THE NATURE OF FIRM GROWTH[*]

Benjamin W. Pugsley
University of Notre Dame

Petr Sedláček
University of Oxford

Vincent Sterk
University College London

November 2017

Abstract

Only half of all startups survive past the age of five and surviving businesses grow at vastly different speeds. Using micro data on employment in the population of U.S. businesses, we estimate that the lion's share of these differences is driven by ex-ante heterogeneity across firms, rather than by ex-post shocks. We embed such heterogeneity in a firm dynamics model and study how ex-ante differences shape the distribution of firm size, "up-or-out" dynamics, and the associated gains in aggregate output. "Gazelles" –a small subset of startups with particularly high growth potential– emerge as key drivers of these outcomes. Analyzing changes in the distribution of ex-ante firm heterogeneity over time reveals that gazelles are driven towards extinction, creating substantial aggregate losses.

Keywords: Firm Dynamics, Startups, Macroeconomics, Big Data

JEL Codes: D22, E23, E24

[*] We thank Costas Arkolakis, Mark Bils, Richard Blundell, Christian Bayer, Vasco Carvalho, Steven Davis, Urban Jermann, Greg Kaplan, Fabien Postel-Vinay and Kjetil Storesletten, as well as seminar and conference participants at the Barcelona GSE summer forum, the 2017 CAED conference in Seoul, Cambridge University, University of California Davis, CIRPÉE/CIRANO/CIREQ Montréal, Duke University, the Federal Reserve Bank of San Francisco, the Federal Reserve Board, the University of Melbourne, Monash University, Ifo Munich, the University of Kent, the University of Nottingham, University of Sydney, Tilburg University, the 2017 SED meetings in Edinburgh, and the University of Queensland for helpful comments. We also are grateful for excellent research assistance by Harry Wheeler. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

1 Introduction

High-growth firms are widely seen as pivotal contributors to economic prosperity, if only for the large number of jobs that they create, see e.g. Haltiwanger, Jarmin, Kulick, and Miranda (2016). But what is it that distinguishes such firms from others that stay small throughout their lives? One view is that, following entry, firms are hit by ex-post shocks to productivity or demand; some startups are lucky and grow into large firms. An alternative view is that there are ex-ante differences in the growth profiles of startups. Some types of startups are poised for growth, for example due to a highly scalable technology or business idea, whereas others are destined to stay small. Although both views seem plausible, there is little empirical evidence on the relative importance of the two in shaping firm dynamics.

While their origins are not yet fully understood, firm dynamics have long been recognized in the literature as a key determinant of macroeconomic outcomes (Hopenhayn and Rogerson (1993), Melitz (2003), Klette and Kortum (2004)). More recently, Decker, Haltiwanger, Jarmin, and Miranda (2016) have documented a downward trend in the skewness of firm growth rates, and put forward the idea that a disappearance of high-growth firms might have driven the slump in U.S. employment and productivity growth, observed over the last decade. However, the origins and implications of the trend are still unclear. Possibly, the U.S. no longer offer a fertile ground for entrepreneurs to create high-potential startups, founded on ambitious business models. Clearly, this would have important repercussions for the U.S. macro economy. Alternatively, the trend could reflect a mere change in the distribution of ex-post shocks faced by individual firms, which might largely wash out at the aggregate level.

This paper uses the Longitudinal Business Database (LBD), an administrative panel covering nearly all private employers in the United States from 1976 to 2012, to dissect the firm growth process and changes thereof. We follow startups for twenty years after they enter and estimate the extent to observed differences across firms are driven by ex-ante heterogeneity and to what extent they are formed by ex-post shocks.[1] We do so

[1] Another important dimension of heterogeneity, on which we do not focus in this paper, relates

using both a reduced-form model and a structural firm dynamics model, both of which allow for heterogeneous ex-ante profiles as well as different types of ex-post shocks. The reduced-form model has the benefit of simplicity and yields analytical formulas which help us understand the identification of the key parameters, whereas the structural firm dynamics model accounts for endogenous selection. In addition, the structural model allows us to distinguish between types of startups with different ex-ante growth and survival profiles, to analyze how their prevalence has changed over time, and to quantify the aggregate consequences of such changes.

Our central piece of empirical evidence is the cross-sectional autocovariance function of business-level employment by age. We thereby take inspiration from the earnings dynamics literature, which has long recognized that autocovariances help to distinguish shocks from deterministic profiles (see e.g. MaCurdy, 1982; Abowd and Card, 1989; Guvenen, 2009; Guvenen and Smith, 2014). Perhaps surprisingly, the literature on firm dynamics does not have a similar tradition. To the best of our knowledge, the basic autocovariance structure of employment by age has not been systematically documented. Instead, the firm dynamics literature has emphasized the profiles of average size and exit by age, see e.g. Haltiwanger, Jarmin, and Miranda (2013), Hsieh and Klenow (2014) and Akcigit, Alp, and Peters (2017).[2] We also target these important moments in the structural model, but highlight the wealth of additional information that is embodied in the autocovariance structure.

A central finding of our study is that ex-ante heterogeneity accounts for a large share of the cross-sectional dispersion in employment, conditional on age, ranging from more than ninety percent in the first year after entry to around forty percent twenty years later. This finding relates to several earlier studies. Abbring and Campbell (2005) use sales data of Texas bars in the first year after entry. They specify a sophisticated model tailored to this industry and estimate that pre-entry scale decisions account for about 40 percent of the variation in sales in the first year. Campbell and De Nardi (2009) and

to the role of supply versus demand factors. For evidence on this, see e.g. Hottman, Redding, and Weinstein (2016) and Foster, Haltiwanger, and Syverson (2016).

[2] Cabral and Mata (2003) document the evolution of the skewness of the size distribution with age.

Hurst and Pugsley (2011) present survey evidence that many nascent entrepreneurs do not expect their business to grow large.[3] Sedláček and Sterk (2017) document strong cohort effects in firm-level employment, depending on the state of the business cycle in the year of entry. The importance of the composition of the firm population is also emphasized by Pugsley and Şahin (2016), who document a strong trend in the U.S. towards older firms, which is the result of accumulating startup deficits.

The structural firm dynamics model we employ in our analysis follows the tradition of Hopenhayn (1992), Melitz (2003), and Luttmer (2007), and features endogenous entry and exit and general equilibrium forces. We introduce a multi-dimensional idiosyncratic process into this framework, to allow not only for persistent and transitory ex-post shocks, but also for heterogeneity in ex-ante growth and survival profiles. This relatively rich process aligns with the reduced-form evidence and is needed to obtain a good fit with the empirical autocovariance structure. As such, our empirical evidence points towards models allowing for ex-ante differences in growth across firms, along the lines of e.g. Luttmer (2011).

