alternative explanation for differences in firm size distributions, based on financial constraints. In industries where firms are more financially constrained, firm growth is more difficult. Firms need to use accumulated profits and own funds to invest and expand, limiting the chances of small and young firms to reach their optimal sizes. This would generate firm size distributions that have relatively fewer large firms. As these few firms would be much larger than the other firms, this would be compatible with a smaller Pareto parameter. See also Angelini and Generale (2008).

Other characteristics that may influence the firm size distribution are as follows. First, industries with more unstable market shares tend to be less concentrated and thus display size distributions with thinner tails (i.e., larger Pareto parameters). Thus, industry instability, or the instability in firm market shares or relative sizes, is a candidate determinant of the shape of the upper tail of industry size distributions (Halvarsson, 2013). A second candidate is industry uncertainty, or the standard deviation of firm growth. The idea, suggested by Daunfeldt and Elert (2013), is that in industries with heightened uncertainty, entry may fall and a thickening effect on the firm size distribution should be expected. A third characteristic is the age of firms in an industry. Daunfeldt and Elert (2013) summarize the literature showing that industries with more mature firms would have more firms beyond the minimum efficient scale and more larger firms, thus with a thicker firm size distribution (i.e., smaller Pareto parameter).

This is a limited, non-exhaustive list of characteristics that may influence the firm size distribution. The selection above is based on relevance and data availability, so that we can provide empirical evidence on the effect of these characteristics of cross-industry as well as time-varying firm size distribution parameters.

3 Data and construction of variables

The main data source used in this paper is the *Relação Anual de Informações Sociais (RAIS Annual Roll of Social Informationl)*. In Brazil, all tax-registered establishments have to report to the Ministry of Labor a set of information on every single labor contract they signed in the previous year, including workers' wages, the date of hiring and dismissal (if it happened), and the education level of workers. Establishments have a unique identifier and inform their industry at the 5-digit level using the *Classificação Nacional de Atividade Econômica (CNAE)*, a classification that is compatible with the International Standard Industrial Classification (ISIC). To avoid working with too fine aggregation levels, we follow the literature and work with the 3-digit aggregation level (Halvarsson, 2013). Although some of our empirical results are only for the manufacturing sector, the bulk of the analysis is for all industries (234 in total

[3]Kettle and Kortum (2004) set up a model of technological change and firm dynamics that is based on R&D expenditures by competing incumbent and entrant firms. The model predicts a stable, skewed firm size distribution and a neutral relationship between R&D expenditures (relative to revenue) and the size distribution. Since we do not have data on R&D expenditures, we do not consider their predictions in the empirical part of the paper.

at the 3-digit level).[4]

RAIS does not provide information on sales or valued added, so our size metric is the number of employees. As we know the employment spell of workers in the firm, the size variable for each firm is the average number of employees during the year.[5] Many theoretical models that rationalize the shape of the firm size distribution are for firms that survive over time. To be consistent with these models, our empirical analysis is conducted only for surviving firms within the most recent period in our data: 2015 to 2019.

As mentioned in the Introduction, we employ a two-stage approach whose second stage consists of a set of regressions in which the shape parameters estimated in the first stage are regressed against a set of explanatory variables. Table 1 exhibits these variables, their formulas, and sources of information.

We do not have access to microdata on firms' revenue, a common measure of firm size. We opted to work with the firms' payroll, which is available in our data and is highly correlated with revenue.[6] Based on this, we construct a set of variables using microdata on firms' payroll. All variables are aggregated at the 3-digit industry level and the time unit is the year. A smaller set of variables could be calculated using aggregated published data from a survey of firms in the manufacturing sector. Named *Pesquisa Industrial Anual (PIA)*, it is an annual sample for the entire manufacturing sector in Brazil and collects information on firms' revenue, value added, assets, payroll, and investments.

Considering the variables displayed in Table 1, as discussed in section 2, industries with higher levels of specific human capital should display thicker tails in their firm size distribution. As RAIS contains information on the schooling level of the firms' workers, we are able to compute measures of human capital from microdata. We opted to construct what we think is a reasonable proxy for industry-specific human capital, namely the proportion of workers at the industry level with at least a university degree.

Section 2 indicates that the capital intensity and financial restrictions influence the shape of the firm size distribution. The former is measured as the ratio between capital stock (total fixed assets and the imputed value for capital renting and leasing) and total revenue at the industry level. The latter is a proxy for financial restrictions faced by firms and is measured at the industry level as the ratio between net investment expenditures and cash flow, which is measured as the difference between total value added and total payroll, as in Bond et al. (2002). As it gauges the amount of investments that is made with the firm's own resources, it intends to capture the degree of financial dependence of the firm and, in this sense, can be seen as a proxy for the restrictions the firms face to obtain credit in the market.

As mentioned in section 2, two variables that receive significant attention in the literature are industry instability and industry uncertainty, with opposing effects on the Pareto scale parameter. The first is the sum of the absolute changes in firms' payroll and is intended to capture the degree of industry instability. Industries with higher instability are expected to have fewer larger firms, thus thinning the tails of the industry size distribution. The second variable, industry uncertainty, attempts to gauge the degree of volatility at the industry level and is measured by the standard deviation of the growth in firms' payroll over the years. It is expected to have a thickening effect on the tails of the industry size distribution.

As previously mentioned, the age of firms is a potential factor that can affect the firm size distribution. Indeed, one should expect that the more mature are the firms in an industry, the thicker the tails of its size distribution. To measure the age dimension, we use the average age

[4] In order to measure the firm size distribution more reliably, we exclude all 3-digit indutries with less than 200 firms in a given year. This filter deletes 18 industries.

[5] Note that this implies that the size variable becomes a (strictly positive) non-integer scalar.

[6] In most of the extensive literature on modelling firm productivity; it is a fixed proportion of revenues, e.g., Loecker and Syverson (2021).

Table 1: Explanatory variables for the second stage

Variable	Description	Formula	Source
Industry human capital	Proportion of workers with at least the university degree	$\frac{\sum_{i=1}^{n_{j,t}} \sum_{k=1}^{N_{i,j,t}} 1\{s(k,i,j,t) \geq 15\}}{\sum_{i=1}^{n_{j,t}} N_{i,j,t}}$	RAIS
Industry physical capital	Total assets as a share of total revenue (industry level)	$\frac{Assets_{j,t}}{Revenues_{j,t}}$	PIA
Financial dependence	Net investments as a share of net valued added (industry level)	$\frac{Net\ investiments_{j,t}}{Value\ added_{j,t} - Payroll_{j,t}}$	PIA
Industry instability	Sum of absolute changes in industry payroll	$\sum_{i=1}^{n_{j,t}} \lvert \Delta \frac{payroll_{i,j,t}}{\sum_{i=1}^{n_{j,t}} payroll_{i,j,t}} \rvert$	RAIS
Industry uncertainty	Standard deviation (SD) of growth in industry payroll	$SD_{j,t}[ln(\frac{payroll_{i,j,t}}{payroll_{i,j,t-1}})]$	RAIS
Industry age	Average age of firms in the industry	$\frac{1}{n_{j,t}} \sum_{i=1}^{n_{j,t}} age_{i,j,t}$	RAIS
Industry age squared	The square of average age of firms in the industry	$(\frac{1}{n_{j,t}} \sum_{i=1}^{n_{j,t}} age_{i,j,t}/10)^2$	RAIS
Industry size	Number of firms in industry as a share of all fims (in logs)	$\frac{ln n_{j,t}}{ln \sum_{j=1}^{n_t} n_{j,t}}$	RAIS
Industry growth	Yearly change in log industry payroll	$\Delta \sum_{i=1}^{n_{j,t}} ln(payroll_{i,j,t})$	RAIS

