Physica A 541 (2020) 123188

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Multiply broken power-law densities as survival functions: An alternative to Pareto and lognormal fits

Roman Tomaschitz

Sechsschimmelgasse 1/21-22, A-1090 Vienna, Austria

ARTICLE INFO

Article history:
Received 27 June 2019
Received in revised form 22 September 2019
Available online 21 October 2019

Keywords:
Size distribution of firms
Multi-parameter distribution
Rank–size relation
Multiply broken power law
Varying power-law index
Nonlinear least-squares regression

ABSTRACT

Survival functions defined by multiply broken power-law densities are introduced to model firm-size data sets of four historical censuses reported in Montebruno et al., Physica A 523 (2019) 858. The survival functions (complementary cumulative distributions) are obtained by least-squares regression. The probability distributions are inferred from the analytic survival functions. The mean firm growth, the growth volatility and entropy evolution are calculated over a 30-year period covered by the censuses. A representation of the survival functions as single power-law density with varying exponent depending on firm size is derived. Index functions (i.e. rescaled logarithmic derivatives of survival functions) are used to demonstrate that neither Pareto power laws nor lognormal densities can accurately reproduce the tails of the firm-size distributions. Like the survival functions, the empirical rank–size relations of the four censuses also admit representation by broken power-law densities. A generating mechanism for empirically obtained broken power-law densities, based on the Fokker–Planck equation, is proposed as well.

1. Introduction

The purpose of this paper is to introduce multi-parameter distributions capable of reproducing firm-size data sets [1,2] over the full empirical size range. The distributions considered are broken power laws, composed of power-law segments with smooth analytic transitions, and their parameters are determined by least-squares regression. Empirical wealth [3–11] and city-size [12–22] distributions can be described in like manner, and more generally, the formalism developed is applicable to any data sets exhibiting approximately straight power-law segments in log–log plots, see the reviews [23–26] for a variety of possible applications.

Motivated by the observation that single power laws, lognormals and other standard distributions can rarely fit the complete empirical data sets, so that one has to resort to tail distributions (by truncating the data sets), there have recently been several suggestions on how to model firm- and city-size distributions over the full data range. One option is to partition the size range into intervals, choosing a suitable standard distribution (power law, lognormal, etc.) in each interval and joining them smoothly at the interval boundaries which become variable fit parameters [13,15,16]; see also Refs. [10,11] for piecewise joint distributions of capital income. Another method is to use linear combinations of a specific type of density, e.g. a finite series of lognormals [21] or generalizations thereof [19]. Distributions composed of a finite series of Hermite polynomials and a lognormal density have been tried in Ref. [17] for size distributions of US firms. A special case of broken power law, based on the inverse tangent, has been used in Ref. [12] for Australian and New

E-mail address: tom@geminga.org.

https://doi.org/10.1016/j.physa.2019.123188
0378-4371/

Zealandian city-size distributions. Due to their relative simplicity and adaptability, broken power laws of various shapes are also frequently used to model wideband spectra of astrophysical sources, cf. Ref. [27] and references therein.

We will analyze the firm-size data sets of historical population censuses of 1851–61–71–81 in England and Wales [1,2], in particular the time evolution of the firm-size distribution (firm size by number of employees) during this period. The survival functions and firm-size rank distributions of the four censuses can quite efficiently be represented by broken power-law densities, uniformly accurately over the full firm-size range. Before discussing specific examples of these densities in Sections 2 and 5, we outline here their general structure.

Multiply broken power-law densities are assembled as finite products of power-law factors,

$$\rho(x) = Ax^{\beta_0}\prod_{k=1}^{n}\left(1+(x/b_k)^{\beta_k/|\eta_k|}\right)^{\eta_k}, \tag{1.1}$$

defined on a positive semi-infinite interval. The amplitudes A, b_k are positive, and $b_k \ll b_{k+1}$. The exponents $\beta_{k\geq 1}$ are positive, β_0 is real (and subject to integrability constraints if the density is defined on $(0,\infty)$), and the exponents η_k can be positive or negative. Density (1.1) consists of $n+1$ approximate power-law segments, $\propto x^{\beta_0}$, $x^{\beta_0+\beta_1\mathrm{sign}(\eta_1)}$, …, $x^{\beta_0+\sum_{k=1}^{n}\beta_k\mathrm{sign}(\eta_k)}$, in the intervals $x \ll b_1$, $b_1 \ll x \ll b_2$, …, $b_n \ll x$, respectively. The amplitudes b_k define the break points between the power-law segments, and the exponents η_k determine the extend of the transitional regions. The densities (1.1) are analytic on the positive real axis and have asymptotic power-law decay for $x \gg b_n$ as indicated. Exponential and lognormal cutoffs of multiply broken power-law densities will be discussed in Section 2.

Pareto power laws can fit the upper tails (i.e. the intermediate size range) of various empirical distributions [23–26] quite well, but not always so much the lower tails which tend to decay faster than the power law suitable for the upper tail, see the examples depicted in Ref. [24], selected from a broad range of research areas. In Ref. [1], Pareto fits were performed to the firm-size tail distributions of the mentioned four Victorian censuses, and lognormal densities were considered as well, with partially inconclusive results as to whether the tails of the survival functions (covering firm sizes of above ten employees) are Pareto or lognormally distributed.

An elementary method to illustrate deviations from Pareto and lognormal distributions is provided by Index functions, which are defined by the rescaled logarithmic derivative of the respective regressed density (survival function, probability density, rank distribution, etc.), $\mathrm{Index}\,[\rho(x)] = -x\rho'(x)/\rho(x)$. Pareto distributions admit constant Index functions, and the Index functions of lognormal densities appear as straight lines with positive slope in linear-log representation. The logarithmic derivative of the broken power-law density (1.1) reads

$$\frac{\rho'(x)}{\rho(x)} = \frac{d(x)}{x}, \quad d(x) := \beta_0 + \sum_{k=1}^{n}\frac{\mathrm{sign}(\eta_k)\beta_k}{1+(b_k/x)^{\beta_k/|\eta_k|}}, \tag{1.2}$$

so that $\mathrm{Index}\,[\rho(x)] = -d(x)$. Thus, in an interval where this function is nearly constant (or a lin-log straight line with positive slope), the broken power-law density $\rho(x)$ in (1.1) can be approximated by a single power law (or a lognormal density), but in general this will not be the case and happens only over a limited range of firm sizes x. As for the mentioned historical firm-size censuses, we find, in all four cases, that neither lognormal nor Pareto densities can give accurate descriptions of the tails of the survival functions, especially the lower tails are largely off the mark in Pareto fits.

We will also discuss a generative mechanism for multiply broken power-law densities, that is, the practical design of a Fokker–Planck equation admitting a regressed density of type (1.1) or a derivative thereof as stationary limit distribution. The mean growth rate (drift coefficient) and growth rate volatility (diffusion coefficient) defining this equation are firm-size dependent empirical functions quantifying disproportionate firm growth.

This paper is organized as follows. In Section 2.1, broken power-law densities of type (1.1) are used to model the survival functions of the firm-size data sets of the 1851–61–71 Censuses in England and Wales [1,2] over the full empirical size range (see Figs. 1–3 in Section 2). In Section 2.2, we consider multiply broken power laws with Weibull and lognormal cutoffs at large firm size and give an example of a broken power law with lognormal cutoff by performing a least-squares fit to the data set of the survival function of the 1881 Census (see Fig. 4 in Section 2).

In Section 2.3, the least-squares regression of the firm-size survival functions of the 1851–61–71–81 Censuses is explained. A weighted χ^2 functional is employed to obtain a uniformly accurate fit of the broken power-law density defining the survival function of each census. The fitting parameters and two goodness-of-fit indicators, the determination coefficient and standard error of the fits, are estimated.

In Section 3, the discrete firm-size probability distributions (mass functions, PDFs) of the four Victorian censuses are calculated from the corresponding analytic survival functions (see Figs. 5–8 in Section 3), and an analytic approximation of the discrete distributions is derived. The time evolution of the firm-size probability distributions and survival functions of the censuses is studied, as well as the evolution of the mean firm growth, its volatility and the entropy evolution between 1851 and 81 (see Figs. 9, 10 in Section 3). The drift and diffusion coefficients of a Fokker–Planck equation that admits a prescribed empirical PDF as stationary solution are also discussed in this section.

A representation of the survival functions as single power-law density with varying exponent (dependent on firm size) is introduced in Section 4.1, and the variation of the power-law exponent is calculated for each census (see Fig. 11 in Section 4). Index functions, related to this varying exponent via the logarithmic derivative of the survival function, cf. (1.2), are discussed in Section 4.2. The purpose of Index functions is to locally quantify deviations of the survival functions

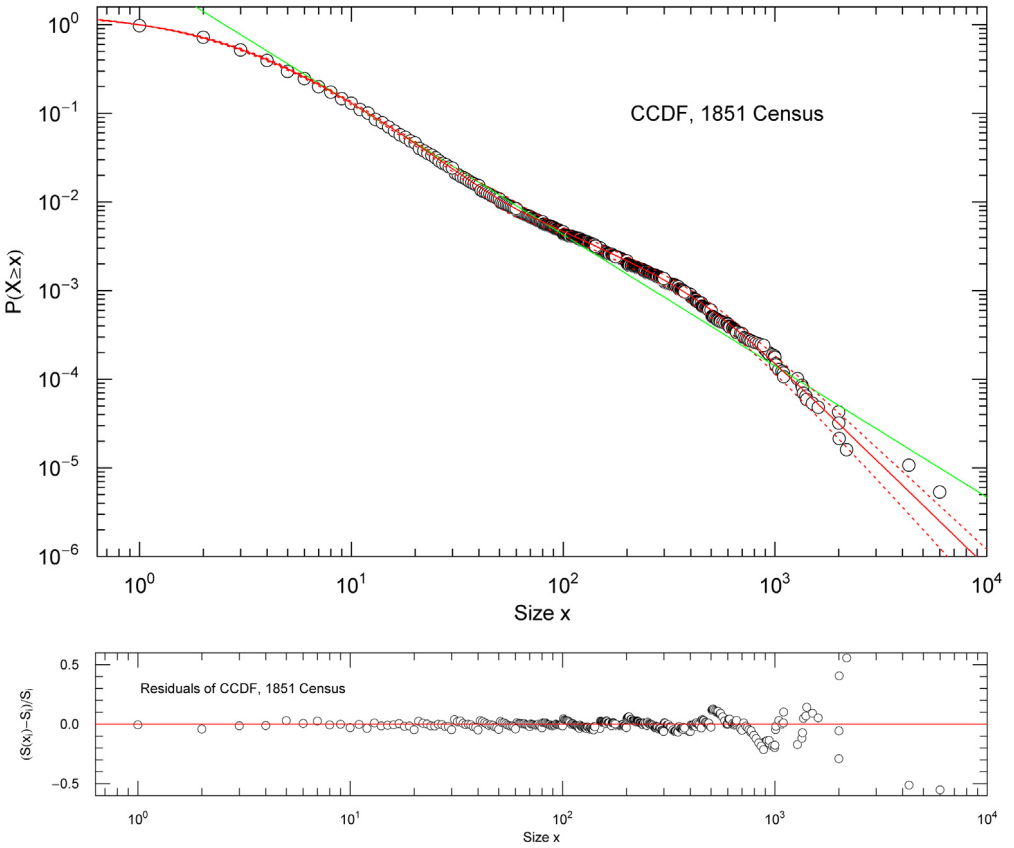

Fig. 1. Firm-size survival function (complementary cumulative distribution CCDF) of the 1851 Census. Data points from Ref. [2]. The solid red curve depicts the least-squares fit of the survival function $S(x) = P(X \geq x)$ in (2.1), which is a broken power law with fit parameters recorded in Table 1. The green straight line is a Pareto power-law fit of the tail distribution (for firm sizes $x \geq 10$) obtained in Ref. [1]. Evidently, the Pareto fit cannot reproduce the curvature of the empirical distribution in the $10^2 \leq x \leq 10^3$ interval nor the steepening slope in the subsequent decade. The dotted red curves define the 1σ error band. The residuals of the χ^2 fit are depicted in the lower panel, cf. Section 2.3.

from Pareto and lognormal densities; linear-log plots of the firm-size Index functions of the 1851–61–71–81 Censuses illustrate such deviations (see Fig. 12 in Section 4).