After taking the structural model to the data, we show that ex-ante heterogeneity is not only an important determinant of size dispersion, but also of the well-documented "up-or-out" dynamics. That is, the fact that many young firms shut down while surviving businesses grow quickly is in large part driven by ex-ante heterogeneity. The impact of this materializes via selection on ex-ante growth profiles: firms with little growth potential exit, allowing firms with high potential to blossom. Indeed, we find that selection on ex-ante heterogeneity, as well as its interaction with ex-post shocks, makes the age profile of average size substantially more upward-sloping. Associated with this steeper slope is a large gain in aggregate output. By contrast, ex-post shocks alone create only small selection effects and hence by themselves matter little for aggregate output.

We also examine specifically the contribution of startups with high growth potential, known as "gazelles" in the literature. The model allows us to back out the distribution of

[3] Guzman and Stern (2015) and Belenzon, Chatterji, and Daley (2017) show that firm growth is partly predictable based on observable characteristics at the time of startup.

ex-ante growth profiles, which we exploit to identify a subset of high-potential startups with projected annual growth of more than twenty percent in the first five years after entry. We find that such ex-ante gazelles account for only about five percent of all startups. Nonetheless, they contribute greatly to the positive slope of the age profile of average size, and to the associated gains in aggregate output.

Finally, we use the model to understand the sources and consequences of an apparent structural change in the growth dynamics of U.S. firms. We consider two subsamples of the data (1976-1996 and 1992-2012). Across these two subsamples, the autocovariance matrix has remained remarkably stable, and so has the profile of exit by age. However, what has changed is the profile of average size by age. This profile has flattened, implying less growth on average in the recent sample. This finding relates to the evidence presented by Hsieh and Klenow (2014) who document that the average size profile in India and Mexico is much flatter than in the U.S., and find large implications for aggregate productivity.

Examining the flattening of the average size profile more closely, we observe that it occurred in a staggered manner. That is, it happened due to flatter profiles of incoming cohorts of startups since the late 1980's, rather than as a simultaneous decline in the size of older firms. This observation suggests that a change in the distribution of ex-ante growth profiles was responsible for the flattening.

To study the underlying changes and their implications directly, we re-estimate the model on the two subsamples. We then evaluate how the ex-ante growth profiles of startups has changed over time. We find a substantial decline in the prevalence of ex-ante gazelles in the population of startups, and that the growth profile of gazelles beyond age fifteen has flattened. These changes together account for about half of the flattening of the average size profile across all firms, despite the fact that gazelles make up only a small fraction of all startups.

Our findings thus confirm with the concerns raised by Decker, Haltiwanger, Jarmin, and Miranda (2016) on the disappearance of high-growth firms. Moreover, our results show that this phenomenon is primarily due to a change in ex-ante profiles of startups, dating back three decades. Finally, we find that the aggregate output loss implied by the change in firm dynamics between the two samples is about 4.5 percent, with larger losses to follow if the observed trend continues.

The remainder of this paper is organized as follows. Section 2 presents the data, the reduced-form model, and initial estimates of the importance of ex-ante heterogeneity for size dispersion. Section 3 describes the structural firm dynamics model and the parametrization procedure. Baseline results from the structural model are presented in Section 4, after which Section 5 presents the results from the split-sample analysis. Finally, Section 6 concludes.

2 Reduced-form evidence

This section estimates to what extent cross-sectional variation in employment is driven by ex-ante heterogeneity and to what extent it is formed by ex-post shocks. We begin by describing our data set and the central piece of empirical evidence used in the estimation: the autocovariance function of logged employment, at the establishment- and firm-level. Next, we specify and estimate a flexible employment process incorporating both ex-ante heterogeneity and ex-post shocks. We show analytically that all the relevant model parameters can be identified from the autocovariance function, and we use the analytical formulas to understand which features of the data drive the results.

2.1 Data

The analysis is based on administrative micro data on employment in the United States, taken from the from Census Longitudinal Business Database (LBD). The data cover almost the entire population of employers over the period between 1979 and 2012. As the unit of analysis we consider logged employment in both establishments and firms.[4] We construct a panel of employment at the establishment- and firm-level in the year of

[4] Establishments are the physical units of a firm, located a specific addresses. A firm can consist of one or multiple establishments. The data are a snapshot taken in the month of March of each year. The age of an establishment is computed as the current year, minus the first year an establishment came into existence. The age of a firm is computed as the age of its oldest establishment.

startup (age zero) up to age nineteen. Prior to the analysis, we take out a fixed effect for the birth year of the establishment (or firm) and for its industry classification at the 4-digit level. In order to streamline the discussion, we will use the term "business" whenever we refer to both establishments and firms simultaneously.

2.2 The autocovariance structure of employment

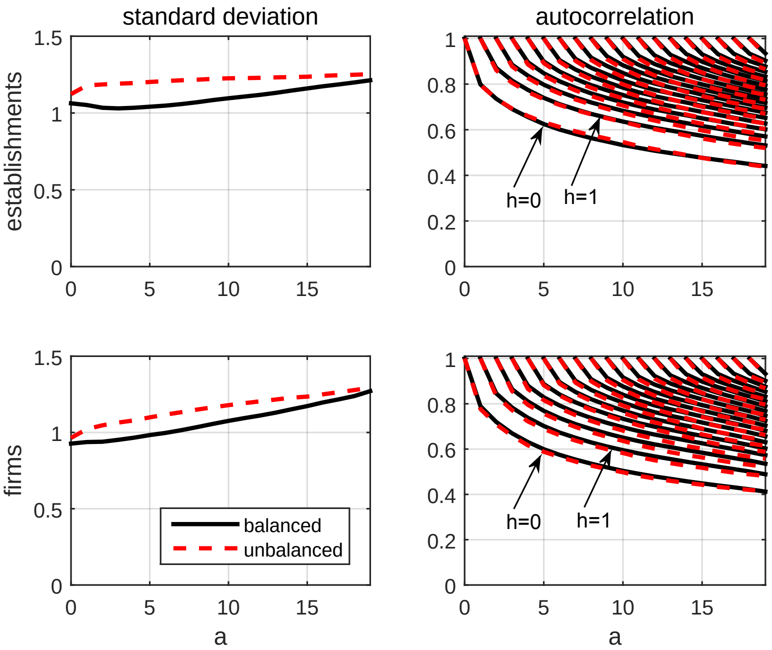
Figure 1 presents our main piece of empirical evidence: the cross-sectional autocovariance structure of logged employment, conditional on age (a). In order to understand this structure more easily, we break down the autocovariances into standard deviations, displayed in the left panels, and autocorrelations, shown in the right panels. The figure presents this information for both establishments (top panels), and for firms (bottom panels), as well as for a balanced panel, containing businesses surviving at least up to age 19, and an unbalanced panel, including all businesses in our data set. Clearly, differences in autocovariances between the balanced and unbalanced panels originate primarily from different cross-section dispersion by age, while the autocorrelations are remarkably similar across the two panels.