Notes: Firms are indexed by i and workers by k. $n_{j,t}$ represents the number of firms in industy j in year t and n_t is the total number of firms. $N_{i,j,t}$ represents the number of workers in firm i in industry j in year t. The Δ symbol denotes the time difference between two adjacent years. $1\{.\}$ denotes the indicator function that assumes value one (zero) if the its argument is true (false). s is the schooling level of worker i. RAIS corresponds to Relação Anual de Informações Sociais and PIA to Pesquisa Industrial Anual.

of firms at the industry level and its square to capture potential non-linear age effects.

Last, as in Halvarsson (2013), we include two control variables that may affect the shape of the firm size distribution, namely the size as well as the growth of the industry. To control for differences in industry size, we use a variable that corresponds to the share of firms in the industry with respect to all firms in any particular year. As for differences in industry growth, the control variable is measured as the log difference in the industry's total payroll between years.

4 Empirical methods

We follow Halvarsson (2013) in applying a two-stage empirical strategy to study the importance of a potential set of determinants of the firm size distribution at the industry level. In the first stage, the shape parameters of the firm size distribution across industries and years are estimated. In the second stage, these estimated parameters are used as the dependent variable in regressions whose explanatory variables are the determinants of the shape of the firm size distribution.[7] In the sequel, we present the details of each stage.

4.1 First stage

Firms are first sorted in descending order with respect to their size y, i.e., $y_{(1),j,t} > y_{(2),j,t} > ... > y_{(n),j,t}$ for industry j defined at the 3-digit level in year t. The relative position $1, 2, ..., n$ is the rank of the firm in the firm size distribution.

It is assumed that the firm size distribution follows a Power Law (or Pareto distribution), whose cumulative distribution function can be written as:

$$F(y) = 1 - (y/y_{min})^{-\zeta}, \tag{1}$$

where $\zeta \in (0, \infty)$ is the shape parameter and $y_{min} > 0$ corresponds to the lower boundary of the distribution. Under a Pareto distribution, there is a log-linear approximation relating a firm's rank to its size, given by:

$$ln(rank) = -\zeta\, ln[(size/size_{min})]. \tag{2}$$

We estimate the ζ coefficient using Gabaix and Ibragimov (2011) regression method, which is based on the rank log-linear result in (2), with the 1/2 adjusted dependent variable to reduce small sample bias. As proposed by Gabaix (2009), the regression allows testing for deviations from the Pareto distribution by including an additional coefficient associated with a quadratic term. If this coefficient is not significant, one may conclude that the Pareto distribution is a reasonable approximation to the firm size distribution. More specifically, the regression is:

$$ln(rank_{i,j,t} - 1/2) = \alpha_{j,t} - \zeta_{j,t}\, ln(size_{i,j,t}) + \gamma_{j,t}\, ln^2(size_{i,j,t} - s^*) + \varepsilon_{i,j,t}, \tag{3}$$

where $rank_{i,j,t}$ is the rank of firm $i = 1, ..., n$ in industry j in year t and $size_{i,j,t} \geq sizemin_{j,t}$. In order not to affect the estimate of $\zeta_{j,t}$, Gabaix (2009) suggests an adjustment to the quadratic size variable, based on the shift parameter $s^* = \frac{cov(ln^2(size_{i,j,t}), \ln(size_{i,j,t}))}{2var(\ln(size_{i,j,t}))}$. Note that the above regression is fit to data only for firms' sizes that are above $sizemin_{j,t}$, the minimum boundary size at the industry-year level.

Testing the significance of $\hat{\gamma}$ must take into account the positive autocorrelation of the dependent variable, which render the regression standard error biased. Under the null hypothesis, the correct formula is such that a significance test uses $\hat{\zeta}^2/\sqrt{2n}$ as standard error and the null is rejected at the 5% significance level if $\left|\hat{\gamma}_{j,t}\right| > 1.96\hat{\zeta}^2/\sqrt{2n}$.

When using the above model and estimating the Pareto distribution parameters, one must determine the boundary parameter $sizemin_{j,t}$. Clauset et al. (2009) propose a search algorithm to find the boundary point that minimizes a specific distance measure. We apply their method with the following steps: (1) estimate $\zeta_{j,t}$ using as minimum size each percentile of the $size_{i,j,t}$ distribution; (2) for each of the 100 estimated $\hat{\zeta}_{j,t}$, calculate the Kolmogorov-Smirnov (KS) test

[7] Daunfeldt and Elert (2013) also use a similar two-stage approach to investigate whether a set of industry- level characteristics matters for the validity of Gibrat's Law.

comparing the empirical distribution with a Pareto distribution using the estimated parameter; (3) select the $\hat{\zeta}_{j,t}$ with the smallest KS test statistic, provided its p-value is greater than 5%.

We complement the Clauset et al. (2009) method using an additional distribution adjustment test. Goerlich (2013) highlights that the Clauset et al. (2009) method requires simulated p-values from the use of a parameter estimate to generate the null distribution in the KS test. This limits the use of the KS test in the above procedure. Goerlich (2013) suggests a Lagrange Multiplier (LM) test with a Pareto distribution null hypothesis, much like the well-known Jarque-Bera test for Normal residuals in the classical linear regression setting. Using maximum likelihood in a family of Power Law distributions, where the Pareto distribution is nested, the Goerlich LM statistic follows a Chi-square 1 d.f. distribution in large samples and is given by

$$LM = nz^2 \frac{(\hat{\zeta}+2)(\hat{\zeta}+1)^4}{\hat{\zeta}}$$

where $z = \frac{\hat{\zeta}}{\zeta+1} - \frac{1}{n}\sum_{i=1}^{n} 1/size_{i,j,t}$. Consistent with the above procedure, we use a second run of Clauset et al. (2009) algorithm, with the LM test used in place of the KS test.

In sum, we implement the Clauset et al. (2009) algorithm using the KS statistic to select the minimum boundary size at industry-year level. Having selected the minimum size, we estimate $\zeta_{j,t}$ based on the Gabaix and Ibragimov (2011) regression method. As a robustness analysis, we also provide estimates of the $\zeta_{j,t}$'s when the Clauset et al. algorithm uses the LM test.