Further examples of multiply broken and lognormally cut power-law densities are studied in Section 5, where they are used to model the rank–size relations and rank Index functions of the four Victorian firm-size censuses (see Figs. 13–18 in Section 5). In Section 6, we present our conclusions.

2. Survival functions of historical firm-size distributions

2.1. Survival functions as broken power laws

We consider historical firm-size data sets from England and Wales [1,2], based on population censuses of 1851–61–71, and use a multiply broken power law of type (1.1) to model the survival function (complementary cumulative distribution, CCDF),

$$S(x) = P(X \geq x) = s(x)/s(1),$$
$$s(x) = \frac{1}{(1+(x/b_1)^{\beta_1/\eta_1})^{\eta_1}}(1+(x/b_2)^{\beta_2/\eta_2})^{\eta_2}\frac{1}{(1+(x/b_3)^{\beta_3/\eta_3})^{\eta_3}}. \tag{2.1}$$

The variable x labels firm size by number of employees. The survival functions of the 1851–61–71 Censuses depicted in Figs. 1–3 are obtained by least-squares fits of the broken power law (2.1) to the data sets, covering the complete empirical size range. (The regression is discussed in Section 2.3.) The fitting parameters $(b_k, \beta_k, \eta_k)_{k=1,2,3}$ are positive and $b_1 \ll b_2 \ll b_3$. The density (2.1) can be approximated by four power-law segments, $\propto 1$, $x^{-\beta_1}$, $x^{-\beta_1+\beta_2}$, $x^{-\beta_1+\beta_2-\beta_3}$, in the intervals $x \ll b_1$, $b_1 \ll x \ll b_2$, $b_2 \ll x \ll b_3$ and $b_3 \ll x$, respectively. The amplitudes b_k indicate the location of the transitions (break points), and the exponents η_k and ratios β_k/η_k determine the extent of the transitional regions,

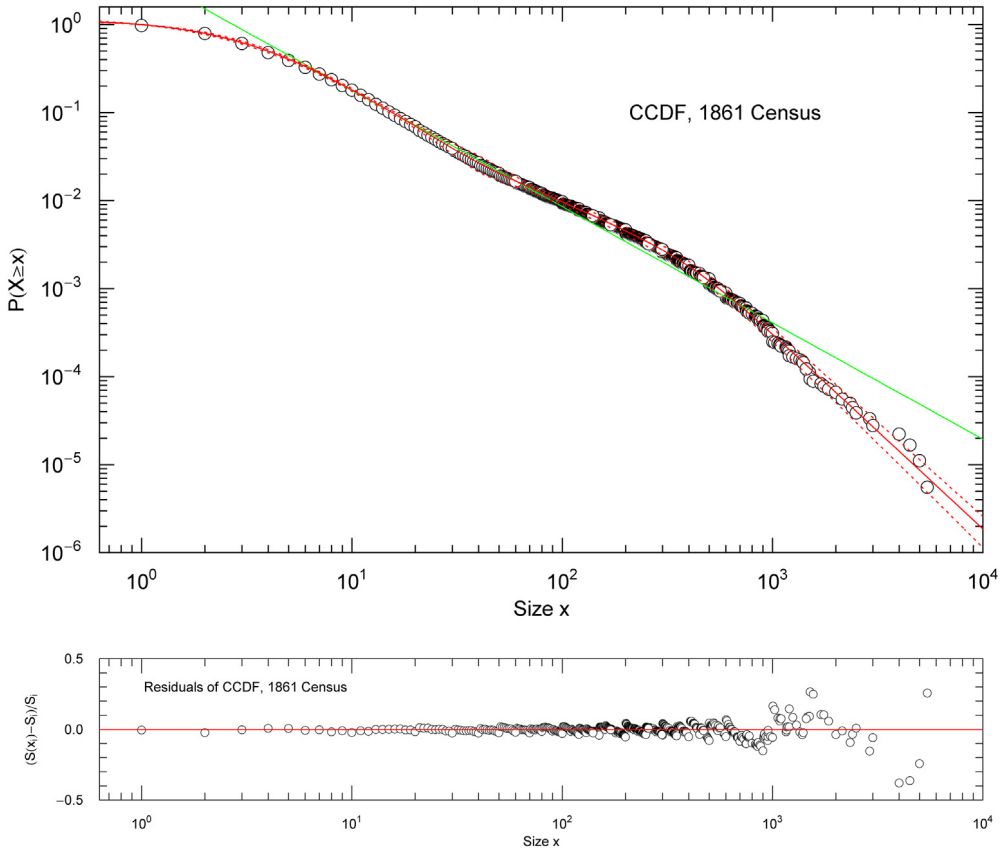

Fig. 2. Firm-size survival function (CCDF) of the 1861 Census. The caption to Fig. 1 applies. Data points from Ref. [2], the solid red curve depicts the survival function (2.1) with parameters in Table 1. For comparison, the green straight line is the Pareto tail fit (for firm size $x \geq 10$ employees) in Ref. [1]. The dotted red curves define the 1σ error band.

i.e. whether the transition from one power law to another is sudden or gradual. The enumerated power laws appear as approximately straight segments in the log–log plots depicted in Figs. 1–3, connected by curved transitions. The fitting parameters $(b_k, \beta_k, \eta_k)_{k=1,2,3}$ are listed in Table 1, the least-squares χ^2 functional is assembled in Section 2.3.

The PDFs of the censuses can be derived from the regressed survival functions, which is done in Section 3. Figs. 1–3 also show Pareto fits to the data sets derived in Ref. [1], which will be discussed in Section 4.2, together with the Index functions of the censuses. For the latter, we will need the logarithmic derivative of the survival function (2.1), $s'(x)/s(x) = d_S(x)/x$,

$$d_S(x) = -\frac{\beta_1}{1+(b_1/x)^{\beta_1/\eta_1}} + \frac{\beta_2}{1+(b_2/x)^{\beta_2/\eta_2}} - \frac{\beta_3}{1+(b_3/x)^{\beta_3/\eta_3}}. \tag{2.2}$$

This is just the derivative (1.2) specialized to density (2.1).

The historical data sets published in Refs. [1,2] also include firm-size data of the 1881 Census in England and Wales. The firm-size survival function of this census is studied in the next section, where we consider broken power-law densities with Weibull and lognormal cutoffs.

2.2. Multiply broken power laws with superexponential, subexponential and lognormal cutoffs

2.2.1. Asymptotic Weibull decay

An exponential cutoff of the broken power-law density $\rho(x)$ in (1.1) is generated by adding a Weibull factor to the product (1.1), $\rho_W(x) = \rho(x)\exp(-(x/a)^\alpha)$. The amplitude a and exponent α are positive, and $b_n \ll a$ (where b_n is the amplitude in the last factor of density (1.1)). The Weibull exponent $\alpha > 0$ defines compressed ($\alpha > 1$, superexponential) or stretched ($\alpha < 1$, subexponential) decay. Density $\rho_W(x)$ consists of $n+1$ approximate power-law segments enumerated after (1.1) (the last one being $x^{\beta_0+\sum_{k=1}^n \beta_k \mathrm{sign}(\eta_k)}$ in the interval $b_n \ll x \ll a$) and decays exponentially for $x \gg a$. The logarithmic derivative of $\rho_W(x)$ reads $\rho'_W(x)/\rho_W(x) = d_W(x)/x$, with $d_W(x) := d(x) - \alpha(x/a)^\alpha$ and $d(x)$ in (1.2). A superexponential cutoff could be useful in modeling the empirical firm-frequency distributions in Refs. [1,2]; here, we focus on firm-size distributions.

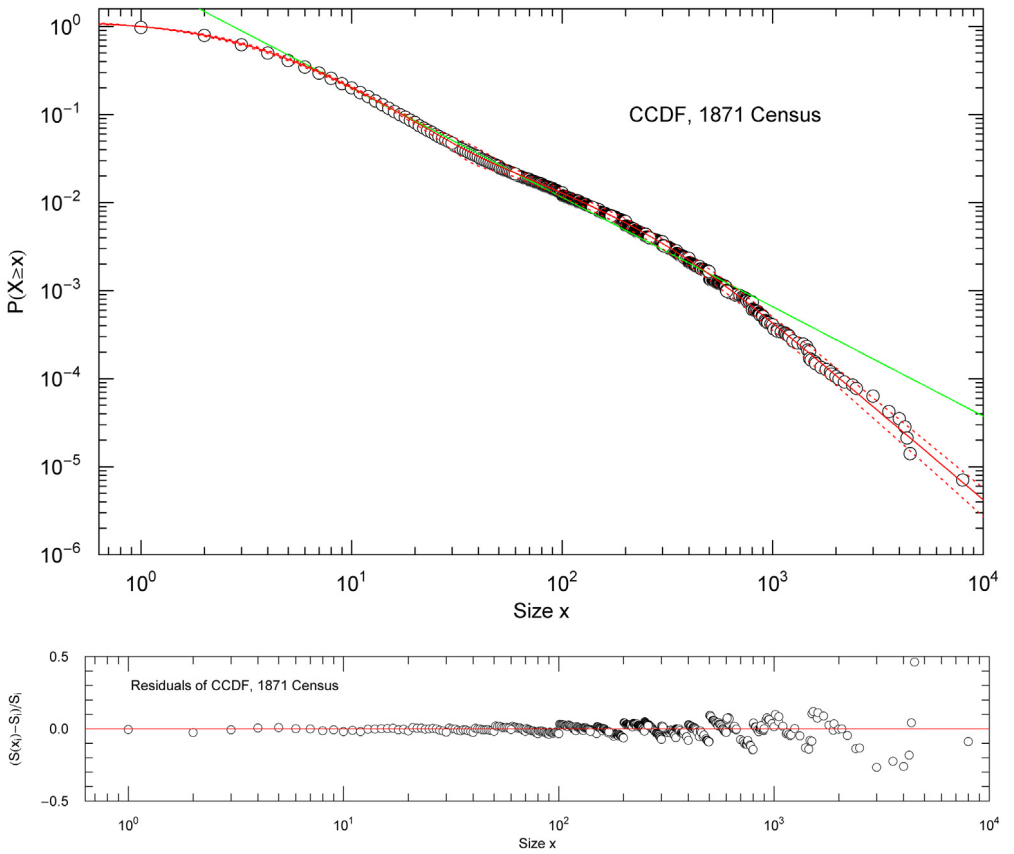

Fig. 3. Firm-size survival function (CCDF) of the 1871 Census. The caption to Fig. 1 applies. Data points from Ref. [2], the solid red curve depicts the survival function (2.1) with parameters in Table 1. The green straight line is the Pareto fit of the tail distribution (for $x \geq 10$) in Ref. [1]. The dotted red curves indicate the 1σ error band. Residuals are shown in the lower panel.

2.2.2. Lognormal decay

A cutoff weaker than stretched exponential (but still stronger than any power-law decay) is generated by adding an asymptotic lognormal factor to density $\rho(x)$ in (1.1),

$$\rho_{\mathrm{LN}}(x) = \rho(x)(1+(x/b_n)^{\beta_n/|\eta_n|})^{-\delta|\eta_n|\log x}. \tag{2.3}$$

The parameters (b_n, β_n, η_n) are defined after (1.1) and $\delta > 0$. This density consists of n power-law segments (the last one being $x^{\beta_0+\sum_{k=1}^{n-1}\beta_k \mathrm{sign}(\eta_k)}$ in the interval $b_{n-1} \ll x \ll b_n$, see after (1.1)) and decays lognormally, $\propto x^{\beta_0+\sum_{k=1}^{n}\beta_k \mathrm{sign}(\eta_k)-\delta\beta_n \log(x/b_n)}$, for $x \gg b_n$. A lognormal (or Weibull) cutoff factor added to the product (1.1) preserves the analyticity of the density on the positive real axis. By the way, density (1.1) has a branch cut along the negative real axis and is non-analytic at $x = 0$. The logarithmic derivative of the lognormally cut density $\rho_{\mathrm{LN}}(x)$ in (2.3) reads $\rho'_{\mathrm{LN}}(x)/\rho_{\mathrm{LN}}(x) = d_{\mathrm{LN}}(x)/x$, where

$$d_{\mathrm{LN}}(x) := d(x) - \delta\beta_n \frac{\log x}{1+(b_n/x)^{\beta_n/|\eta_n|}} - \delta\,|\eta_n| \log(1+(x/b_n)^{\beta_n/|\eta_n|}), \tag{2.4}$$

with $d(x)$ defined in (1.2).