Let us first focus on the cross-sectional standard deviations by age, shown in the left panels. Standard deviations are between 1 and 1.4 log points for both establishments and firms, indicating large size differences even at young ages. Also, the cross-sectional dispersion generally increases with age and this is true for both the balanced and unbalanced panels. The latter indicates that the observed fanning out of the size distribution with age is not purely driven by selective exit of certain businesses.[5]

The right panels of Figure 1 depict the associated autocorrelations of logged employment across businesses, by age. Keeping the initial age h fixed, the autocorrelations decline convexly with lag length $a - h$. Importantly, the autocorrelations appear to stabilize at relatively high levels. For instance, the autocorrelations between logged employment at ages zero and ten and zero and nineteen are 0.55 and 0.44, respectively. On the other hand, for a fixed lag length, the autocorrelations are increasing in age.

[5] The exception to this pattern is the flat age profile of cross-sectional dispersion for establishments below age five in the balanced panel.

Figure 1: Standard deviations and autocorrelations of log employment by age



Note: The left panels show cross-sectional standard deviations of log employment by age (a) for establishments (top left panel) and firms (bottom left panel). The right panels show autocorrelations of log employment between ages a and $0 \leq h \leq a$ for establishments (top right panel) and firms (bottom right panel). "Balanced" refers to a panel of establishments (firms) which survived at least up to age 19, while "unbalanced" refers to a panel of all establishments (firms).

For instance, the correlation of log employment between age zero and age nine is 0.56, whereas the corresponding correlation between age ten and nineteen is 0.73. These empirical patterns contain important information on the relative importance of ex-ante heterogeneity and ex-post shocks, as we will discuss below in detail.

2.3 Employment process

To understand what we can learn from the autocovariances about the importance of ex-ante versus ex-post heterogeneity, we now consider a reduced-form model of employment which includes both sources of heterogeneity. As will become clear, the model is flexible enough to provide a good description of the observed patterns in the data. It also nests as special cases reduced-form representations of several prominent structural firm dynamics models in the literature, such as the models of Hopenhayn and Rogerson

(1993) and Melitz (2003).

Let $n_{i,a}$ be the employment level of an individual business i at age a and consider the following process for this variable:

$$\begin{aligned} \ln n_{i,a} &= \ln n_{i,a}^{EXA} + \ln n_{i,a}^{EXP}, && && (1) \\ \ln n_{i,a}^{EXA} &= u_{i,a} + v_{i,a}, && \text{(ex-ante component)} \\ \ln n_{i,a}^{EXP} &= w_{i,a} + z_{i,a}, && \text{(ex-post component)} \end{aligned}$$

where

$$\begin{array}{lllll} u_{i,a} = & \rho_u u_{i,a-1} + \theta_i, & u_{i,-1} \sim iid(\mu_{\tilde{u}}, \sigma^2_{\tilde{u}}), & \theta_i \sim iid(\mu_\theta, \sigma^2_\theta), & \rho_u \in [0,1), \\ v_{i,a} = & \rho_v v_{i,a-1}, & v_{i,-1} \sim iid(\mu_{\tilde{v}}, \sigma^2_{\tilde{v}}), & & \rho_v \in [0,1), \\ w_{i,a} = & \rho_w w_{i,a-1} + \varepsilon_{i,a}, & w_{i,-1} = 0, & \varepsilon_{i,a} \sim iid(0, \sigma^2_\varepsilon), & \rho_w \in [0,1), \\ z_{i,a} \sim & iid(0, \sigma^2_z). & & & \end{array}$$

Here, all shocks are drawn from distributions which are i.i.d. across time and across firms, and we let μ denote a mean and σ^2 a variance.

In the above process, $\ln n_{i,a}^{EXA} = u_{i,a} + v_{i,a}$ captures the *ex-ante* component, which is governed by three stochastic, business-specific parameters which are drawn independently just prior to startup, at age $a = 0$. The parameter θ_i is a permanent component which accumulates gradually with age at rate ρ_u. The second parameter, $u_{i,-1}$, is a transitory ex-ante draw which allows for the possibility that the path of the ex-ante component starts away from zero. The third parameter, $v_{i,-1}$, is a second initial condition which is allowed to die out at its own speed, as the business ages.

In the long run, the ex-ante component reaches a steady state level given by $\ln n_{i,\infty}^{EXA} = \theta_i/(1-\rho_u)$. Since this level differs across businesses, the process admits heterogeneity in long-run steady states. Moreover, since initial conditions differ across businesses, we allow for heterogeneity in the paths from initial employment towards the steady states. Finally, since the process includes two separate initial conditions, each with their own persistence parameter, we allow businesses to gravitate towards their steady-state levels at different speeds. We thus allow for rich heterogeneity in ex-ante growth profiles.

The *ex-post* shocks enter the model via a second component, $\ln n_{i,a}^{EXP} = w_{i,a} + z_{i,a}$. Here, $w_{i,a}$ captures persistent ex-post shocks, and is modeled as an autoregressive process of order one, with i.i.d. innovations given by $\epsilon_{i,a}$ and a persistence parameter denoted by ρ_w. The initial level of $w_{i,a}$ is normalized to zero for all businesses. We further introduce purely transitory ex-post shocks via an i.i.d. component denoted by $z_{i,a}$. The process for the ex-post component is constructed such that the unconditional mean is zero at any age, so that it does not capture any of the heterogeneity in ex-ante profiles.

The process postulated above nests various specifications commonly used in the firm dynamics literature to model firm-level shocks. For example, Hopenhayn and Rogerson (1993) assume an AR(1) for firm-level productivity, with a common constant across firms and heterogeneous initial draws. In their baseline model without distortions, the firm-level shocks map one-for-one into employment. We obtain their specification by setting $\sigma_u = \sigma_\theta = \sigma_z = 0$ and $\rho_v = \rho_w$. By contrast, Melitz (2003) and Hsieh and Klenow (2009) allow, like us, for heterogeneity in steady-state levels, but abstract from ex-post shocks and assume that steady states are immediately reached. We obtain their process by setting $\sigma_u = \sigma_v = \sigma_z = 0$, which implies that $\ln n_{i,a} = \theta_i$ at any age.[6] Our baseline process also aligns with models with richer heterogeneity ex-ante profiles and/or ex-post shocks, as proposed by for example Luttmer (2011) and Arkolakis (2016) and Arkolakis, Papageorgiou, and Timoshenko (forthcoming).