4.2 Second stage

The estimates of the shape parameter from the first stage ($\hat{\zeta}_{j,t}$) are used as the dependent variable in regressions where industries at the 3-digit level $j = 1, ..., 234$ constitute the cross sectional dimension and $t = 2015, ..., 2019$ are the sample years. The explanatory variables described in section 3 are a set of determinants of the industry-level firm size distributions and are arranged in the vector $X_{j,t}$. The second-stage regression is specified as:

$$\hat{\zeta}_{j,t} = \theta + X_{j,t-1}\beta + \mu_j + \lambda_t + u_{j,t}, \tag{5}$$

where θ is the intercept, μ_j is an industry-specific fixed effect that absorb any time-invariant characteristic that is idiosyncratic to the industry, λ_t corresponds to time dummies that capture aggregate business cycle effects, and $u_{j,t}$ is a disturbance term assumed to be strictly exogenous. Note that the explanatory variables enter the regression in a one-year lagged fashion as an attempt to circumvent potential problems of simultaneity. As a result, we lose one year of data. Our parameter of interest is β, which measures the partial correlation of the determinants with the shape of the firm size distribution.

As equation 5 shows, the dependent variable is an estimate of unknown industry shape parameters, which are obtained in the first stage. As is well known, estimation of this equation by OLS tends to be inefficient since the dependent variable is likely to be estimated with errors of measurement that induce heteroskedasticity and display autocorrelaction. As an attempt to correct for that, we employ the Driscoll and Kraay (1998)'s procedure, which delivers standard errors in the second stage regression that are heteroskedasticity- and autocorrelation-consistent and robust to various forms of cross sectional and temporal dependence (Hoechle, 2007).

In practice, to estimate equation 5 we use four different specifications. The first is weighted least square (WLS) where the weights are the inverse of the standard error of the slope coefficient from the first stage. The second specification uses Driscoll and Kraay (2008) standard errors (DriscOLS), which are robust to general forms of cross-section and time dependence. The third method, which is also based on Driscoll and Kraay standard errors, is a panel regression in which industry-specific effects are introduced and treated as fixed effects. The fourth method considers the industry-specific effects as fixed effects. Second stage regressions are run only with estimates of $\zeta_{j,t}$ for which the Power Law was not rejected in the first stage regressions.

5 Results

In this section, we first present the estimates of the first stage regression, which allows us to characterize the shape of the (tail of the) firm size distribution. Next, we discuss the results of the second stage regressions, where the slope parameter of the first stage is regressed onto a set of explanatory variables. The analysis of the second stage results is conducted both for the overall sample of all sectors and for a sample of the industrial sector, for which we have a larger set of explanatory variables.

5.1 First stage results

We start by presenting descriptive statistics for the minimum size of the firm size distribution across the industries and years in our sample of surviving firms within the period 2015-2019. The statistics presented in this subsection are computed for the samples when the minimum size is selected either by the KS test or by the LM test.

Table 2a exhibits descriptive statistics of the distribution of minimum sizes. The mean (median) centile of the distribution selected as best fit as minimum size for the Pareto distribution when using the KS criterion is 74.4 (77), while for the LM it is much higher at 88.9 (93). The 1^{st} and 99^{th} percentiles are also higher for the latter criterion. The descriptive statistics for the minimum size in terms of the corresponding number of employees are presented in Table 2b. As expected, this table shows that the minimum sizes based on the LM test are always higher than those of the KS test. For instance, the mean (median) minimum size for the former is 321.1 (56.7) workers, while it is 68,6 (16.5) for the latter criterion. Overall, these results show that adherence to the Pareto distribution tends to occur for the higher parts of the firm size distribution across industries and years, especially when the LM criterion is used.
Table 2c presents descripti~~v~~ve statistics for the estimated coefficient $\hat{\zeta}_{j,t}$ in regression (3) for which the Pareto distribution is not rejected at the 5% significance level for the coefficient of the quadratic term, $\hat{\gamma}_{j,t}$. From a total of 1079 industries and years in our sample, the significance test for $\hat{\gamma}_{j,t}$ is not rejected for 69.1% (84.2%) when using the KS (LM) minimum size criterion. As it can be seen in Table 2c, the mean (median) of $\hat{\zeta}_{j,t}$ based on the KS test is 1.24 (1.20) and ~~the~~ 1^{st} (99^{th}) percentile is 0.60 (2.35). The statistics based on the LM test including the standard deviation are always larger than the ones based on the KS test. This can be confirmed in Figure 1, which presents the kernel densities of the $\hat{\zeta}_{j,t}$'s based on each test. Interestingly, the KS mode is close to one, where the Power Law is known as Zipf's Law.
Scatterplots can help illustrate the relative difference by industry-year between the $\hat{\zeta}_{j,t}$ estimated by the two methods. Figure 2 shows the parameter estimates, indicating that in most cases the estimate under a minimum size chosen by the KS criterion is smaller than the LM criterion. Note that the LM criterion tends to select larger (log) sizes as the best fit minimum size, as seen in Figure 3 (overall distribution) and Figure 4 (industry-year comparison). It appears that a larger minimum size is associated with a larger (thinner tail) distribution parameter.

An overall picture of the results for the power law exponent can be seen in Figure 5. It presents the estimated $\hat{\zeta}_{j,t}$'s and their 95% confidence intervals across the industry-year samples based on the KS minimum size criterion and when the the power law is not rejected. The horizontal line at the unity value represents the Zipf's law. As it can be seen from the figure, there is a substantial number of cases for which the the Zipf's law cannot be rejected. More precisely, the 5% significance-level test for $\zeta = 1$ does not reject the Zipf's law in 42% of the cases. Considering manufacturing only industries, the Zipf's law is not rejected in 45.6% of cases. Halvarsson (2013) reports a corresponding statistic for Sweden of 73% (445 out of 611 cases), which indicates that the Zipf's law regularity found in the literature (e.g., Axtell, 2001; Fujiwara et al., 2004) is relevant but not so strong in Brazil.

To be complete, Table 2d presents summary statistics for the quadratic coefficient in regression (3). It shows that the median estimate is negative, which indicates that for a large number of cases the firm size distribution has thinner tails than the Power Law. In contrast, for some regressions the quadratic coefficient is positive, meaning a convex deviation from the Power Law that thickens the tails of the distribution.