2.2.3. A survival function with lognormal cutoff

As an example of a broken power-law density admitting lognormal decay, we fit historical firm-size data of the 1881 Census in England and Wales [2] with the survival function $S(x) = s(x)/s(1)$,

$$s(x) = \frac{1}{(1+(x/b_1)^{\beta_1/\eta_1})^{\eta_1}}(1+(x/b_2)^{\beta_2/\eta_2})^{\eta_2(1-\delta\log x)}, \tag{2.5}$$

which is a special case of density (2.3). The parameters $(b_k, \beta_k, \eta_k)_{k=1,2}$ and δ are positive. This survival function is composed of two power-law segments $\propto 1$, $x^{-\beta_1}$ in the intervals $x \ll b_1$, $b_1 \ll x \ll b_2$ and has an asymptotic

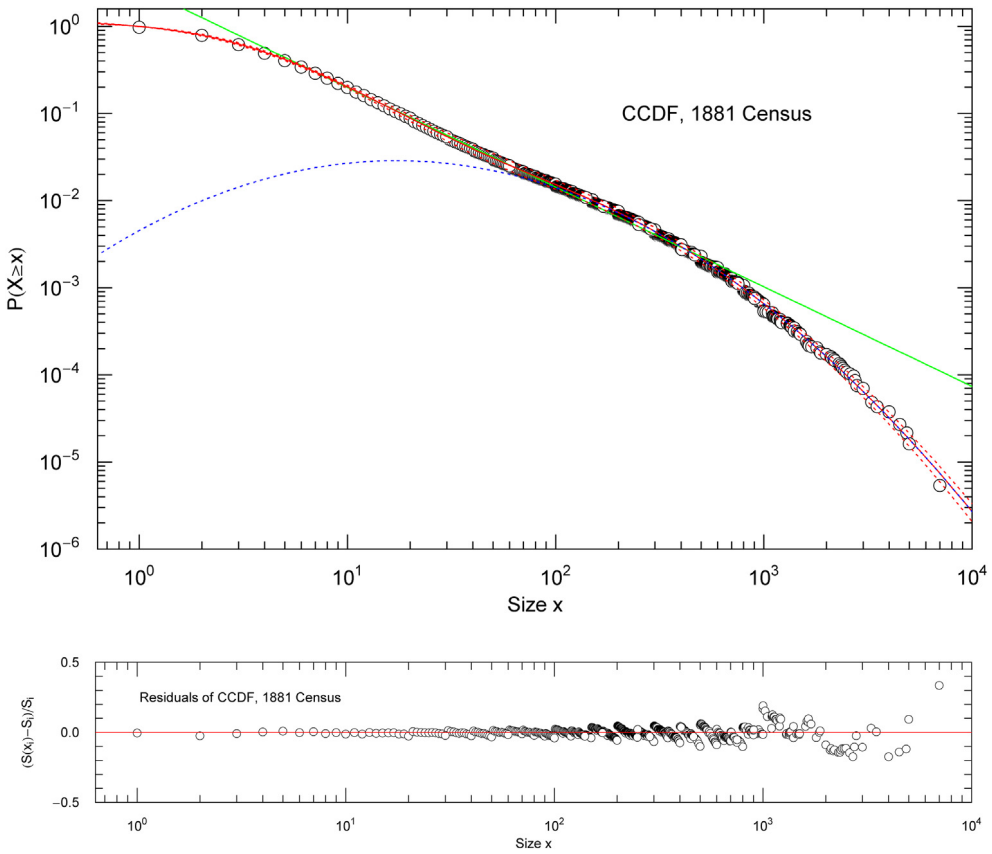

Fig. 4. Firm-size survival function (CCDF) of the 1881 Census. Data points from Ref. [2]; the solid red curve depicts the survival function (2.5), which is a broken power law with lognormal cutoff. The asymptotic limit of the survival function (2.5) is a lognormal density (dotted blue parabola). The fitting parameters are listed in Table 1. The green straight line is the Pareto tail fit (for $x \geq 10$) in Ref. [1]. The dotted red curves show the 1σ error band.

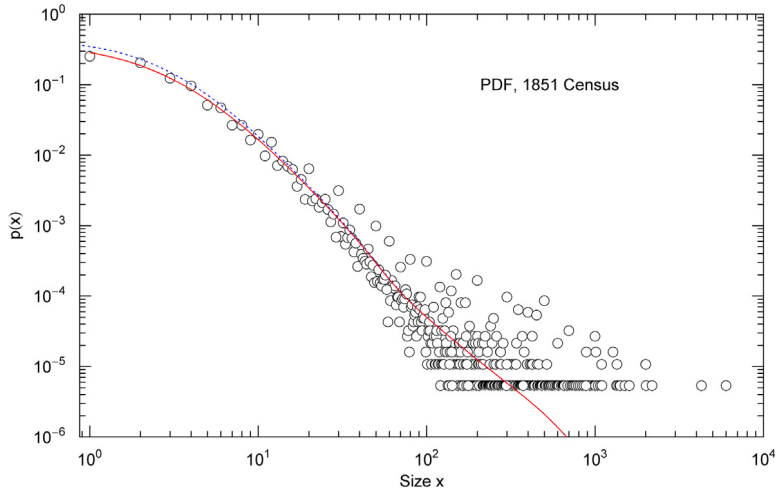

Fig. 5. Firm-size probability distribution function (PDF) of the 1851 Census. Data points from Ref. [2]. The solid red curve depicts the discrete PDF $p(x)$ in (3.1) defined on positive integers (firm size by number of employees), with a cubic polynomial interpolation for non-integer x. The dotted blue curve is the continuous PDF $\hat{p}(x)$ in (3.2), which is an efficient approximation of $p(x)$ for $x \geq 20$. Because of the large spread of the data points, the PDF is calculated from the survival function in Fig. 1 rather than by least-squares regression. The goodness-of-fit parameters are listed in Table 1.

Table 1
Parameters of the survival functions of Victorian firm-size censuses, based on data sets in Ref. [2], cf. Section 2. The fitting parameters $(b_k, \beta_k, \eta_k)_{k=1,2,3}$ of the survival function $S(x)$ in (2.1) of the 1851–61–71 Censuses and the parameters $(b_k, \beta_k, \eta_k)_{k=1,2}$ and δ of the lognormally cut survival function (2.5) of the 1881 Census were obtained by least-squares regression, cf. Figs. 1–4; the standard deviations are indicated as subscripts, and the 1σ error band is depicted in the figures. Also recorded are the number N of data points and the minimum of the least-squares functional χ^2 in (2.7). (Degrees of freedom: dof $= N - 9$ for the 1851–61–71 Censuses and dof $= N - 7$ for the 1881 Census.) The standard error SE and determination coefficient R^2 of the fits, cf. Section 2.3, are listed for the survival function S as well as for the corresponding PDF p in (3.1). Also indicated are the firm-size expectation value $\mu[X]$, standard deviation $\sigma[X]$ and entropy $S[p]$, which are calculated with the discrete PDF $p(n)$ in (3.1) truncated at the maximum firm size (MaxSize) of the respective census, cf. Section 3 and Fig. 10.

	1851 Census	1861 Census	1871 Census	1881 Census
b_1	$3.0912_{\pm 0.12}$	$3.2858_{\pm 0.12}$	$3.5485_{\pm 0.13}$	$3.0628\ _{\pm 0.11}$
β_1	$1.7489_{\pm 0.011}$	$1.5193_{\pm 0.010}$	$1.4853_{\pm 0.010}$	$1.3707\ _{\pm 0.0087}$
η_1	$1.4184_{\pm 0.12}$	$0.9751_{\pm 0.12}$	$1.0623_{\pm 0.12}$	$0.91437\ _{\pm 0.10}$
b_2	$53.396_{\pm 4.4}$	$40.335_{\pm 4.2}$	$37.290_{\pm 3.8}$	$59.036\ _{\pm 6.6}$
β_2	$0.66934_{\pm 0.031}$	$0.46341_{\pm 0.022}$	$0.48293_{\pm 0.022}$	$1.7369\ _{\pm 0.13}$
η_2	$0.11384_{\pm 0.26}$	$0.084516_{\pm 0.29}$	$0.079012_{\pm 0.32}$	$0.80025\ _{\pm 0.34}$
δ	–	–	–	$0.13159\ _{\pm 0.0019}$
b_3	$470.89_{\pm 51.}$	$414.91_{\pm 34.}$	$343.82_{\pm 31.}$	–
β_3	$1.2328_{\pm 0.17}$	$1.1636_{\pm 0.10}$	$1.0213_{\pm 0.088}$	–
η_3	$0.41610_{\pm 0.13}$	$0.36648_{\pm 0.11}$	$0.46288_{\pm 0.098}$	–
N	390	488	468	576
χ^2	2.01	1.12	1.14	1.10
SE[S]	1.57×10^{-3}	8.19×10^{-4}	9.21×10^{-4}	7.97×10^{-4}
$1 - R^2[S]$	4.28×10^{-4}	1.17×10^{-4}	1.35×10^{-4}	1.24×10^{-4}
SE[p]	2.04×10^{-3}	1.11×10^{-3}	1.17×10^{-3}	1.02×10^{-3}
$1 - R^2[p]$	1.14×10^{-2}	5.62×10^{-3}	6.42×10^{-3}	5.85×10^{-3}
MaxSize	6000	5444	8000	7000
$\mu[X]$	6.479	9.238	10.64	11.78
$\sigma[X]$	30.87	43.32	54.10	62.70
$S[p]$	2.471	2.782	2.877	2.895

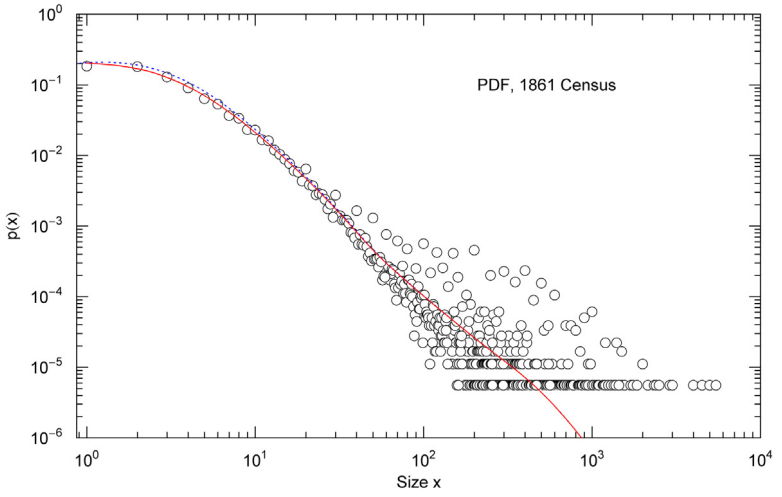

Fig. 6. Firm-size PDF of the 1861 Census. Data points from Ref. [2]. The solid red curve shows the discrete PDF $p(x)$ in (3.1) with a polynomial interpolation for non-integer x. The dotted blue curve is the analytic approximation $\hat{p}(x)$ in (3.2) applicable for large firm size. The PDF is calculated from the survival function (2.1) in Fig. 2 by way of (3.1).

lognormal cutoff $\propto x^{-\beta_1+\beta_2-\delta\beta_2\log(x/b_2)}$ for $x \gg b_2$. The least-squares fit of $S(x)$ in (2.5) is depicted in Fig. 4, the fitting and goodness-of-fit parameters are stated in Table 1.