2.4 Parameter identification

We now demonstrate the usefulness of the autocovariance matrix in quantifying the role of ex-ante versus ex-post heterogeneity. We do so by showing analytically that

[6] Our process also nests specifications commonly assumed in the econometrics literature on dynamic panel data models, see for example Arellano and Bond (1991). This literature typically assumes an autoregressive process, like Hopenhayn and Rogerson (1993), but allow for heterogeneity in the constant θ_i and thus in steady-state levels. Commonly, however, θ_i is differenced out and hence no estimate is provided for σ_θ, a key parameter in our application. Moreover, the panel data econometrics literature commonly assumes that $\rho_u = \rho_v = \rho_w$. In our application, it turns out that this assumption is too restrictive to provide a good fit of the observed autocovariance matrix. Our results thus caution against the use of standard panel data estimators when applied to employment dynamics of young establishments.

all the key parameters can be identified from the autocovariance matrix. Given the process postulated above, the covariance of employment of a business at age a and at age $h = a - j$, where $0 \leq j \leq a$ is the lag length, can be expressed as:

$$\begin{aligned} Cov\left(\ln n_{i,a}, \ln n_{i,a-j}\right) &= \rho_u^j \rho_u^{2(a-j+1)} \sigma_{\tilde{u}}^2 + \rho_v^j \rho_v^{2(a-j+1)} \sigma_{\tilde{v}}^2 \\ + \left(1-\rho_u^{a+1}\right)\left(1-\rho_u^{a-j+1}\right) \frac{\sigma_\theta^2}{\left(1-\rho_u\right)^2} &+ \rho_w^j \frac{1-\rho_w^{2(a-j+1)}}{1-\rho_w^2} \sigma_\varepsilon^2 + 0^j \sigma_z^2. \end{aligned} \tag{2}$$

This result is derived in Appendix A.1. The autocovariance function is a nonlinear function of the persistence and variance parameters of the components of the underlying process. Given that in total there are eight such parameters, we need an autocovariance matrix with at least eight elements for identification.[7]

To understand the identification, it is useful to consider the autocovariance at an infinite lag length, i.e. letting the age a approach infinity keeping the initial age $h = a - j$ fixed:

$$\lim_{a \to \infty} Cov\left(\ln n_{i,a}, \ln n_{i,h}\right) = \frac{1-\rho_u^{h+1}}{\left(1-\rho_u\right)^2} \sigma_\theta^2.$$

When σ_θ equals zero, i.e. when there is no heterogeneity in steady-state levels, the autocovariance is zero. Thus, long-horizon autocovariances contain valuable information on the presence of ex-ante heterogeneity in steady-state levels. In Figure 1, autocorrelations appear to stabilize at long lag lengths, i.e. at high levels of a given $h = a - j$, suggesting that such heterogeneity is indeed a feature of the data.

We can obtain further insight into the identification by considering the fit of an AR(1) process with a homogeneous constant.[8] This restricted process features no heterogeneity in steady state levels ($\sigma_\theta = 0$) and hence autocovariances become zero at infinite lag lengths. In order for the model to fit the high autocovariances observed in the data, the persistence parameter ρ_w needs to be close to one. As can be seen from the second-to-last term in Equation (2), however, this implies that the autocovariance

[7] Note that the mean parameters μ_θ, $\mu_{\tilde{u}}$ and $\mu_{\tilde{v}}$ are not identified by the autocovariance function. These parameters, however, are also not needed to quantify the importance of ex-ante versus ex-post heterogeneity.

[8] The considered AR(1) process is consistent with the Hopenhayn and Rogerson (1993) model without adjustment costs.

function becomes close to being linear in age, a, and in lag length, j. Figure 2 shows that such linear patterns are in contrast with the data, in which the autocovariance function is convex in age and concave in lag length.[9]

In contrast to the AR(1) model, our richer baseline process admits steady-state heterogeneity. This relaxes the need for the persistence parameters of being close to one in order to match the long-run autocovariances in the data. Identification of the various components of the process derives from the fact that each has its own specific impact on how the autocovariance function depends of age and the lag length, as can be seen from Equation (2).

2.5 Estimation procedure

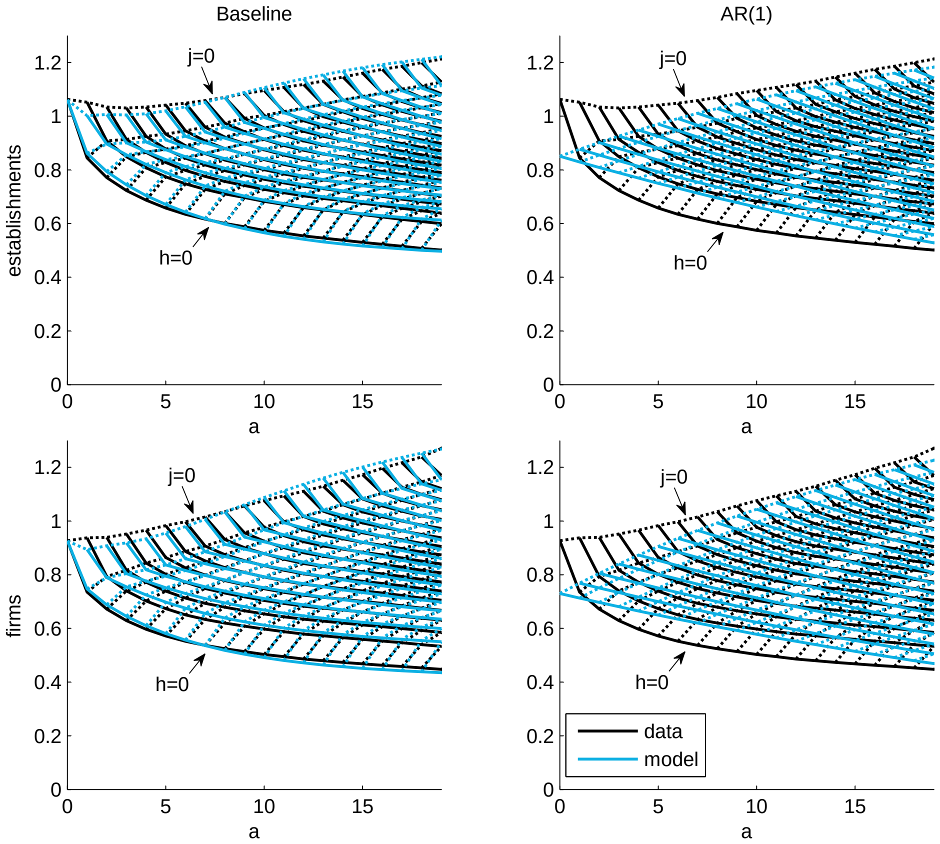
We estimate the parameters of the process using a minimum distance procedure, as proposed by Chamberlain (1984). Specifically, we minimize the sum of squared deviations of the upper triangular parts of the autocovariance matrix implied by the process, from its counterpart in the data. Because there is a very large number of observations underlying each element in the empirical autocovariance matrix, we assign equal weights to all elements in the estimation procedure. See Appendix A.2 for more details. Our baseline results apply to the balanced panel data set.