Table 2a – Descriptive Statistics – Industry Minimum Size Centiles

Selection Method	Mean	Median	Std Dev	1% Perc.	99% Perc.
KS test	74.40	77	15.22	30	98
LM test	88.88	93	11.58	45	99

Table 2b – Descriptive Statistics – Industry Minimum size

Selection Method	Mean	Median	Std Dev	1% Perc.	99% Perc.
KS test	68.62	16.50	288.35	3.61	1107.93
LM test	321.09	56.69	1313.99	3.75	4246.81

Table 2c – Descriptive Statistics – Zeta estimates

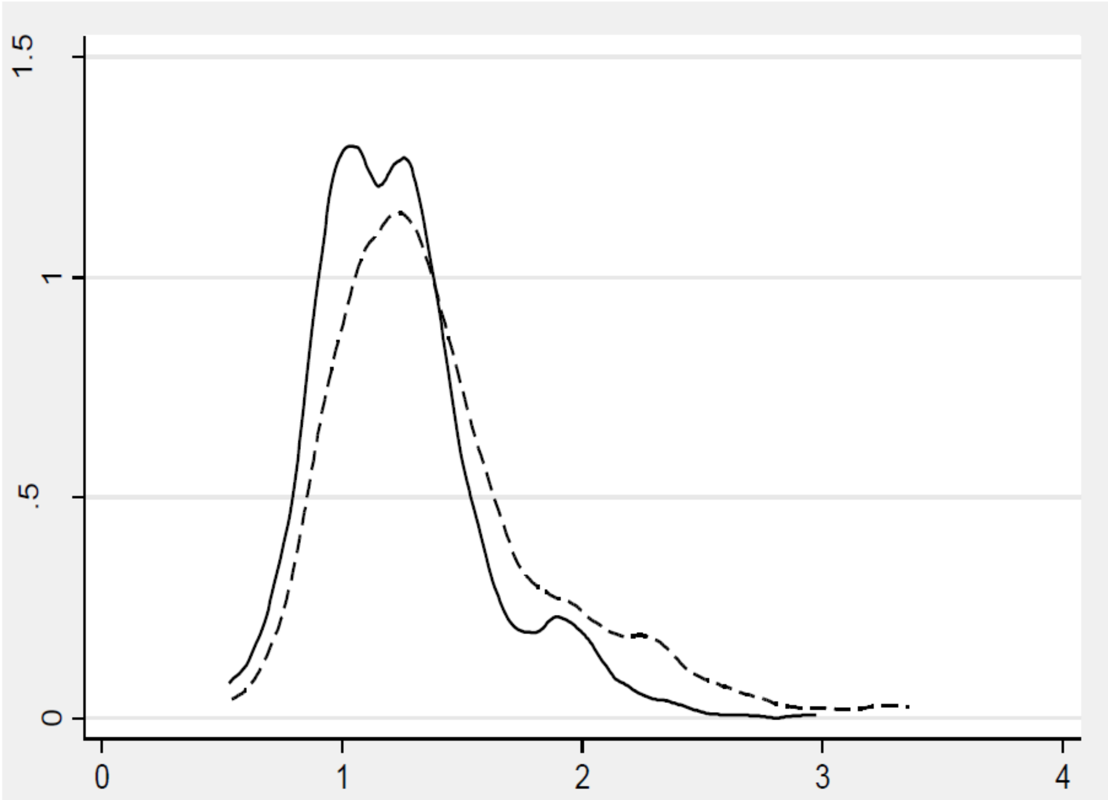
Selection Method	Mean	Median	Std Dev	1% Perc.	99% Perc.
KS test	1.24	1.20	0.37	0.60	2.35
LM test	1.49	1.32	0.99	0.63	4.20

Table 2d – Descriptive Statistics – Gamma estimates

Selection Method	Mean	Median	Std Dev	1% Perc.	99% Perc.
KS test	-0.05	-0.04	0.12	-0.41	0.30
LM test	-0.55	-0.04	9.39	-6.44	0.91

Notes: the KS and LM tests refer to the criterion used to determine the minimum size of the firm size distribution across the 1,079 industry-year samples. The firm size distribution (panel b) is measured by the average number of employees in a year. Zeta (Gamma) refers to the linear (quadratic) coefficient in regression (3) in the text. The Zeta coefficients on panel (c) refer to those industries-year where the Gamma coefficient is not significant, not rejecting the Pareto hypothesis (746 samples for KS and 908 samples for LM).

Figure 1 – Kernel densities of Zetas from regression (3), by minimum size selection criterion



Note: Solid Line – KS criteria; Dashed Line –LM criteria.

Figure 2 – Zetas from regression (3) – industry comparison by minimum size selection criteria