The logarithmic derivative of the survival function (2.5) reads $s'(x)/s(x) = d_S(x)/x$, where

$$d_S(x) = -\frac{\beta_1}{1+(b_1/x)^{\beta_1/\eta_1}} + \frac{\beta_2}{1+(b_2/x)^{\beta_2/\eta_2}} - \delta\beta_2\frac{\log x}{1+(b_2/x)^{\beta_2/\eta_2}} - \delta\eta_2\log(1+(x/b_2)^{\beta_2/\eta_2}), \tag{2.6}$$

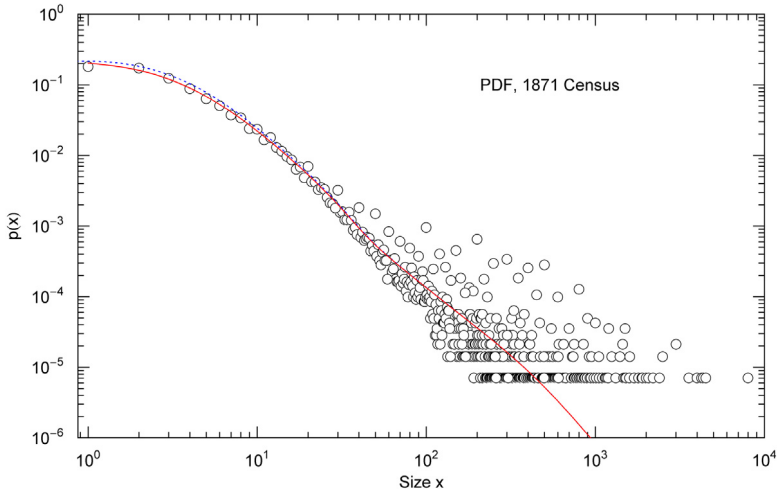

Fig. 7. Firm-size PDF of the 1871 Census. Data points from Ref. [2]. The solid red curve shows the discrete PDF $p(x)$ in (3.1) with a polynomial interpolation. The dotted blue curve is the continuous approximation $\hat{p}(x)$ of the PDF, cf. (3.2). The PDF is calculated from the survival function (2.1) plotted in Fig. 3.

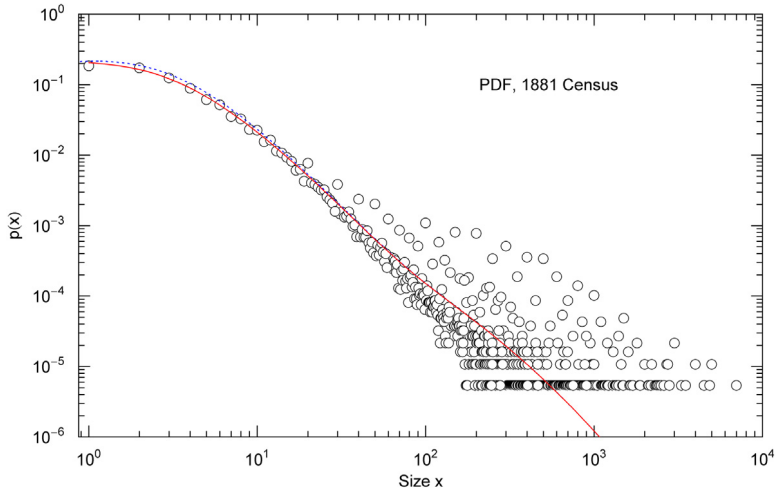

Fig. 8. Firm-size PDF of the 1881 Census. Data points from Ref. [2]. The solid red curve shows the discrete PDF $p(x)$ in (3.1) with a polynomial interpolation. The dotted blue curve is the continuous approximation $\hat{p}(x)$ of the PDF, cf. (3.2). The PDF is calculated from the survival function (2.5), which is a broken power law with lognormal cutoff, cf. Fig. 4.

obtained by specializing the logarithmic derivative in (2.4). The Index function of this census, $\mathrm{Index}\,[S(x)] = -d_S(x)$, and the Pareto approximation depicted in Fig. 4 will be discussed in Section 4.2, and the firm-size PDF will be derived in Section 3. In the next section, we explain the least-squares regression of multiply broken power-law densities, used for the survival functions (2.1) and (2.5) in Figs. 1–4.

2.3. Nonlinear least-squares regression of the survival function

The χ^2 functional of the least-squares fits to the data sets $(x_i, S_i)_{i=1,\ldots,N}$ depicted in Figs. 1–4 reads

$$\chi^2 = \sum_{i=1}^{N} \frac{1}{S_i^2}(S(x_i) - S_i)^2, \tag{2.7}$$

where x_i labels firm size (number of employees) in ascending order. The S_i define the empirical survival function $S_i = \sum_{k=i}^{N} p_k$, where p_i is the empirical probability attached to size x_i. The firm-size data sets $(x_i, p_i)_{i=1,\ldots,N}$ of the 1851–61–71–81 Censuses are taken from Ref. [2]; the x_i are positive integers and the probabilities are normalized, $\sum_{i=1}^{N} p_i = 1$.

The least-squares fits in Figs. 1–4 are performed with the survival function $S(x;(b_k,\beta_k,\eta_k)_{k=1,2,3})$ in (2.1) for the 1851–61–71 Censuses and with $S(x;(b_k,\beta_k,\eta_k)_{k=1,2},\delta)$ in (2.5) for the 1881 Census. We have here explicitly indicated the fitting parameters, cf. Table 1. The residuals determining the χ^2 functional (2.7) are weighted with the empirical S_i. The weight factors are crucial to obtain a uniformly accurate fit of the data sets $(x_i,S_i)_{i=1,\ldots,N}$ over the firm-size range covered by data points, $1\le x_i<10^4$, since the empirical probabilities p_i as well as the complementary cumulants S_i decay rapidly, varying over five logarithmic decades, cf. Figs. 1–4. That is, without the weight factors, the residuals in the fourth and fifth decade would give a negligible contribution to the χ^2 functional, as compared with residuals in the first two decades. The rescaled residuals $(S(x_i)-S_i)/S_i$ constituting series (2.7) at the minimum of χ^2 are depicted in the lower panels of Figs. 1–4.

The standard error of the fits, $\mathrm{SE}[S]=(\sum_{i=1}^N(S(x_i)-S_i)^2/N)^{1/2}$ and the determination coefficient $R^2[S]$ are listed in Table 1. For the latter, we adopt definition (1) of Ref. [28], $R^2[S]=1-\sum_{i=1}^N(S(x_i)-S_i)^2/(N\sigma^2)$, with sample variance $\sigma^2=\sum_{i=1}^N(S_i-\overline{S})^2/N$, $\overline{S}=\sum_{i=1}^N S_i/N$. Other definitions of the determination coefficient enumerated in Ref. [28] are not equivalent, as the regression based on $S(x)$ in (2.1) or (2.5) is nonlinear.

The stated goodness-of-fit parameters are defined as global averages over the complete distribution and, therefore, do not always reflect the local quality of the fit, especially in cases where the data points vary over several orders of magnitude as in Figs. 1–4. For instance, the data points of the lower tails in Figs. 2–4 differ by a factor of two or more from the Pareto fits. These large errors do not significantly affect the goodness-of-fit parameters (of the Pareto fit indicated in the figures by the green straight line), since the probabilities attached to data points of the lower tail are very small (and thus negligible in averages) as compared with probabilities in the upper tail where the Pareto straight-line fit is a good approximation. A local goodness-of-fit criterion is the varying width of the error band, cf., e.g., Ref. [29], indicated by the dotted red curves in Figs. 1–4. (Cross-correlations of the fitting parameters are not included in the depicted error bands; the inverse covariance matrix is diagonal, composed of second-order derivatives of $\chi^2/2$ with respect to the fitting parameters.)

In logarithmic coordinates (i.e. log–log representation), single power laws are linear, and lognormals are second-order polynomials defining parabolas as depicted in Fig. 4. They can be fitted by using linear least-squares regression. In this case, the minimum of the χ^2 functional is calculated by solving a linear system obtained by equating the derivatives of χ^2 with respect to the fitting parameters to zero. In contrast, the parametrization of the multiply broken power laws (1.1) is nonlinear and remains so in logarithmic coordinates. The minimization of the χ^2 functional (2.7) amounts to solving a nonlinear system, which requires a good initial guess of the fit parameters. To obtain an initial guess of the multi-parameter density $\rho(x)$ in (1.1), we make use of the condition $b_k\ll b_{k+1}$ on the amplitudes defining the factors of the product in (1.1). Due to this condition, the factor defined by amplitude b_{k+1} does not significantly affect the density $\rho(x)$ in the range $x\le b_k$, so that an initial guess of $\rho(x)$ can be found by successively adding factors (to cover an ever increasing size range) and visually fitting the factors (which are asymptotic power laws, i.e. log–log straight lines) one by one. This usually leads to a good initial guess of the fitting parameters, which is needed when minimizing a nonlinear multi-parameter χ^2.

3. Firm-size probability distributions inferred from regressed survival functions

Once the survival function has been found by a least-squares fit, the discrete probability distribution (PDF, mass function) is obtained as

$$p(i)=S(i)-S(i+1), \tag{3.1}$$

defined on positive integers i labeling firm size, and $\sum_{i=1}^\infty p(i)=1$ since $S(1)=1$, cf. (2.1) and (2.5). In Figs. 5–8, the firm-size PDFs of the 1851–61–71–81 Censuses are depicted as solid red curves, with a cubic polynomial interpolation of the discrete probabilities $p(i)$. The analytic survival functions $S(x)$ of the censuses are stated in (2.1) (for 1851–61–71) and in (2.5) (for 1881), with parameters in Table 1.

The continuous PDF $\hat{p}(x)$ defined by the survival function via $S(x)=\int_x^\infty\hat{p}(x)\mathrm{d}x$ on the interval $[1,\infty)$ reads

$$\hat{p}(x)=-S'(x)=-\frac{S(x)}{x}d_S(x), \tag{3.2}$$

with $d_S(x)$ in (2.2) for the 1851–61–71 Censuses and in (2.6) for the 1881 Census. This PDF is depicted in Figs. 5–8 as dotted blue curve. $\hat{p}(x)$ is normalized and gives a continuum approximation to the probability mass function $p(i)$ in (3.1), provided that the firm sizes are sufficiently large. (Since $S^{(n)}(x)/S(x)=\mathrm{O}(1/x^n)$, the difference $p(i)$ in (3.1) is close to the derivative $-S'(i)$ for large i.) The density $\hat{p}(x)$ accurately reproduces the discrete $p(i)$ for firm sizes $i,x\ge 20$.

Because of the large spread of the data points (x_i,p_i) in Figs. 5–8, the firm-size PDF $p(i)$ is not inferred by data regression but calculated from the fitted survival function via (3.1). The goodness-of-fit indicators $\mathrm{SE}[p]$ and $R^2[p]$ in Table 1 are calculated as outlined in Section 2.3 for the survival function, with the replacements $S(x_i)\to p(x_i)$ and $S_i\to p_i$.