2.6 Model fit and parameter estimates

Let us begin by inspecting the model fit of our baseline specification, depicted in the left panels of Figure 2. The top left panel shows the empirical and model-generated autocovariance function for establishments and the bottom left panel shows the same for firms. In both cases the fit is very good, correctly capturing the convexly declining pattern of the autocovariances in the lag length, given the initial age h, and the concavely increasing pattern in age given the lag length $j > 0$. Finally, the model fits the non-monotonic pattern in cross-sectional dispersion by age.

[9] Note that $\rho_w = 1$ would introduce a random walk component, consistent with Gibrat's law. However, violation of Gibrat's law in the data has been documented in the literature, in particular among new and young firms, see e.g. Haltiwanger, Jarmin, and Miranda (2013).

Figure 2: Autocovariance matrices: reduced-form models versus data



Note: Autocovariance of log employment between age $a = h + j$ and age $h \leq a$ in the data, and in the baseline model (left panels) and an AR(1) model (right panels). Results are shown for firms (top panels) and establishments (bottom panel). Autocovariances in the data are from the balanced panel.

The corresponding parameter estimates are shown in Table 1. A key feature of our baseline process is the presence of dispersion in long-run steady states, governed by σ_θ and ρ_u. The point estimates imply a standard deviation of long-run steady-state employment levels of 0.76 for establishments and 0.71 for firms. These values are substantial when considering that the overall cross-sectional dispersion of twenty year old businesses is about 1.4 (see Figure 1).

As discussed above, we also fit an AR(1) process for illustrative purposes. The fit of this model is considerably worse compared to the baseline, with a root mean squared error that is two to four times as high as in our baseline (see the bottom line of Table 1). In the literature, the failure of an AR(1) to fit the data well has been established

Table 1: Parameter estimates from reduced-form model

	Establishments		*Firms*	
	Baseline	AR(1)	Baseline	AR(1)
ρ_u	0.2059		0.2183	
	(0.0015)		(0.0018)	
ρ_v	0.8415	0.9752	0.8323	0.9771
	(0.0010)	(0.0001)	(0.0014)	(0.0001)
ρ_w	0.9489	0.9752	0.9625	0.9771
	(0.0003)	(0.0001)	(0.0003)	(0.0001)
σ_θ	0.6031		0.5545	
	(0.0014)		(0.0015)	
$\sigma_{\widetilde{u}}$	2.0461		1.7425	
	(0.0174)		(0.0145)	
$\sigma_{\widetilde{v}}$	0.7378	0.9069	0.6951	0.8304
	(0.0017)	(0.0009)	(0.0021)	(0.0009)
σ_ε	0.2554	0.2610	0.2548	0.2676
	(0.0004)	(0.0002)	(0.0004)	(0.0003)
σ_z	0.2623		0.2716	
	(0.0006)		(0.0006)	
RMSE	0.0100	0.0387	0.0120	0.0259

Note: *RMSE* is the root-mean squared error of the autocovariance matrix in the model, relative to the data.

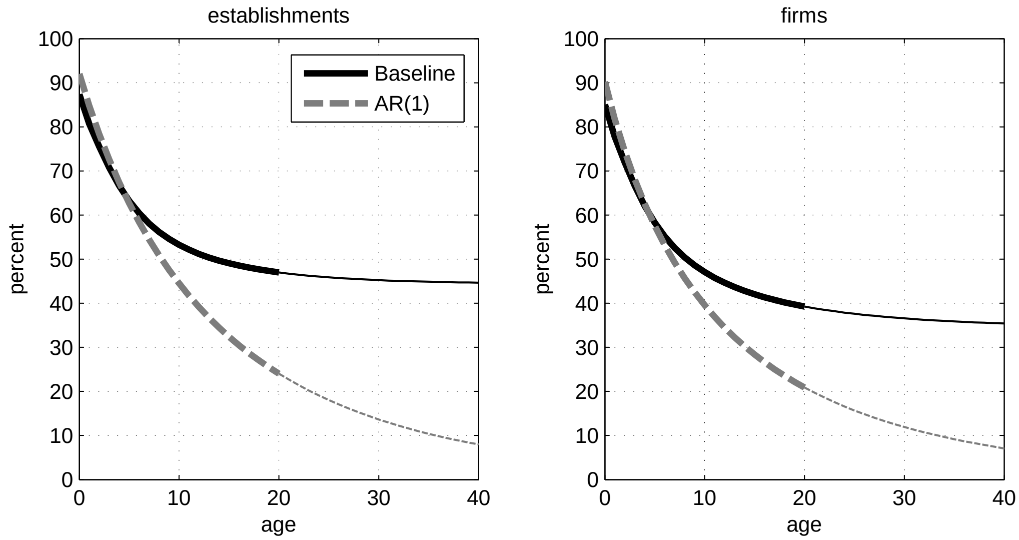
by Lee and Mukoyama (2015), who study manufacturing plants.

2.7 The importance of ex-ante and ex-post heterogeneity

With the estimated model at hand, we can quantify the relative importance of ex-ante profiles and ex-post shocks for the cross-section dispersion in employment. This is done based on Equation (2). With the lag length j set to zero, this equation provides a decomposition of the variance of size (log employment), at any given age a, into the contributions of the ex-ante and ex-post components. Figure 3 plots the fraction of the total variance that is accounted for by the ex-ante component. Thick lines denote the age groups used in the estimation, i.e. age zero to nineteen, whereas thin lines represent an extrapolation for businesses at age 20 or above using the point estimates.[10]

[10] The lines in the figure are point estimates. We have also computed confidence bands for this decomposition, but these are extremely narrow due to the very large number of data points used in the estimation. This is also reflected in very small standard errors around the point estimates for the

Figure 3: Contribution of ex-ante heterogeneity to cross-sectional employment dispersion



Note: Contribution of the ex-ante component, $\ln n_{i,a}^{EXA}$, to the cross-sectional variance of log employment, by age. Thin lines denote age groups not directly used in the estimation. The decomposition is based on Equation (2) with $j = 0$.