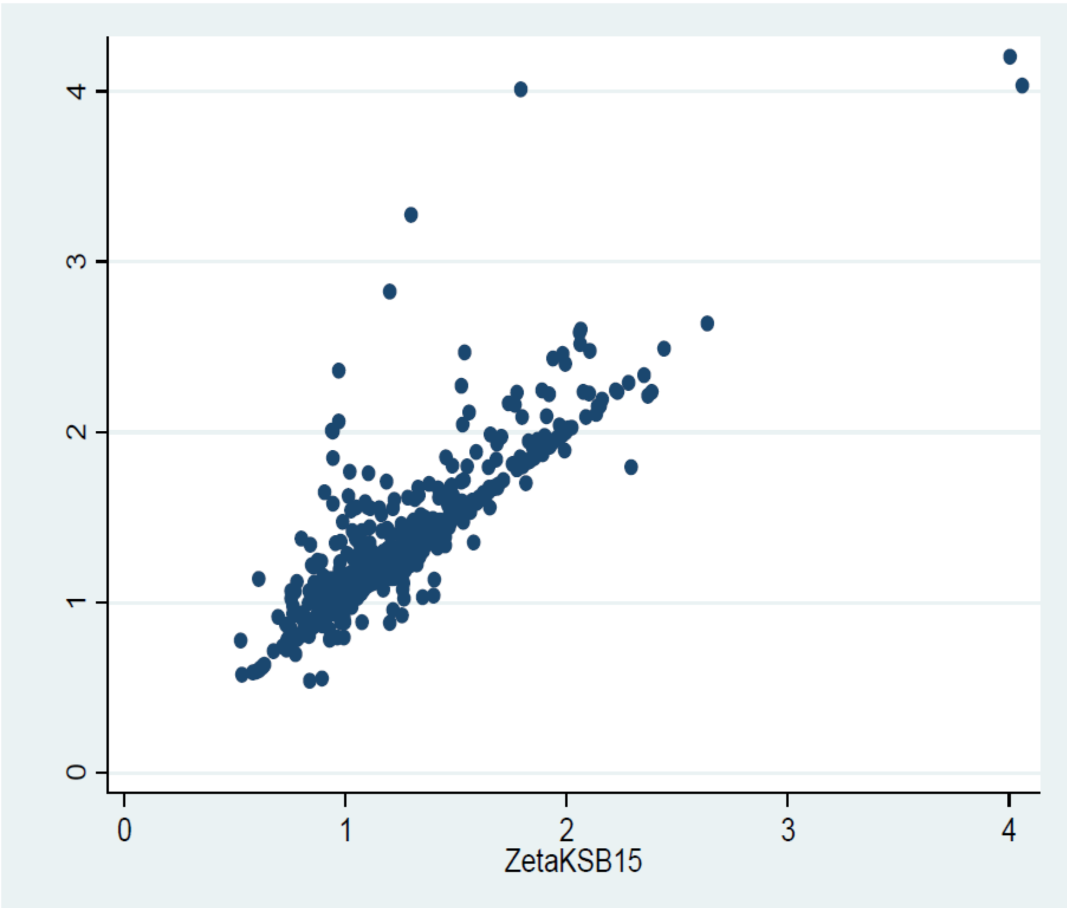
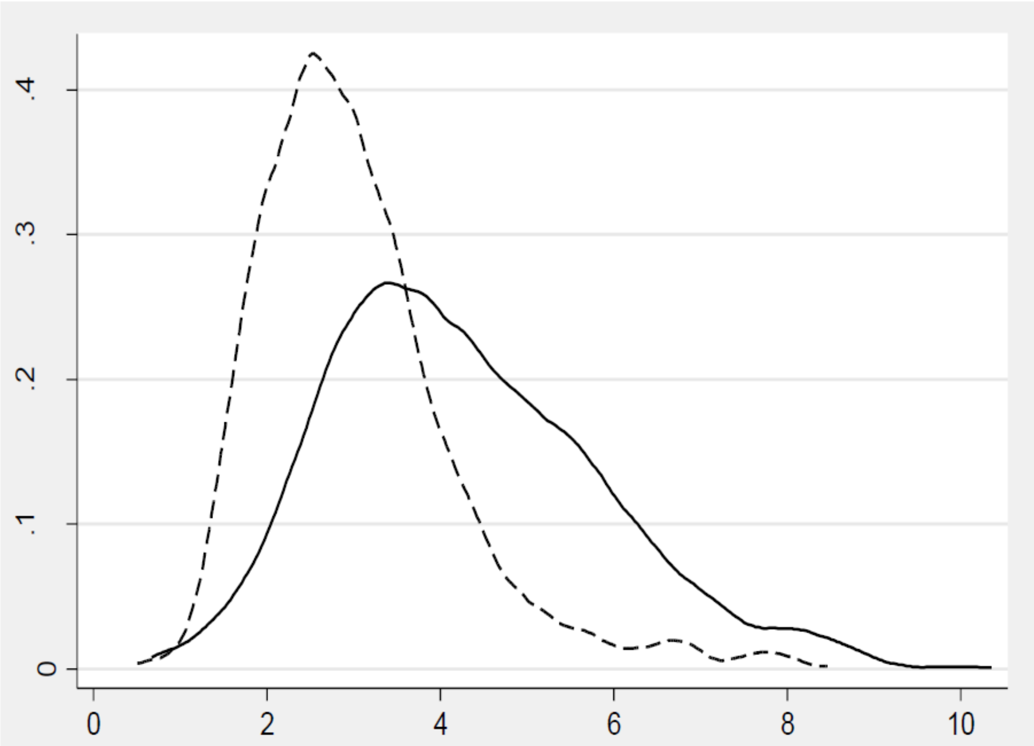
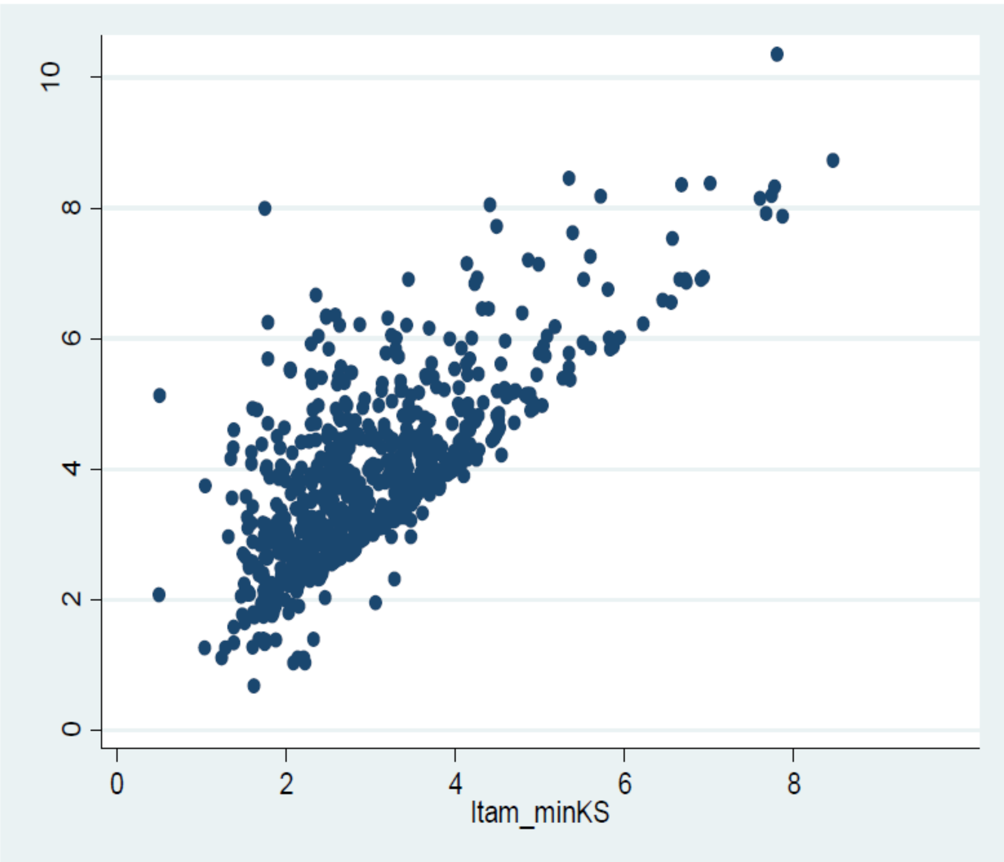


Figure 3 – Kernel densities of log minimum size by selection criterion



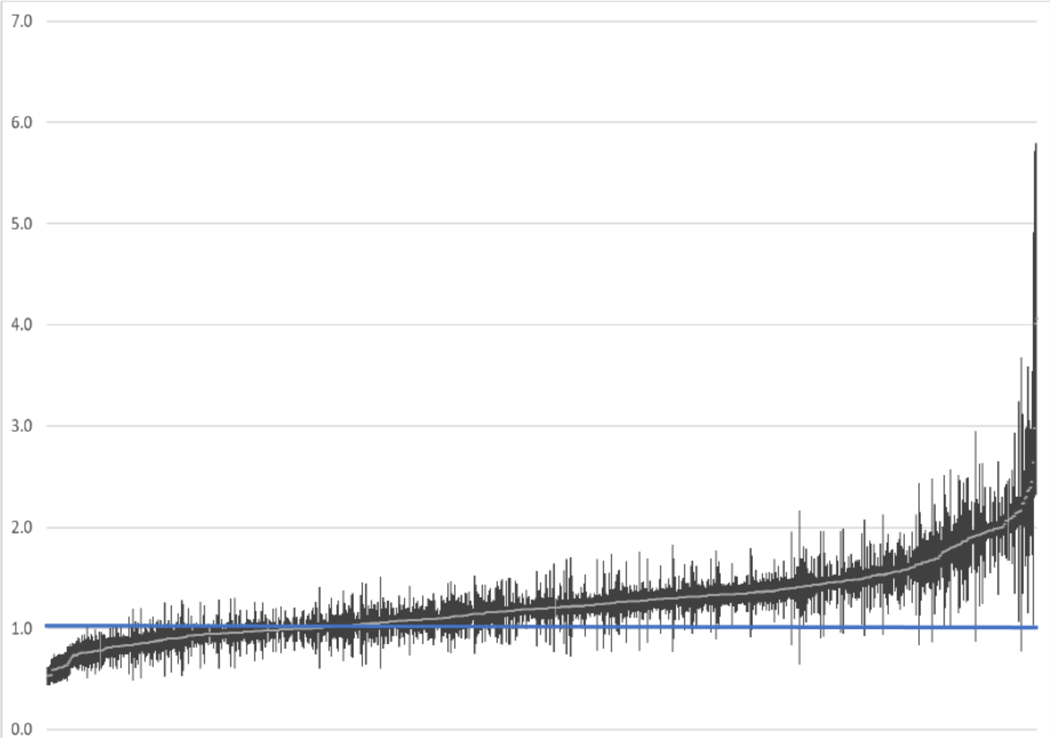
Note: Solid Line – KS criteria; Dashed Line –LM criteria.