In Fig. 9, we compare the firm-size PDFs (3.1) of the four Victorian censuses as well as the corresponding survival functions. Apparently, a stationary state has not been attained. Fig. 10 shows the time evolution of the mean firm growth in the period $1851-81$. Expectation value and root mean square (RMS) are calculated with the mass function (3.1); the series $\mu=\sum_{n=1}^{\mathrm{MaxSize}}np(n)$ is truncated at the largest firm size of the respective census, cf. Table 1, which

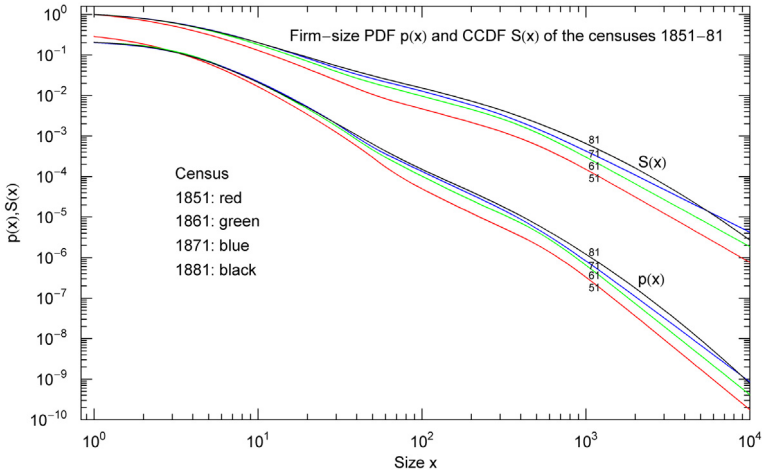

Fig. 9. Time evolution of the firm-size probability distribution $p(x)$ and survival function $S(x)$ of the 1851–61–71–81 Censuses. The survival function of each census is inferred from the least-squares fits in Figs. 1–4. A cubic polynomial interpolation is used to depict the corresponding probability mass functions (3.1), see also Figs. 5–8. Even though the distributions look similar, a steady state is not yet attained. In particular, the lower tails of the survival functions (2.1) of the 1851–61–71 Censuses are power laws, whereas the survival function (2.5) of the 1881 Census has a lognormal lower tail, cf. Figs. 1–4.

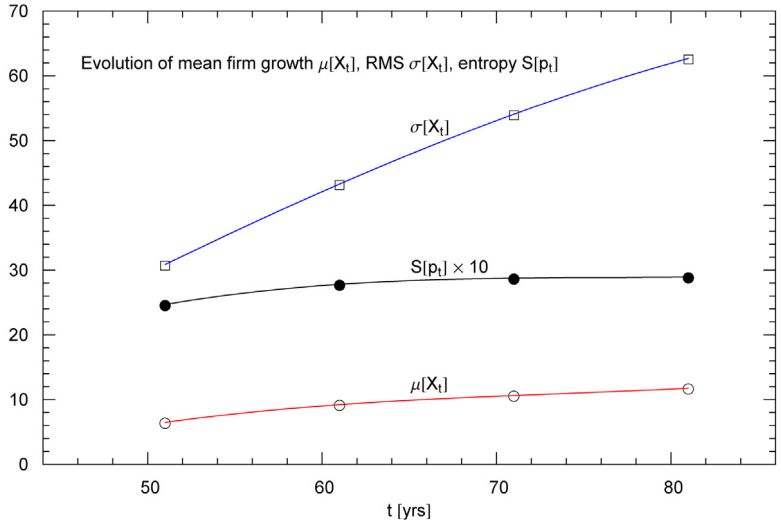

Fig. 10. Time evolution of the firm-size expectation value $\mu[X_t]$, standard deviation $\sigma[X_t]$ and entropy $S[p_t]$ in the interval 1851–81, cf. Section 3. Data points taken from Table 1. In the case of $\mu[X_t]$ and $\sigma[X_t]$, the ordinate labels firm size (by number of employees), whereas $S[p_t]$ is dimensionless. Depicted is a cubic polynomial interpolation of the data points.

does not significantly affect the normalization: $10^{-6} < 1 - \sum_{n=1}^{\text{MaxSize}} p(n) < 10^{-5}$ for all four censuses. The entropy, $S[p] = -\sum_{n=1}^{\text{MaxSize}} p(n)\log p(n)$, is also listed in Table 1 for each census. The entropy growth rate per decade (which is 0.311, 0.095 and 0.018, respectively, for the three decades covered by the censuses) rapidly decreases. This suggests that the entropy is close to a stationary plateau value, cf. Fig. 10, a constant entropy being a necessary condition for a steady state. However, an approximately constant entropy does not suffice to conclude that a stationary state has been reached. The entropy function is nearly unaffected by the lower tails, as the probabilities in this region are very small, giving a negligible contribution to the above stated average defining entropy. As is apparent from the time evolution of the PDF and survival function in Fig. 9, a steady state is not yet attained within the 30-year interval covered by the censuses; see also the evolution of the rank–size relation in Fig. 17 (Section 5).

Finally we sketch a generating stochastic mechanism for the analytic PDF $\hat{p}(x)$ in (3.2), using a one-dimensional Fokker–Planck equation,

$$p_{,t}(x,t) = -(\mu(x)xp(x,t))_{,x} + (1/2)(\sigma^2(x)x^2p(x,t))_{,x,x}, \tag{3.3}$$

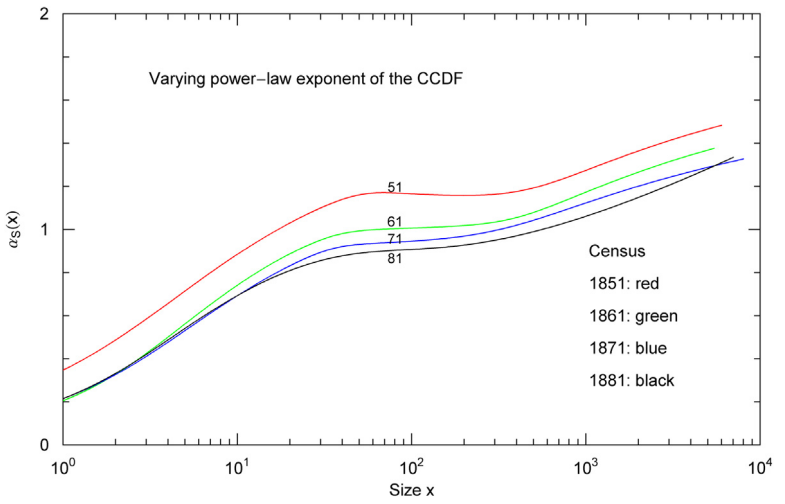

Fig. 11. Varying power-law exponent of the firm-size survival functions of the 1851–61–71–81 Censuses, cf. Section 4.1. The exponent $\alpha_S(x) = -\log S(x)/\log x$ shows substantial variation over the empirical firm-size range. The curves are plotted up to the largest firm size in the respective data set (labeled MaxSize in Table 1). The survival functions $S(x)$ are depicted in Figs. 1–4.

where $\mu(x)$ and $\sigma^2(x)$ denote the firm-size dependent mean and variance of the firm growth rate, cf., e.g., Refs. [30–32]. The associated stochastic equation reads $\mathrm{d}X_t/X_t = \mu(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t$ with standard Wiener process W_t. Substitution of the stationary limit $\hat{p}(x)$ of the PDF $p(x,t)$ into (3.3) gives

$$(\sigma^2(x))' + (\frac{2}{x} + \frac{\hat{p}'(x)}{\hat{p}(x)})\sigma^2(x) = \frac{2}{x}(\mu(x) + \frac{c}{x\hat{p}(x)}), \tag{3.4}$$

with integration constant c. We assume the PDF of the 1881 Census to be close to the stationary limit and use this PDF for $\hat{p}(x)$ in (3.4), although the time evolution depicted in Fig. 9 does not provide much evidence for this, as mentioned above.

Solving (3.4) for $\sigma^2(x)$ gives

$$\sigma^2(x) = \frac{\sigma_0^2 x_0^2 \hat{p}_0}{x^2\hat{p}(x)} + \frac{2}{x^2\hat{p}(x)}(\int_{x_0}^{x} x\mu(x)\hat{p}(x)\mathrm{d}x + c(x - x_0)), \tag{3.5}$$

with integration constant $\sigma_0^2 = \sigma^2(x_0)$ and $\hat{p}_0 = \hat{p}(x_0)$. Since $\hat{p}(x)$ in (3.2) is defined and normalized on the interval $[1,\infty)$, it is convenient to put $x_0 = 1$. Assuming $\mu(x)\hat{p}(x)$ and $\sigma^2(x)\hat{p}(x)$ to be of order $\mathrm{O}(x^{-1-\varepsilon})$, $\varepsilon > 0$, for large firm size, which requires $c = 0$ in (3.4) and (3.5), we obtain $\langle\sigma^2(x)\rangle = \sigma_0^2 x_0\hat{p}_0 + 2\langle\mu(x)\rangle$, where $\langle\sigma^2(x)\rangle := \int_{x_0}^{\infty}\sigma^2(x)\hat{p}(x)\mathrm{d}x$ and analogously for $\langle\mu(x)\rangle$. (This follows from (3.5) via integration by parts.)

An empirically inferred growth rate volatility $\sigma(x)$ substituted into (3.4) (with $c = 0$) determines the mean $\mu(x)$. Conversely, a mean growth rate $\mu(x)$ and integration constant σ_0^2 can be specified in (3.5), which determines the variance. (The growth rate can become negative but is constrained by the positivity of the variance.) Empirical growth rate estimates are not available for the four Victorian censuses. Typically, in fact quite universally, the growth rate volatility turns out to be a decaying power law $\sigma(x) \propto x^{-\delta}$ with exponent around 0.2. An exponent of $\delta = 0.16$ was inferred for US manufacturing firms [33], where the power law is applicable for firm sizes above ten employees; see also Refs. [34,35] for similar recent estimates of this exponent based on Chinese and Iberian firms. The above stated stochastic process is unambiguously defined once an analytic stationary PDF $\hat{p}(x)$ and a growth rate volatility $\sigma(x)$ are specified, both of which can be determined empirically.

A constant mean growth rate and variance in (3.3) defines geometric Brownian motion, reflecting Gibrat's law of size-independent growth, cf., e.g., Refs. [36–38]. In this case, the Fokker–Planck equation is exactly solvable, admitting a lognormal fundamental solution which does not converge to a stationary limit, cf., e.g., Ref. [39]. A size-dependent mean growth rate and growth rate variance indicate deviations from Gibrat's law and lognormality. The PDFs in Figs. 5–8 are decidedly not lognormal over the full empirical size range, their survival functions being defined in (2.1) and (2.5). (Lognormals appear as parabolas in log–log coordinates, see Fig. 4.)

In Section 4, we will discuss (by making use of Index functions) to which extend (i.e. in which firm-size ranges) the survival functions of the 1851–61–71–81 Censuses in Figs. 1–4 can be approximated by single power laws and lognormals. Stochastic mechanisms generating stable power-law tails whilst maintaining a constant growth rate are studied in Refs. [30–32], with emphasis on the Zipf power law, $p(x) \propto 1/x^2$, leading to an inversely proportional survival function. The focus of this paper is on modeling complete distributions over the full empirical size range, rather than on tails.

4. Index functions

4.1. Varying power-law index of the survival function

An arbitrary density $S(x)$ can be written as power law with varying exponent and constant amplitude, $S(x) = Ax^{-\alpha_S(x)}$. The exponent, $\alpha_S(x) = -\log(S(x)/A)/\log x$, depends on the amplitude which can be freely chosen. For instance, the power-law representation of a lognormal density is $n(x) = n(1)x^{-(\alpha+\beta\log x)}$ (with real α and positive β), the exponent appearing as straight line in linear-log plots. The amplitude is chosen so that the exponent is non-singular at $x = 1$.

The normalization of the survival functions (2.1) and (2.5) is $S(1) = 1$, which suggests to put $A = 1$, $S(x) = x^{-\alpha_S(x)}$. Any other choice of amplitude would make the exponent diverge at $x = 1$. The varying exponent thus reads

$$\alpha_S(x) = -\frac{\log S(x)}{\log x}, \tag{4.1}$$

and $\alpha_S(1) = -S'(1)$, obtained by expanding the ratio (4.1) at $x = 1$ in linear order. Fig. 11 depicts the varying exponent $\alpha_S(x)$ of the firm-size survival functions of the 1851–61–71–81 Censuses, calculated with $S(x)$ in (2.1) or (2.5) and parameters in Table 1. The varying exponents show similar features within the empirical size range: an increasing slope in the first two decades, followed by an approximately flat plateau and a subsequent moderate increase.