Figure 3 shows that for businesses in the year of startup, that is at age zero, the ex-ante component accounts for about 85 percent of the cross-sectional variance in size. The remainder is due to ex-post shocks that materialized in the first year. Considering older age groups, the contribution of ex-ante heterogeneity declines, but remains high. At age twenty, ex-ante factors account for 47 percent of the size variance among establishments, and around 40 among firms. In the data, more than seventy percent of the businesses are twenty years old or younger. Our results show that, among these businesses, ex-ante factors are a key determinant of size. Increasing age towards infinity, the contribution of ex-ante heterogeneity stabilizes at around 45 percent for establishments and 35 percent for firms. Therefore, even among very old businesses ex-ante factors contribute to a large chunk of the dispersion in size.

Figure 3 also plots the decomposition for the AR(1) process. While the estimated contribution of ex-ante heterogeneity among young firms is comparable to the baseline, discrepancies arise beyond age five. In the long run, the contribution of ex-ante hetero-

parameters, as can be observed from Table 1.

geneity converges to zero. This happens by construction, as the AR(1) does not admit heterogeneity in steady-state levels.

3 Structural model

In this section we estimate a structural firm dynamics model, which has several advantages relative to the reduced-form analysis. First, the structural model accounts for selective entry and exit, which might affect the estimated importance of ex-ante and ex-post heterogeneity. Importantly, firm selection is a multifaceted process which occurs along various dimensions of heterogeneity. Second, the structural model allows us to compute aggregates, and quantify the importance of ex-ante and ex-post heterogeneity for aggregate outcomes. Finally, since the structural model speaks not only to the autocovariance structure, but also to the profiles of average size and exit, by age, it enables us to fully characterize the population of startups according to their ex-ante growth and survival potential. We estimate the model for firms, and report results for establishments in Appendix B.4.

3.1 The model

We consider a closed general equilibrium economy with heterogeneous firms and endogenous entry and exit, as in Hopenhayn and Rogerson (1993). Following Melitz (2003) and others, each firm is monopolistically competitive and faces a demand schedule which is downward-sloping in the price they set. To model heterogeneity across firms, we embed an idiosyncratic process with the same structure as in the reduced-form analysis, thereby allowing for differences in both ex-ante profiles and ex-post shocks.

Households. The economy is populated by an infinitely-lived representative household who owns the firms and supplies a fixed amount of labor in each period, denoted by $\overline{N}$. Household preferences are given by $\sum_{t=0}^{\infty} \beta^t C_t$, where $\beta \in (0,1)$ is the discount factor. C_t is a Dixit-Stiglitz basket of differentiated goods given by:

$$C_t = \left(\int_{i \in \Omega_t} \varphi_{i,t}^{\frac{1}{\eta}} c_{i,t}^{\frac{\eta-1}{\eta}} \right)^{\frac{\eta}{\eta-1}},$$

where Ω_t is the measure of goods available in period t, $c_{i,t}$ denotes consumption of good i, η is the elasticity of substitution between goods, and $\varphi_{i,t} \in [0, \infty)$ is a stochastic and time-varying demand fundamental specific to good i. We consider a stationary economy from now on and simplify notation by dropping time subscripts.

The household's budget constraint is given by $\int_{i \in \Omega} p_i c_i = W\overline{N} + \Pi$, where p_i denotes the price of good i, W denotes the nominal wage and Π denotes firm profits. Utility maximization implies a demand schedule given by $c_i = \varphi_i \left(p_i/P\right)^{-\eta} C$, where P is a price index given $P \equiv \left(\int_{i \in \Omega} \varphi_i p_i^{1-\eta} \right)^{\frac{1}{1-\eta}}$, so that total expenditure satisfies $PC = \int_{i \in \Omega} p_i c_i$.

Incumbent firms. There is an endogenous measure of incumbent firms, each of which produces a unique good. Firms are labeled by the goods they produce $i \in \Omega$. The production technology of firm i is given by $y_i + f = n_i$, where y_i is the output of the firm, n_i is the amount of labor input (employment) and f is a fixed cost of operation common to all firms, denominated in units of labor. It follows that firms face the following profit function:

$$\pi_i = p_i y_i - W n_i.$$

Additionally, given the market structure, each firm faces a demand constraint given by

$$y_i = \varphi_i \left(p_i/P\right)^{-\eta} Y, \tag{3}$$

which is the demand schedule of the household combined with anticipated clearing of goods markets, which implies $c_i = y_i$ and $Y = C$.

At the beginning of each period, a firm may be forced to exit exogenously with probability $\delta \in (0, 1)$. If this does not occur, the firm has the opportunity to exit endogenously and avoid paying the fixed cost. If the firm chooses to remain in operation, it must pay the fixed cost and in turn it learns its demand fundamental φ_i. Given its production technology and demand function, the firm sets its price p_i (and implicitly y_i, n_i and π_i) to maximize the net present value of profits. The price-setting problem is static and the firm sets prices as a constant markup over marginal costs W:

$$p_i = \frac{\eta}{\eta - 1} W.$$

We let labor be the numeraire so that $W = 1$, and define the real wage $w \equiv W/P$ as the price of labor in terms of the Dixit-Stiglitz consumption basket C. Using this result, we can express profits as $\pi_i = \varphi_i w^{-\eta} C \chi - f$, where $\chi \equiv \frac{(\eta-1)^{\eta-1}}{\eta^{\eta}}$, and labor demand as $n_i = \varphi_i \left(\frac{\eta}{\eta-1}\right)^{-\eta} w^{-\eta} C + f$. Note that fluctuations in the demand fundamental directly map into the firms' employment.

The demand fundamental φ_i is a function of an exogenous underlying Markov state vector, denoted $\mathbf{s}_i$. The value of a firm at the moment the exit decision is taken, denoted V, can now be expressed as:

$$V(\mathbf{s}_i) = \max \left\{ \mathbb{E}\left[\pi\left(\mathbf{s}_i'\right) + \beta(1-\delta) V\left(\mathbf{s}_i'\right) \middle| \, \mathbf{s}_i \right], 0 \right\}.$$

In the above equation $\mathbf{s}_i'$ denotes the value of the state after the continuation decision is taken. Accordingly, we can express the profit, output, employment and exit policies as $\pi_i = \pi(\mathbf{s}_i')$, $y_i = y(\mathbf{s}_i')$, $n_i = n(\mathbf{s}_i')$, and $x_i = x(\mathbf{s}_i)$, respectively.