Figure 4 –industry comparison of log minimum size selected by LM and KS criteria



Note: Vertical Axis: log minimum size selected by LM criteria. Horizontal Axis: log minimum size selected by KS criteria.

Figure 5 – Zeta estimates from regression (3) and their confidence intervals.



Note: yellow line- Zeta estimates from regression (3) – KS minimum size criterion. Vertical bars: 95% confidence intervals. Solid blue line indicates the Zipf's Law parameter (Zeta=1). 42% of industry – years we cannot reject the Zipf´s Law hypothesis.

5.2 Second stage results

Table 3 presents the results of the second stage regression (5) for the sample of all industries. Table 4 contains the results for the sample of subsectors of the manufacturing industry, for which we were able to construct two additional explanatory variables, namely proxies for physical capital intensity and financial frictions. In interpreting the results it is important to bear in mind that a positive (negative) coefficient is associated with a thinning (thickening) effect on the (tails of the) firm size distribution. We recall that all explanatory variables enter the regressions with one-year lag as an attempt to circumvent potential endogeneity problems. Estimates are presented under two parameter estimation methods and two alternative assumptions on the heterosckedasticity and autocorrelation of the error term in equation (5). We use pooled LS and fixed effects (FE) models, where the latter attempts to control for any unobserved, long-run (invariant under our 5-year time window) differences across industries. We use both homoskedastic and non-serially correlated errors parameter covariance matrix and Driscroll and Kray (1998) standard errors. All models use weights for the industries based on the sample size used to estimate the dependent variable in the first stage.

5.2.1 All industries

Table 3 shows that the human capital variable has a negative and statistically significant effect across all specifications of the second stage regression. This result corroborates the prediction of Rossi-Hansberg and Wright (2007) that higher use of human capital at the industry level should thicken the firm size distribution. Fixed effects controls increase the coefficient (in absolute terms) but the result is robust to different estimation methods.

Industry instability displays a positive and statistically significant coefficient in both the weighted least square model (Pooled LS) and the specification with Discroll and Kraay (2008) standard errors (DriscLS). However, the sign of the coefficient is reversed in the panel specifications, although the coefficient estimates become insignificant at conventional significance levels. These results, which are in line with those found by Halvarsson (2013) for Sweden, do not fully endorse the prediction pointed out in section 3 that industry instability should have a thinning effect on the tails of the firm size distribution.

Table 3 – Estimates from second stage regressions – All industries

	LS	DriscLS	DriscFE	FE
Human Capital	-0.377	-0.377	-0.716	-0.716
	(0.051)**	(0.023)**	(0.047)**	(0.079)**
Instability	0.918	0.918	-0.010	-0.010
	(0.106)**	(0.051)**	(0.045)	(0.046)
Uncertainty	-0.910	-0.910	-0.116	-0.116
	(0.111)**	(0.109)**	(0.064)	(0.092)
Age	-0.005	-0.005	0.004	0.004
	(0.004)	(0.002)*	(0.003)	(0.008)
$(Age/10)^2$	0.003	0.003	0.075	0.075
	(0.028)	(0.014)	(0.034)*	(0.057)
Growth	-0.155	-0.155	-0.009	-0.009
	(0.110)	(0.190)	(0.031)	(0.038)
Size	-0.535	-0.535	10.475	10.475
	(0.976)	(1.053)	(3.166)**	(12.106)
Constant	0.422	0.539	0.235	0.222
	(0.100)**	(0.060)**	(0.105)*	(0.120)
R^2	0.21	0.34	.	0.19

Notes: Zeta estimates from regression (3) based on distribution with minimum size selected by the KS criteria. Sample size 746 observations * p<0.05; ** p<0.01

The point estimates of the coefficient for industry uncertainty are negative for all four specifications, though, as for the instability parameter, they lose significance in the panel models. A thickening effect was expected and obtained in the pooled data estimates, but the introduction of industry-specific unobserved controls in the panel models does not allow a confirmation of this prediction. In the case of uncertainty (instability), industries with higher than average uncertainty have lower (higher) than average parameter, but an increase in within industry uncertainty (instability) is not systematically associated with a decrease (an increase) in the distribution scale parameter.

The age of industries should have a thickening effect on the firm size distribution. Parameter signs and values indicate that such effect is estimated in the data for the least squares model but not for the panel estimates. However, most the coefficients for industry age are not significant on statistical grounds. Thus, at least for the sample of all industries, it seems that the age dimension does not affect the shape of the firm size distribution.

Inconclusive results appear for the control variables: for industry growth, all coefficients are clearly non-significant across specifications and for industry size most are insignificant and flip sign between the pooled and the panel specifications.

5.2.2 Manufacturing sample

Moving to the results for the sample of manufacturing industries, Table 4 shows that the effect of physical capital intensity is negative and statistically significant for PooledLS and DriscLS but becomes very small and insignificant for the panel models. Controlling for subsector specific effects (either random or fixed) thus ousts the effect of the degree of physical capital intensity on the shape of the firm size distribution.

The sign pattern of the effect of financial dependence is similar to that of physical capital, though it does not display statistical significance for the WLS and DriscOLS specifications and it is marginally significant for one of the panel models. It seems therefore that greater financial dependence which is interpreted as higher levels of financial frictions that especially constraint the growth of small firms (Cabral and Mata, 2003) is not determining the shape of firm size distribution, at least in the manufacturing industry in Brazil.

Table 4 – Estimates from second stage regressions – Manufacturing

	LS	DriscLS	DriscFE	Pooled FE
Capital Intensity	-0.060	-0.060	-0.000	-0.000
	(0.017)**	(0.014)**	(0.003)	(0.008)
Financial Depend	-0.038	-0.038	0.027	0.027
	(0.050)	(0.053)	(0.011)*	(0.015)
Human Capital	-1.201	-1.201	-0.114	-0.114
	(0.143)**	(0.025)**	(0.028)**	(0.196)
Instability	0.751	0.751	0.150	0.150
	(0.178)**	(0.146)**	(0.053)**	(0.065)*
Uncertainty	-1.214	-1.214	-0.179	-0.179
	(0.199)**	(0.074)**	(0.140)	(0.115)
Age	0.004	0.004	-0.035	-0.035
	(0.009)	(0.001)**	(0.004)**	(0.012)**
$(Age/10)^2$	-0.135	-0.135	0.093	0.093
	(0.046)**	(0.005)**	(0.046)*	(0.092)
Growth	-0.069	-0.069	-0.013	-0.013
	(0.175)	(0.137)	(0.026)	(0.059)
Size	0.998	0.998	-2.788	-2.788
	(4.098)	(0.831)	(4.039)	(18.170)
Constant	0.974	0.000	0.659	0.608
	(0.161)**	(0.000)	(0.131)**	(0.181)**
R^2	0.51	0.71	.	0.20