4.2. Index functions quantifying deviations from Pareto and lognormal distributions

The Index functional $\text{Index}\,[S(x)] = -xS'(x)/S(x)$ in lin-log representation provides an elementary method to visualize the deviation of an arbitrary density from a power-law or lognormal density [30,31]. If the survival function is a power law, $S(x) \propto x^{-\alpha}$, then $\text{Index}\,[S(x)]$ is constant and coincides with the exponent α. In the case of a lognormal density, $S(x) \propto x^{-\alpha-\beta\log x}$, $\beta > 0$, we find $\text{Index}\,[S] = \alpha + 2\beta\log x$, which is a straight line in lin-log plots. Conversely, if $\text{Index}\,[S]$ is a lin-log straight line with zero or positive slope, then $S(x)$ is a power-law or lognormal density. Thus, the deviation of $\text{Index}\,[S]$ from a constant or straight line in lin-log plots quantifies the extent to which $S(x)$ differs from a power-law or lognormal density.

The Index functional is related to the varying exponent (4.1) of the survival function by $\text{Index}\,[S] = \alpha_S(x) + x\alpha_S'(x)\log x$. The solution $\alpha_S(x)$ of this equation has been discussed in Section 4.1, the constant amplitude being the integration constant. In contrast to the exponent $\alpha_S(x)$, which depends on the choice of the integration constant, $\text{Index}\,[S]$ is a measure of the variation of the power-law index unaffected by the choice of amplitude and determined by the survival function alone. The PDF (3.2) can be written as product $\hat{p}(x) = S(x)\text{Index}\,[S(x)]/x$.

The firm-size Index functions of the 1851–61–71–81 Censuses are defined by the survival functions (2.1) and (2.5) and read $\text{Index}\,[S(x)] = -d_S(x)$, with $d_S(x)$ in (2.2) for 1851–61–71 and $d_S(x)$ in (2.6) for 1881. Fig. 12 depicts lin-log plots of these Index functions, covering the complete empirical data range up to the largest firm-size of each census (listed as MaxSize in Table 1). For comparison, the Index functions of the Pareto straight-line approximations of the survival functions in Figs. 1–4 are depicted as horizontal dotted lines in Fig. 12.

For large firm size $x \geq 2000$, the Index functions of the 1851–61–71 Censuses attain constant limit values, since the asymptotic limit of their survival function (2.1) is a power law. For firm sizes $x < 2000$, the Index functions largely deviate from their asymptotic values. In contrast, since the asymptotic limit of the survival function (2.5) of the 1881 Census is lognormal, the firm-size Index function of this census converges to a straight line with positive slope (dotted in Fig. 12) for large firm size (which means $x \geq 200$ in this case). For firm sizes in the upper tail, $10 \leq x \leq 200$, this Index function substantially deviates from its asymptotic straight-line limit.

In the firm-size range $10 \leq x \leq 10^4$ defining the tail distributions in Ref. [1], the Index functions cannot be approximated by a constant (indicating a power-law CCDF) or straight line with positive slope (indicating a lognormal CCDF). Therefore, neither a Pareto nor a lognormal density is suitable to describe the firm-size tail distributions of the four censuses. This is also evident from the Pareto approximations in Ref. [1] (for firm sizes $x \geq 10$), which are indicated by the log–log straight lines in Figs. 1–4. These Pareto fits fail to model the curvature of the survival functions in the $10^2 \leq x \leq 10^3$ interval, which is clearly visible in Figs. 1–4 (since there is little scatter in the data sets defining the curves), let alone the lower tails in the subsequent decade.

5. Rank–size relations as broken power laws

In Figs. 13–16, broken power laws are used to model the firm-size rank distributions of the 1851–61–71–81 Censuses in England and Wales, based on data sets in Refs. [1,2]. (The corresponding survival functions and PDFs of the censuses are discussed in Sections 2 and 3, cf. Figs. 1–8.) The firm-size rank $R(x)$ is ordered by decreasing firm size x, the largest firm being ranked one. For the rank distributions of 1851–61–71, we employ a power law with one break, a special case of density (1.1),

$$R(x) = Ax^{\beta_0}\frac{1}{(1+(x/b_1)^{\beta_1/\eta_1})^{\eta_1}}. \tag{5.1}$$

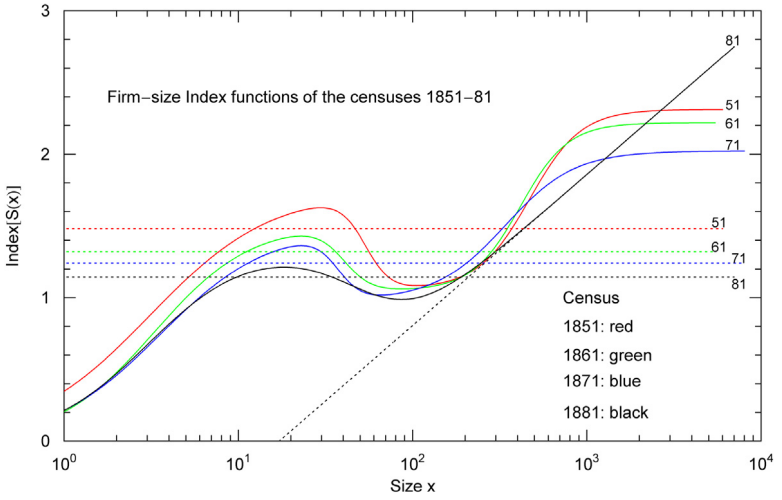

Fig. 12. Firm-size Index functions of the 1851–61–71–81 Censuses. Index $[S] = -xS'(x)/S(x)$ quantifies the deviation of the survival function S from a Pareto or lognormal density, cf. Section 4.2. (The firm-size survival functions $S(x) = P(X \geq x)$ are depicted in Figs. 1–4.) The Pareto power laws derived in Ref. [1] for the survival functions admit constant Index functions (dotted horizontal straight lines). The curves are plotted up to the largest firm size in the respective data set, cf. Table 1. Asymptotically, for large firm size, the survival functions of the 1851–61–71 Censuses are power laws, cf. after (2.1), which is indicated by the straight horizontal segments of the Index curves emerging in the last decade of the firm-size range. In contrast, the survival function of the 1881 Census converges, for large firm size, to a lognormal density, cf. after (2.5) and Fig. 4, whose Index function is a straight line with positive slope (black, dotted).

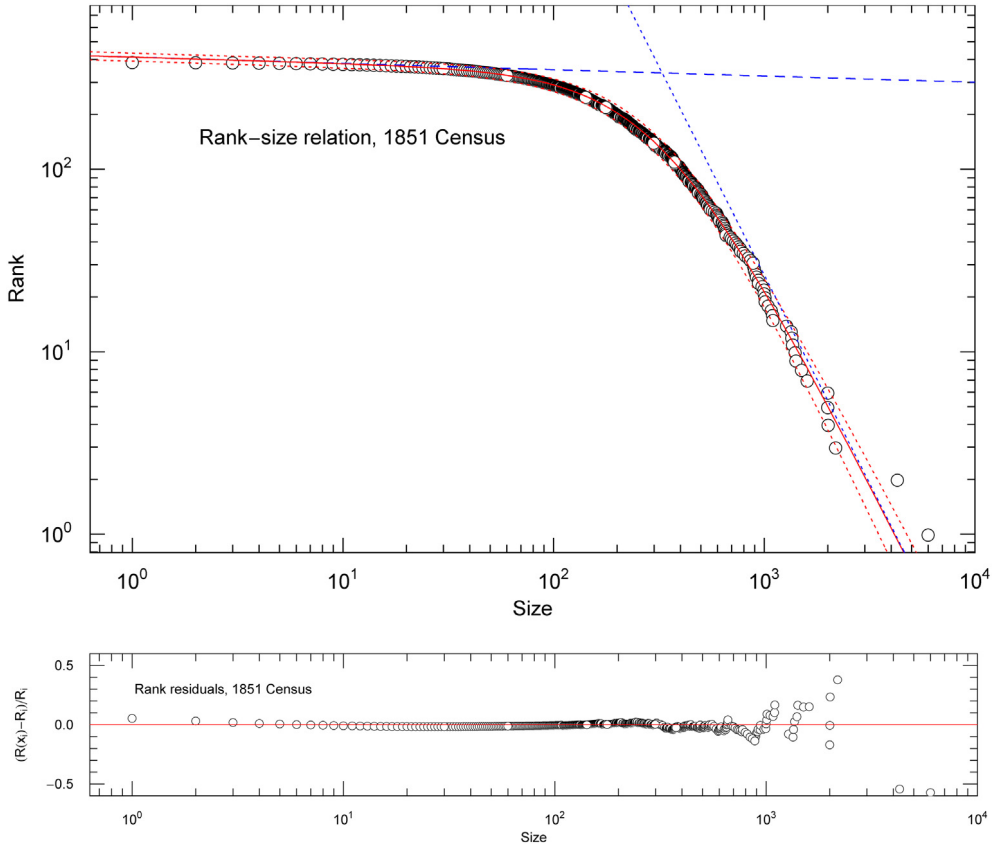

Fig. 13. Rank distribution of the 1851 Census, cf. Section 5. The rank is ordered by decreasing firm size. Data points from Ref. [2]. The least-squares fit (solid red curve) is performed with the broken power law $R(x)$ in (5.1); the fit parameters are listed in Table 2. The dashed and dotted blue straight lines show the asymptotic limits of $R(x)$. The dotted red curves indicate the 1σ error band. The residuals are depicted in the lower panel.

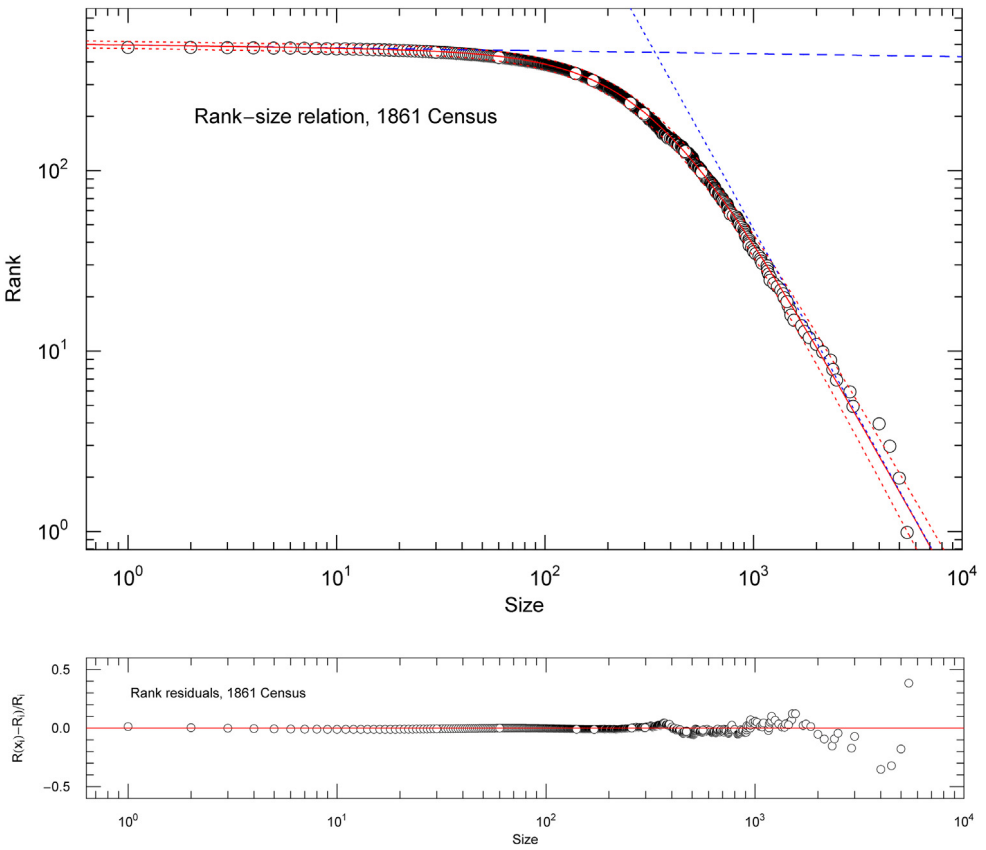

Fig. 14. Rank–size relation of the 1861 Census. Data points from Ref. [2], the caption of Fig. 13 applies. The solid red curve shows the regressed broken power law $R(x)$ in (5.1) with parameters in Table 2. The dashed and dotted blue straight lines are the asymptotic power-law limits. The dotted red curves depict the 1σ error band.