Firm entry. Firm entry is endogenous and requires paying an entry cost f^e, denominated in units of labor. After paying the entry cost at the beginning of a period, the firm observes its initial level of $\mathbf{s}_i$, at which point it becomes an incumbent. Free entry implies the following condition:

$$wPf^e = \int V(\mathbf{s}) \, G(d\mathbf{s}),$$

where G is the distribution from which the initial levels of $\mathbf{s}_i$ are drawn.

Aggregation and market clearing. Let $\mu\left(\mathbf{S}\right)$ be the measure of firms in $\mathbf{S} \in \mathcal{S}$, where is $\mathcal{S}$ is the Borel $\sigma-$algebra generated by $\mathbf{s}$. Given the exit policy, $\mu\left(\mathbf{S}\right)$ satisfies:

$$\mu\left(\mathbf{S}'\right) = \int \left[1 - x\left(\mathbf{s}\right)\right] F\left(\mathbf{S}'|\mathbf{s}\right)\left[\mu\left(d\mathbf{s}\right) + M^e G\left(d\mathbf{s}\right)\right],$$

where M^e denotes the measure of entrants and $F\left(\mathbf{S}'|\, s\right)$ is consistent with the transition law for $\mathbf{s}_i$. The total measure of active firms is given by:

$$\Omega = \int \mu\left(d\mathbf{s}\right).$$

Labor market clearing implies that total labor supply equals total labor used for production, for the fixed cost, and for the entry cost:

$$\bar{N} = \int y\left(\mathbf{s}'\right)\mu\left(d\mathbf{s}'\right) + \int f\left[1 - x\left(\mathbf{s}\right)\right]\left[\mu\left(d\mathbf{s}\right) + M^e G\left(d\mathbf{s}\right)\right] + M^e f^e.$$

Stochastic driving process. In line with the reduced-form analysis we integrate the following exogenous idiosyncratic process for the demand fundamental $\varphi_{i,t}$:

$$\begin{array}{lcllll}
\ln\varphi_{i,t} & = & u_{i,t} + v_{i,t} + w_{i,t} + z_{i,t} & & & \\
u_{i,t} & = & \rho_u u_{i,t-1} + \theta_i, & u_{i,-1} \sim iid(\mu_{\tilde{u}}, \sigma^2_{\tilde{u}}) & \theta_i \sim iid(\mu_\theta, \sigma^2_\theta) & \rho_u \in [0,1) \\
v_{i,t} & = & \rho_v v_{i,t-1}, & v_{i,-1} \sim iid(\mu_{\tilde{v}}, \sigma^2_{\tilde{v}}) & & \rho_v \in [0,1) \\
w_{i,t} & = & \rho_w w_{i,t} + \varepsilon_{i,t}, & w_{i,-1} = 0 & \varepsilon_{i,a} \sim iid(0, \sigma^2_\varepsilon) & \rho_w \in [0,1) \\
z_{i,t} & \sim & iid(0, \sigma^2_z), & & &
\end{array}$$

where we re-introduced time indices. Note that the firm-level state is given by $\mathbf{s}_{i,t} = [u_{i,t}, v_{i,t}, w_{i,t}, z_{i,t}]$. The above process implies that the level of demand faced by a firm is determined by both a idiosyncratic ex-ante profile, captured by $u_{i,t}$ and $v_{i,t}$, as well as ex-post shocks, which enter via $w_{i,t}$ and $z_{i,t}$.

Discussion: adjustment costs and selection. Ex-post demand shocks are a standard feature of firm dynamics models, since demand conditions may change for various reasons that are beyond the control of the firm. Considering ex-ante heterogeneity across firms also has a strong tradition in the literature. While in certain models ex-ante heterogeneity across firms materialize immediately (see e.g. Melitz, 2003), other studies consider a gradual accumulation of difference, for instance through customer base accumulation (see e.g. Arkolakis, 2016; Luttmer, 2011; Drozd and Nosal, 2012; Gourio and Rudanko, 2014; Perla, 2015). While our baseline model allows only for "passive" accumulation of ex-ante differences, we consider endogenous adjustment costs in Appendix B.3.[11] Importantly, incorporating adjustment costs does not change the main results.

As in any firm dynamics model with endogenous entry and/or exit, a key channel via which heterogeneity may impact on aggregate outcomes is selection. Since we integrate a multi-dimensional idiosyncratic process into the model, selection occurs along several different competing margins. Importantly, there is no one-to-one mapping between a particular a value of demand and the survival probability of the respective firm. For example, a currently small and unprofitable startup may survive with high probability if it has sufficiently promising long-run growth potential and only faces poor initial conditions or ex-post shocks.

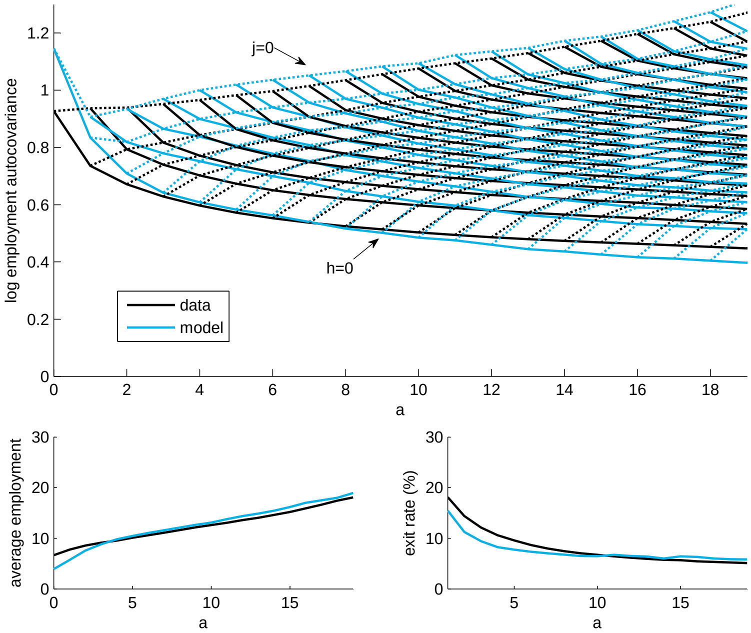
3.2 Parametrization and model fit

We now match the model to our data for firms. Before doing so, we set three parameters a priori, assuming a model period of one year, which corresponds to the frequency of our data. First, the discount factor is set to $\beta = 0.96$, which implies an annual real interest rate of about four percent. Second, we set the elasticity of substitution between goods to $\eta = 6$, which is in the ballpark of values common in the literature. Third, we set the entry cost f_e such that the ratio of the entry cost to the operational fixed cost is $f_e/f = 0.82$, following estimates of Barseghyan and Dicecio (2011).