Notes: Zeta estimates from regression (3) based on distribution with minimum size selected by the KS criteria. Sample size 270 observations * $p<0.05$; ** $p<0.01$

Like the sample for all industries, the effect of industry-specific human capital displays a negative sign across all specifications. All estimates but the one for the fixed-effect model exhibit statistical significance at the 1% level. This indicates that the (expected) thickening effect of human capital at the industry level may be relevant even when industry-specific fixed effects are controlled for and when richer specifications of the error structure are considered. The decrease (in absolute value) of the coefficient in the fixed effects specifications is remarkable and not observed for the overall sample of industries. It seems thus that unobserved heterogeneity between sectors plays a more important role in absorbing the effect of industry-specific human capital in the manufacturing industry than for the economy at large.

Differently from the results for the overall sample, industry instability now displays a positive and statistically significant coefficient across all specifications. This result is in contrast with the thinning effect that was expected from industries that display more instability.

Regarding the coefficient for industry uncertainty, it exhibits the same pattern in terms of sign and statistical significance as those for the overall sample. This means that its predicted thickening effect on the size distribution may not be fully confirmed for the manufacturing industry as well.

The slope coefficients for age and age square now become statistically significant for most model specifications. However, both coefficients flip sign between the pooled least square models and the pair of panel models. If one focuses on the latter, the estimates indicate that the age dimension has a thinning effect on the firm size distribution that is cushioned as the industries in the manufacturing sector become more mature, against what would be expected by the theory reviewed above.

As in the all industry results, the coefficients for industry size as well as for industry growth flip sign across pair of models and are not statistically meaningful in any specification. One may thus conclude that neither the size nor the growth of industries are associated with the shape of the firm size distribution.

5.2.3 Robustness estimates using LM test minimum size criteria

Before the concluding comments, we explore the robustness of the results by considering our newly proposed method to select the minimum size for the Pareto distribution parameter estimation that uses the LM test instead of the KS test as the criterion to select the minimum size. Tables 5 and 6, which are based on the LM criterion, provide analogues of tables 3 and 4. The qualitative results are mostly the same when the second stage is run based on estimates from the LM criterion, compared with the results when estimates are computed using a minimum size chosen by the KS criterion in tables 3 and 4. The main differences are on significance, where the regression coefficient estimates are sometimes not significant when they are in the KS-based estimates. The negative, thickening effect of industry-specific human capital is observed for the overall and the manufacturing samples, the partially significant results of uncertainty, instability, and capital intensity follow the same pattern as for the KS criterion, and the non-significance of financial dependence, age, growth, and size effects are also observed.

Table 5 – Estimates from second stage regressions – All industries – LM minimum size

	Pooled LS	DriscLS	DriscFE	FE
Human Capital	-0.390	-0.390	-0.588	-0.588
	(0.052)**	(0.028)**	(0.111)**	(0.167)**
Instability	0.841	0.841	-0.068	-0.068
	(0.106)**	(0.094)**	(0.044)	(0.098)
Uncertainty	-0.583	-0.583	-0.093	-0.093
	(0.110)**	(0.098)**	(0.154)	(0.181)
Age	-0.009	-0.009	0.009	0.009
	(0.004)*	(0.001)**	(0.009)	(0.017)
$(Age/10)^2$	0.051	0.051	-0.101	-0.101
	(0.028)	(0.012)**	(0.082)	(0.111)
Growth	-0.099	-0.099	0.052	0.052
	(0.116)	(0.104)	(0.028)	(0.082)
Size	-0.472	-0.472	5.068	5.068
	(0.966)	(0.887)	(6.283)	(25.906)
Constant	0.455	0.455	0.000	0.532
	(0.068)**	(0.038)**	(0.000)	(0.253)*
R^2	0.15	0.28	.	0.04

Notes: Zeta estimates from regression (3) based on distribution with minimum size selected by the LS criteria. Sample size 908 observations * p<0.05; ** p<0.01

Table 6 – Estimates from second stage regressions –Manufacturing– LM minimum size

	Pooled LS	DriscLS	DriscFE	FE
Capital Intensity	-0.043	-0.043	-0.011	-0.011
	(0.021)*	(0.018)*	(0.007)	(0.021)
Financial Depend	0.017	0.017	0.048	0.048
	(0.053)	(0.071)	(0.015)**	(0.043)
Human Capital	-1.105	-1.105	-0.263	-0.263
	(0.180)**	(0.048)**	(0.246)	(0.567)
Instability	0.996	0.996	0.123	0.123
	(0.199)**	(0.185)**	(0.084)	(0.183)
Uncertainty	-0.697	-0.697	-0.261	-0.261
	(0.247)**	(0.050)**	(0.145)	(0.332)
Age	0.008	0.008	0.013	0.013
	(0.010)	(0.005)	(0.021)	(0.035)
$(Age/10)^2$	-0.140	-0.140	-0.416	-0.416
	(0.054)**	(0.027)**	(0.227)	(0.238)
Growth	-0.024	-0.024	0.091	0.091
	(0.222)	(0.169)	(0.041)*	(0.180)
Size	-6.165	-6.165	27.293	27.293
	(4.614)	(1.462)**	(8.943)**	(58.525)
Constant	0.661	0.000	0.950	0.949
	(0.188)**	(0.000)	(0.094)**	(0.544)
R^2	0.32	0.64	.	0.08

Notes: Zeta estimates from regression (3) based on distribution with minimum size selected by the KS criteria. Sample size 347 observations * $p<0.05$; ** $p<0.01$

6 Conclusions

The regularity that the Pareto distribution may be used to model the upper tail of the firm size distribution has proven valid in a number of countries and industries. The Pareto distribution parameter is central in many models of firm competition and industry dynamics. Understanding firm differences from the industry characteristics help predict the relevance of structural transformation variables and granularity differences across industries.

We contribute to the literature by estimating the parameter of the Pareto distribution of the upper tail of the firm size distribution of all industries in the economy, moving beyond manufacturing data. We expand the analysis considering alternative methods to select the minimum firm size, under Clauset et al. (2009) goodness of fit criteria using different statistical distribution tests. Namely, we consider not only the Kolmogorov-Smirnov test proposed by Clauset et al. (2009) but also the Goerlich (2013) LM test. Both tests are used as criteria to select a minimum firm size where the Pareto would be valid. The actual parameter estimate follows the well-known *Rank*-1/2 regression method of Gabaix and Ibragimov (2008) and Gabaix (2009), also used in Halvarsson (2013).