The logarithmic derivative thereof reads, cf. (1.2),

$$\frac{R'(x)}{R(x)} = \frac{d_R(x)}{x}, \quad d_R(x) = \beta_0 - \frac{\beta_1}{1+(b_1/x)^{\beta_1/\eta_1}}. \tag{5.2}$$

The rank distribution (5.1) interpolates between two asymptotic power laws $\propto x^{\beta_0}$, $x^{\beta_0-\beta_1}$, valid for $x \ll b_1$ and $b_1 \ll x$, respectively. As is evident from Figs. 13–15, the transitional region between the power-law limits is quite extended, and the second power law only applies to the lower tail above a firm size of 10^3 employees. The power-law indices and amplitudes are listed in Table 2.

In contrast to the 1851–61–71 Censuses, the lower tail of the empirical rank distribution of the 1881 Census is weakly curved, cf. Fig. 16, which suggests to try a power law with lognormal cutoff for the least-squares fit,

$$R(x) = Ax^{\beta_0}\frac{1}{(1+(x/b_1)^{\beta_1/\eta_1})^{\eta_1\log x}}. \tag{5.3}$$

This is a special case of the lognormally cut power-law density $\rho_{\mathrm{LN}}(x)$ in (2.3). (In this case, the last factor of the product defining $\rho(x)$ in (1.1) is omitted, which means to replace n by $n-1$ in (1.1) and (1.2). We can then put $\delta = 1$ in (2.3) and (2.4), without loss of generality, since this parameter can be absorbed in the fit parameters β_n and η_n.) For large firm size $x \gg b_1$, the asymptotic limit of (5.3) is lognormal, $R(x) \sim Ax^{\beta_0-\beta_1\log(x/b_1)}$, whereas $R(x) \sim Ax^{\beta_0}$ in the opposite regime $x \ll b_1$. (The survival function (2.5) of this 1881 Census also has a lognormal cutoff, see Fig. 4.) The logarithmic derivative of the rank distribution (5.3) reads $R'(x)/R(x) = d_R(x)/x$, where, cf. (2.4),

$$d_R(x) = \beta_0 - \beta_1\frac{\log x}{1+(b_1/x)^{\beta_1/\eta_1}} - \eta_1\log(1+(x/b_1)^{\beta_1/\eta_1}). \tag{5.4}$$

The fitting parameters and goodness-of-fit indicators of the rank distributions (5.1) and (5.3) depicted in Figs. 13–16 are recorded in Table 2. The exponent β_0 is slightly negative and the amplitude A and exponents β_1, η_1 are positive, so that $R(x)$ is decreasing in the interval $[1,\infty)$. Fig. 17 shows the time evolution of the rank–size relation between 1851 and 1881; see also Fig. 9 for the corresponding evolution of the survival function and PDF.

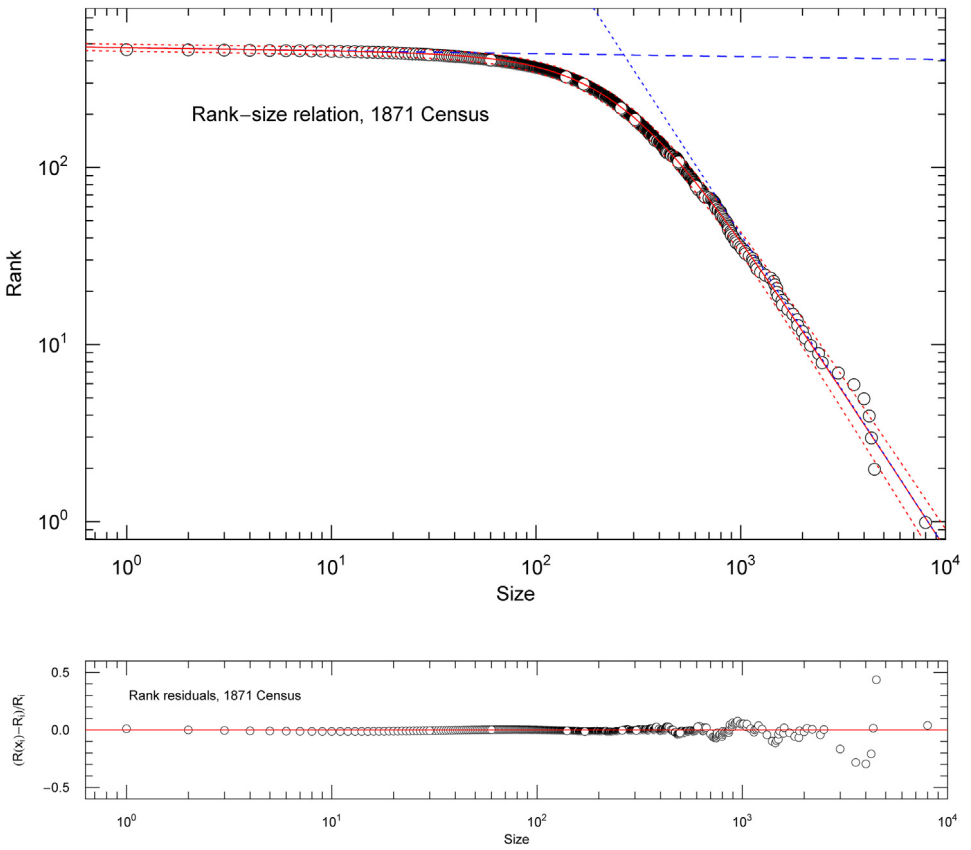

Fig. 15. Rank–size relation of the 1871 Census. Data points from Ref. [2]. As in the previous two examples in Figs. 13 and 14, the least-squares fit is performed with the broken power law (5.1) (solid red curve, fitting parameters in Table 2). The dashed and dotted blue straight lines indicate the asymptotic power-law components. The dotted red curves depict the 1σ error band.

Table 2
Parameters of the rank–size relations $R(x)$ of the 1851–61–71–81 Censuses, cf. Section 5. The firm-size rank distribution of the 1851–61–71 Censuses is the broken power law (5.1), and the rank distribution of 1881 is the lognormally cut power law (5.3), cf. Figs. 13–16. The fit parameters are the amplitude A, the negative power-law exponent β_0 and the parameters (b_1, β_1, η_1) defining the power-law and lognormal factors in (5.1) and (5.3). χ^2 denotes the minimum of the least-squares functional $\chi^2 = \sum_{i=1}^{N}(R(x_i) - R_i)^2/R_i^2$, with data points $(x_i, R_i)_{i=1,\dots,N}$ from Ref. [2]. Also listed are the degrees of freedom, dof $= N - 5$, as well as the standard error SE and determination coefficient R^2 of the fits, cf. Section 2.3.

	1851 Census	1861 Census	1871 Census	1881 Census
A	$412.97_{\pm 21.}$	$496.91_{\pm 23.}$	$475.16_{\pm 22.}$	$583.18_{\pm 24.}$
$-\beta_0$	$0.034480_{\pm 0.0095}$	$0.016258_{\pm 0.0082}$	$0.016819_{\pm 0.0084}$	$0.014877_{\pm 0.0073}$
b_1	$327.11_{\pm 16.}$	$336.45_{\pm 15.}$	$268.40_{\pm 13.}$	$252.73_{\pm 13.}$
β_1	$2.2513_{\pm 0.12}$	$2.0567_{\pm 0.091}$	$1.7575_{\pm 0.074}$	$0.20673_{\pm 0.0075}$
η_1	$1.3957_{\pm 0.098}$	$1.2751_{\pm 0.086}$	$1.0194_{\pm 0.089}$	$0.13114_{\pm 0.014}$
χ^2	1.18	0.698	0.592	0.673
dof	385	483	463	571
SE	2.74	2.40	1.98	2.73
$1 - R^2$	5.94×10^{-4}	2.91×10^{-4}	2.15×10^{-4}	2.69×10^{-4}

The Index function of a rank distribution is defined, as in Section 4.2, by the rescaled logarithmic derivative, $\text{Index}\,[R(x)] = -xR'(x)/R(x) = -d_R(x)$, with $d_R(x)$ in (5.2) for the 1851–61–71 Censuses and in (5.4) for the 1881 Census, see Fig. 18. (Compare also with Fig. 12 showing the Index curves of the survival functions of these censuses.) The rank Index functions of the 1851–61–71 Censuses have constant limits for large as well as small firm size, since the rank distributions (5.1) are asymptotic power laws in both regimes, whereas the Index function of the 1881 Census approaches a lin-log straight line (with positive slope) in the limit of large firm size, reflecting the lognormal cutoff of the rank distribution (5.3) in Fig. 16.

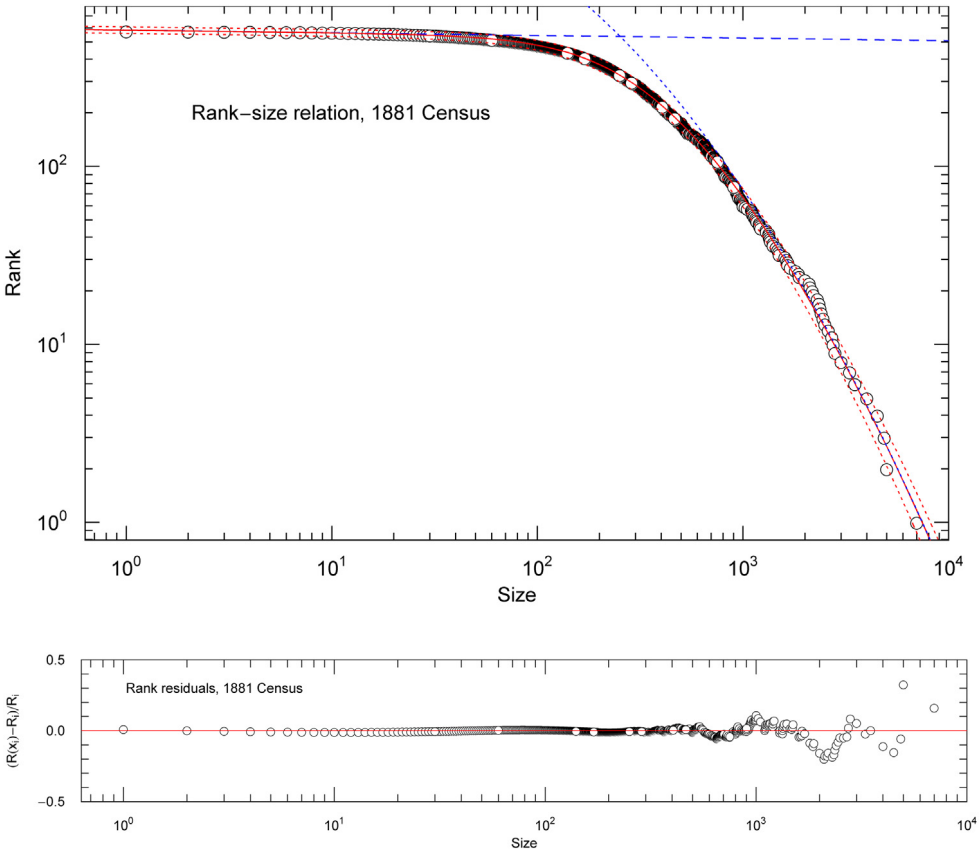

Fig. 16. Rank–size relation of the 1881 Census. Data points from Ref. [2]. Depicted is the least-squares fit of the lognormally cut power-law density $R(x)$ in (5.3) (solid red curve). The fitting parameters are listed in Table 2. In contrast to the rank distributions in Figs. 13–15, which have a straight power-law slope at large firm size (lower tail in the fourth logarithmic decade), the slope of $R(x)$ is weakly curved in this limit. The rank function (5.3) converges, for large firm size, to a lognormal density (indicated by the dotted blue parabolic curve), like the survival function of this census in Fig. 4. The dashed blue straight line is the power-law limit of $R(x)$ in (5.3) valid for small firm size. The dotted red curves show the 1σ error band.

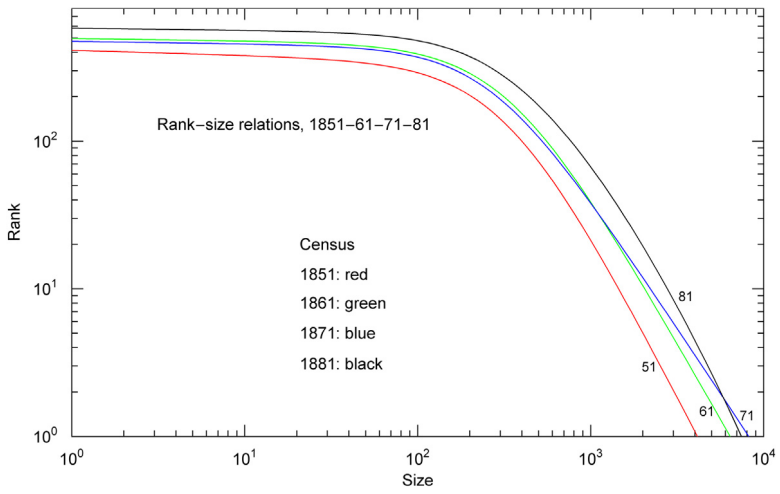

Fig. 17. Time evolution of the rank–size relation over a 30-year period. Depicted are the regressed rank distributions $R(x)$ of the 1851–61–71–81 Censuses, obtained from the least-squares fits in Figs. 13–16. The rank distributions of 1851–61–71 exhibit straight power-law tails (in the fourth size decade), whereas the lower tail of the 1881 rank distribution is asymptotically lognormal, cf. Fig. 16. (The evolution of the survival function and PDF during this period is shown in Fig. 9.)

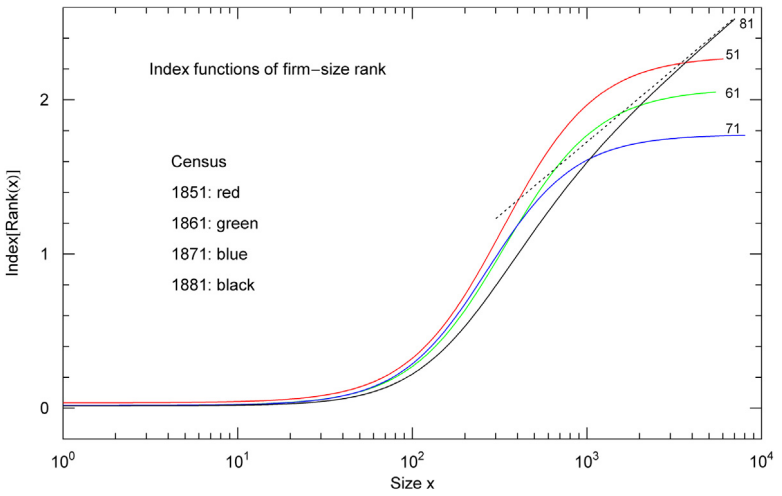

Fig. 18. Index curves of rank distributions. Rank Index functions are defined analogously to the firm-size Index functions in Fig. 12, Index $[R(x)] = -xR'(x)/R(x)$, cf. Section 5. As pointed out in Section 4.2, the Index function of a single power law is constant, whereas the Index function of a lognormal density shows as straight line with positive slope in lin-log representation. Since the rank distributions $R(x)$ in (5.1) of the 1851–61–71 Censuses have power-law limits for small as well as large firm size (see Figs. 13–15), their Index functions are constant in these asymptotic regimes. (The Index curves are depicted up to the largest firm size of the respective data set, cf. Table 1.) The Index function of the 1881 Census is also constant at small firm size, but terminates in an ascending linear slope (dotted black asymptote), since the rank distribution (5.3) of the 1881 Census is asymptotically lognormal, cf. Fig. 16.

6. Conclusion

There has been controversy as to whether the tails of firm-size (city-size, etc.) distributions are Pareto or lognormally distributed, cf., e.g., Refs. [40–43]. The reason commonly cited is that Pareto and lognormal densities are difficult to discriminate due to the slow logarithmic increase of the lognormal exponent, cf. the beginning of Section 4.1. First, defining a tail distribution requires to choose a truncation point in the empirical data set, which involves some arbitrariness even though optimization methods have been developed in this regard [24]. Then there is the frequently used notion of an upper and a lower tail, which already signals that the tail distribution cannot be accurately fitted with a single Pareto power law or lognormal density, otherwise there is no reason for this distinction. The data sets in Figs. 1–4 illustrate this very well: The tail distribution is defined by truncating the empirical distribution at a firm size of ten employees. The depicted Pareto fits of the tails (green straight lines in Figs. 1–4, derived in Ref. [1]) are quite accurate approximations in the second firm-size decade, less so in the third, and they fail in the fourth logarithmic decade, the lower tail. Here, we have argued that the problem with Pareto and lognormal densities is not so much that they are similar, but rather that they give similarly poor fits to the tail distributions, at least to those of the four Victorian firm-size censuses [1,2] studied here, and especially to the lower tails, which is the actual reason why they are difficult to distinguish.

Since Pareto and lognormal densities do not always give accurate fits when invoked as evidence for a Zipf or Gibrat law [24,26,41], we have proposed multiply broken power-law densities that are sufficiently adaptable to reproduce firm-size data over the full empirical size range. The densities are analytic functions, composed of power-law segments connected by smooth transitions, and admit a Weibull or lognormal cutoff if needed, cf. Sections 1 and 2. To demonstrate the efficiency of these multi-parameter distributions, the survival functions and rank distributions of the Victorian 1851–61–71–81 Censuses were modeled as broken power laws by performing least-squares fits to the data sets, cf. Figs. 1–4 and 13–16. The reconstruction of a stochastic process generating an empirically inferred broken power-law density is explained in Section 3.

In contrast to the nearly dispersionless data sets defining the survival functions in Figs. 1–4, the corresponding data sets of the firm-size PDFs in Figs. 5–8 are highly dispersed and would require data binning for regression. Therefore, the firm-size PDFs of the four censuses were calculated from the regressed survival functions. The goodness-of-fit parameters of the survival functions, PDFs and rank–size relations are listed in Tables 1 and 2. The PDFs in Fig. 9 show noticeable time variation within a 30-year interval (as do the corresponding rank distributions in Fig. 17), and they deviate from lognormal densities and Pareto tails, indicating disproportionate firm growth as manifested by size-dependent drift and diffusion coefficients in the Fokker–Planck equation.

Deviations of the fitted survival functions, PDFs and rank distributions from Pareto power laws and lognormals can be quantified by Index functions studied in Section 4.2. The Index curves of the survival functions of the four Victorian censuses are depicted as lin-log plots in Fig. 12 and the Index curves of the rank distributions in Fig. 18. The Index functions of Pareto power laws and lognormal densities are straight lines in lin-log representation, which cannot match the Index curves of the survival functions in Fig. 12 and rank distributions in Fig. 18, not even those of their tails defined

by discarding the first decade of the firm-size range. Therefore, neither Pareto nor lognormal densities can give reasonably accurate approximations to the tail distributions of the firm-size data depicted in Figs. 1–4 (survival functions) and Figs. 13–16 (rank–size relations). Realistic modeling of these data sets requires multi-parameter densities. Here, we have discussed specific examples to that effect, broken power laws as survival functions and rank distributions.

Acknowledgments

My thanks to seven anonymous referees for their extensive suggestions, questions and constructive criticism, which has greatly helped to improve an earlier draft of this paper.

References

[1] P. Montebruno, R.J. Bennett, C. van Lieshout, H. Smith, Physica A 523 (2019) 858.
[2] P. Montebruno, R.J. Bennett, C. van Lieshout, H. Smith, Mendeley data, 2019, http://dx.doi.org/10.17632/86xkkncmw3.1.
[3] Y. Wang, S. You, Physica A 457 (2016) 443.
[4] E. Calderín-Ojeda, F. Azpitarte, E. Gómez-Déniz, Physica A 461 (2016) 756.
[5] P. Soriano-Hernández, M. del Castillo-Mussot, O. Córdoba-Rodríguez, R. Mansilla-Corona, Physica A 465 (2017) 403.
[6] P. Soriano-Hernández, M. del Castillo-Mussot, I. Campirán-Chávez, J.A. Montemayor-Aldrete, Physica A 471 (2017) 733.
[7] M. Jagielski, K. Czyzewski, R. Kutner, H.E. Stanley, Physica A 474 (2017) 330.
[8] M. Campolieti, Physica A 503 (2018) 263.
[9] M. Campolieti, Physica A 534 (2019) 120821.
[10] B. Oancea, T. Andrei, D. Pirjol, Physica A 469 (2017) 486.
[11] B. Oancea, D. Pirjol, T. Andrei, Physica A 492 (2018) 403.
[12] E. Gómez-Déniz, E. Calderín-Ojeda, Physica A 436 (2015) 821.
[13] E. Calderín-Ojeda, Physica A 450 (2016) 385.
[14] Y. Chen, Physica A 443 (2016) 555.
[15] J. Luckstead, S. Devadoss, Physica A 465 (2017) 573.
[16] J. Luckstead, S. Devadoss, D. Danforth, Physica A 474 (2017) 237.
[17] L.M. Cortés, A. Mora-Valencia, J. Perote, Physica A 485 (2017) 35.
[18] S. Da Silva, R. Matsushita, R. Giglio, G. Massena, Physica A 512 (2018) 68.
[19] H.S. Kwong, S. Nadarajah, Physica A 513 (2019) 55.
[20] S. Arshad, S. Hu, B.N. Ashraf, Physica A 513 (2019) 87.
[21] I. Băncescu, L. Chivu, V. Preda, M. Puente-Ajovín, A. Ramos, Physica A 526 (2019) 121017.
[22] A. Ghosh, B. Basu, Physica A 528 (2019) 121094.
[23] M.E.J. Newman, Contemp. Phys. 46 (2005) 323.
[24] A. Clauset, C.R. Shalizi, M.E.J. Newman, SIAM Rev. 51 (2009) 661.
[25] M.V. Simkin, V.P. Roychowdhury, Phys. Rep. 502 (2011) 1.
[26] M.P.H. Stumpf, M.A. Porter, Science 335 (2012) 665.
[27] R. Tomaschitz, Physica A 483 (2017) 438.
[28] T.O. Kvålseth, Am. Stat. 39 (1985) 279.
[29] R. Tomaschitz, Physica A 524 (2019) 130.
[30] X. Gabaix, Q. J. Econ. 114 (1999) 739.
[31] X. Gabaix, Annu. Rev. Econ. 1 (2009) 255.
[32] J. Córdoba, Int. Econ. Rev. 49 (2008) 1463.
[33] M.H.R. Stanley, L.A.N. Amaral, S.V. Buldyrev, S. Havlin, H. Leschorn, P. Maass, M.A. Salinger, H.E. Stanley, Nature 379 (1996) 804.
[34] Y. Zou, Physica A 535 (2019) 122344.
[35] R. Pascoal, M. Augusto, H. Rocha, Physica A 531 (2019) 121797.
[36] J. Eeckhout, Amer. Econ. Rev. 94 (2004) 1429.
[37] B.B. Ahundjanov, S.B. Akhundjanov, Physica A 526 (2019) 120944.
[38] Y. Aydogan, M. Donduran, Physica A 533 (2019) 122066.
[39] A. Mura, M.S. Taqqu, F. Mainardi, Physica A 387 (2008) 5033.
[40] M. Mitzenmacher, Internet Math. 1 (2004) 226.
[41] M. Levy, Amer. Econ. Rev. 99 (2009) 1672.
[42] J. Eeckhout, Amer. Econ. Rev. 99 (2009) 1676.
[43] M. Bee, M. Riccaboni, S. Schiavo, Econ. Lett. 120 (2013) 232.