Our estimates indicate that the Pareto distribution is valid for the tails of the majority of industries in Brazil. The tail itself tend to be high, with the estimated minimum size on the upper quarter centiles of the distributions. The Parameter estimates are mostly over unity, with median estimate of 1.2 under the KS minimum size criteria and 1.3 under the LM minimum size criteria. Interestingly, only in about half of the industry-year distribution we can accept the hypothesis that the distribution follows the special case of the Zipf's distribution, i.e., when the Pareto parameter equals one.

The results across methods are such that the LM test tend to select higher minimum sizes than the more widely used KS test. The associated shape parameter estimates under the LM test tend to be higher than the KS test, pointing to a result that is observed within industry, namely, higher minimum sizes are associated, in general, with higher shape parameter estimates. This empirical regularity can be observed in Clauset et al. (2009) simulations when the selected minimum is smaller or too large compared with the true minimum size.

The parameter estimates across industries are a means to understand the determinants of the firm size distribution. A central hypothesis in the literature posits that industries with higher specific capital, such as human capital, tend to have distributions with larger firms (Rossi-Hansberg and Wright, 2007), i.e., a thickening of the tail distribution with relatively smaller Pareto coefficients. While the literature uses proxies for human capital, such as physical capital intensity (as in Halvarsson, 2013), we are able to measure the schooling level of firms and industries. We also consider other factors that may influence the size distribution, such as financial restrictions, industry uncertainty, volatility, size and maturity.

In general, our estimates suggest that human capital is negatively associated with the shape parameter of the Pareto, as expected in the literature. Our estimates partially confirm the predictions on the uncertainty and instability variables, although the results are insignificant when considering short-term variations of these variables, i.e. in models with fixed effects. The results are generally the same when considering only manufacturing or all industries, suggesting that the known differences between, say, services and manufacturing such as size and human capital intensity do not seem to affect the main results.

Finally, the main results do not change significantly when considering the LM criterion for minimum size. This suggests that while the level of the Pareto parameter estimates may be larger in the LM criterion compared to the KS criterion, their dynamics are sufficiently close so that the main conclusions maintain across regression models using different estimates.

7 References

Angelini, P., Generale, A. (2008): On the Evolution of Firm Size Distributions, American Economic Review, 98: 426-438.

Axtell, R.L. (2001): Zipft Distribution of U.S. Firm Sizes, Science, 293: 1818-20.

Cabral, L., Mata, J. (2003): On the Evolution of the Firm Size Distribution: Facts and Theory, American Economic Review, 93: 1075-1090.

Clauset, A., Shalizi, C.R., Newman, M.E. (2009): Power-Law Distributions in Empirical Data, SIAM Review, 51: 661-703.

Coelho, D., Corseuil, C. H., Foguel, M.N. (2017): Employment Growth of Firms in the Brazilian Economy, in Business Dynamics and Productivity, OECD Publishing, Paris.

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

DeLoecker, J, Syverson, C. (2021): An Industrial Organization Perspective on Productivity. Handbook of Industrial Organization vol.4, chapter 3. New York: Elsevier.

Driscoll, J.C., Kraay, A.C. (1998): Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data, Review of Economics and Statistics, 549-560.

Ericson, R., Pakes, A. (1995): Markov-Perfect Industry Dynamics: A Framework for Empirical Work, Review of Economic Studies, 62: 53-82.

Fujiwara, Y., Di Guilmi, C., Ayoama, H., Gallegati, M., Souma, W. (2004): Do Pareto-Zipf and Gibrat laws hold true? An analysis with European Firms, Physica A: Statistical Mechanics and its Applications, 335: 197-216.

Gabaix, X. (2009): Power Laws in Economics and Finance, Annual Review of Economics, 1: 255-93.

Gabaix, X. (2011): The Granular Origins of Aggregate Fluctuations, Econometrica, 79: 733-332.

Gabaix, X., Ibragimov, R. (2011): Rank – ½: A Simple Way to Improve the OLS Estimation of Tail Exponents, Journal of Business and Economics Statistics, 29: 24-39.

Giovanni, J., Levchenko, A. (2012): Country Size, International Trade and Aggregate Fluctuations in Granular Economies, Journal of Political Economy, 120: 1083-1132.

Goerlich, F.J. (2013): A Simple and Efficient Test for the Pareto Law, Empirical Economics, 45: 1367-1381.

Goubert, C., Itskhoki, O. (2021): Granular Comparative Advantage, Journal of Political Economy, forthcoming.

Growiec, J., Pammolli, F., Riccaboni, M., Stanley, H. E (2008): On the Size Distribution of Business Firms, Economic Letters, 98: 207-212.

Halvarsson, D. (2013): Firm Dynamics: The Size and Growth Distribution of Firms, PhD Thesis, Department of Industrial Economics and Management, Royal Institute of Technology, Stockholm, Sweden.

Henrekson, M., Johansson, D. (1999): Institutional Effects on the Evolution of the Size Distribution of Firms, Small Business Economics, 12: 11-23.

Hoechle, D. (2007): Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence, Stata Journal, 7: 281-312.

Hopenhayn, H. (1992): Entry, Exit, and Firm Dynamics in Long Run Equilibrium, Econometrica, 60: 1127–1150.

Hopenhayn, H. (2014): Firms, Misallocation, and Aggregate Productivity: A Review, Annual Review of Economics, 2014, v.6.

Hsieh, C.T., Olken, B. (2014): The Missing "Missing Middle". Journal of Economic Perspectives, 28: 89-108.

Kondo, I, Lewis, L, Stella, A (2021) Heavy Tailed But Not Zipf: Firm and Establishment Size in the U.S, Census Bureau Discussion Paper CES 21-15.

Ioannides, Y., Overman, H., Rossi-Hansberg, E., Schmidheiny, K. (2008): The Effect of Information and Communication Technologies on Urban Structure, Economic Policy, 23: 201-

242.
Jovanovic, B. (1982): Selection and the Evolution of Industry, Econometrica, 50: 649-670
Lucas Jr., R. (1978): On the Size Distribution of Business Firms, Bell Journal of Economics, 9: 508-523.
Luttmer, E.G.T. (2007): Selection, Growth, and the Size Distribution of Firms, Quarterly Journal of Economics, 122: 1103-1114.
Na, J., Lee, J., Baek, C. (2017): Is the Service Sector Different in Size Heterogeneity?, Journal of Economic Interaction and Coordination 12: 95–120
Resende, M., Cardoso, V. (2022): Firm Size Distribution and Power Laws in Brazil: Some Empirical Evidence, mimeo.
Ribeiro, E. (2007). The Dynamics of Firm Size Distribution, Brazilian Review of Econometrics, 27: 199-223.
Rosen, K., Resnick, M. (1980): The Size Distribution of Cities: An Examination of the Pareto Law and Primacy, Journal of Urban Economics, 8: 165-186.
Rossi-Hansberg, E., Wright, M.L.J. (2007): Establishment Size Dynamics in the Aggregate Economy, American Economic Review, 97: 1639-1666.
Soo, K. (2005): Zipf's Law for Cities: A Cross-Country Investigation, Regional Science and Urban Economics, 35: 239-263.
Tybout, R.J. (2000): Manufacturing Firms in Developing Countries: How Well They Do, and Why?, Journal of Economic Literature 38, 11-44.