The remaining parameters are estimated using the the Simulated Method of Mo-

[11] Our baseline model also abstracts from differences in technologies, another form of heterogeneity often considered in the firm dynamics literature. However, given that we match our model to employment data, our model is observationally equivalent to one with heterogeneity in TFP. Moreover, Hottman, Redding, and Weinstein (2016) and Foster, Haltiwanger, and Syverson (2016) have recently investigated the relative importance heterogeneity in demand versus technology. They conclude that demand factors are a major driver of heterogeneity in the data.

Figure 4: Targeted moments: data and structural model (firms)



Notes: Top panel: Autocovariances of log employment between age $a = h + j$ and age $h \leq a$ in the data and the model, for a balanced panel of firms surviving up to at least age $a = 19$. Bottom left panel: Average employment by age a (unbalanced panel). Bottom right panel: exit rate by age a.

ments. Details of the numerical solution and simulation procedure are provided in Appendix B.1. Again, we target the upper triangular of the autocovariance matrix of logged employment, by age, for a balanced panel of firms surviving up to at least age nineteen. Now, however, we also target the the age profiles of the exit rate and average employment (in an unbalanced panel). In doing so, we assume that all shock innovations are drawn from normal distributions and we normalize the level parameters μ_u and μ_v to zero. In contrast to the reduced-form setup, we further assume that $\rho_v = \rho_w$, which eases the computational burden substantially.[12]

Figure 4 illustrates the fit of the model. The upper panel shows that the model fits the autocovariance matrix very well, although it overshoots on the variance of logged employment at age zero. The lower left panel shows average employment by age. In the data, this profile is upward sloping, increasing between about 7 at age zero to 18 at

[12] This restriction reduces the number of state variables as firms no longer need to keep track of $w_{i,t}$ and $v_{i,t}$ separately. Moreover, Table 1 shows that the reduced form estimates of these persistence parameters are close to each other.

age nineteen. The model captures this pattern well, although it somewhat undershoots on the size of very young firms. The lower right panel shows the annual exit rate by age. In the data, about eighteen percent of firms exit between age zero and one. Subsequently, the exit rate gradually declines, stabilizing at older age categories. The model matches this pattern, predicting relatively high exit rates at young ages, but somewhat undershoots on the exit rates of young firms. Overall, the model provides a good fit of the three sets of empirical comments, considering that 10 parameters are used to target 249 moments.

Additionally, we consider how the model fits the employment distribution by age and size, which is not directly targeted. Figure 5 shows employment shares of different age/size bins, in the model and in the data. Overall, the model fits this distribution well. The model also provides a similarly good fit of the fractions of firms in each of these bins (not shown).

The associated parameter values are shown in Table 2. The fixed cost is estimated to be 0.54, which is about half the wage of a single employee. The exogenous exit rate is estimated to be about 4.1 percent. Thus, a substantial fraction of firms exits for reasons unrelated to their fundamentals. However, Figure 4 makes clear that there is also a substantial amount of endogenous exit, as the overall exit rate in the model varies between 15.5 percent at age zero to 5.8 percent at age nineteen.

The remaining parameters are somewhat difficult to interpret individually, especially since the parameter values are for the unconditional distributions, whereas the equilibrium distributions are truncated by selection. Below, however, we will quantify the model's implications for the importance of ex-ante heterogeneity and make a direct comparison to the reduced-form model along this dimension.

4 The importance of ex-ante versus ex-post heterogeneity

In this section we use the structural model to study the importance of ex-ante heterogeneity for a number of outcomes highlighted in the literature. We first quantify its importance for cross-sectional dispersion in firm size, as we did in the reduced-form

Table 2: Parameter values (firms)

parameter		value
		set a priori
β	discount factor	0.96
η	elasticity of substitution	6.00
f^e	entry cost	0.44
		estimated
f	fixed cost of operation	0.539
δ	exogenous exit rate	0.041
μ_θ	permanent component θ, mean	-1.762
σ_θ	permanent component θ, st. dev.	1.304
$\sigma_{\widetilde{u}}$	initial condition u_{-1}, st. dev.	1.572
$\sigma_{\widetilde{v}}$	initial condition v_{-1}, st. dev.	1.208
σ_ϵ	transitory shock ϵ, st. dev.	0.307
σ_z	noise shock z, st. dev.	0.203
ρ_u	permanent component, persistence	0.393
ρ_v	transitory component, persistence	0.988

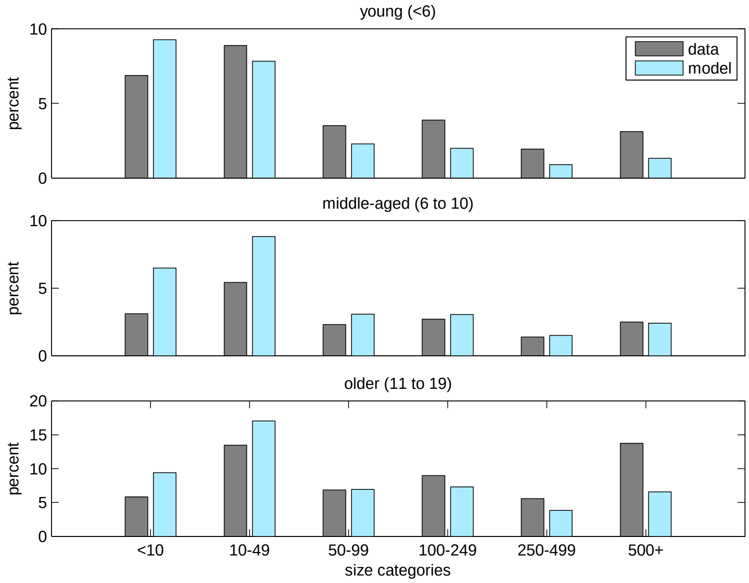
Notes: parameter values. Top three parameters are calibrated as discussed in the main text. The remaining parameters are set such that the model matches the empirical autocovariance of employment and the age profiles of average size and exit rates from age 0 to 19.

model. However, an advantage of the structural model is that it accounts for endogenous selection of firms, which potentially has an impact on this quantification.

Next, we study "up-or-out dynamics", the phenomenon that low-growth firms tend to exit, whereas surviving firms tend to grow quickly. In the literature, such dynamics have been emphasized as a sign of a well-functioning economy and as an important contributor to aggregate output and productivity, see for example Haltiwanger, Jarmin, and Miranda (2013). We use our model to examine the sources of up-or-out dynamics, by quantifying the importance of ex-ante heterogeneity for the age profiles of the exit rate and of average size. This helps us understand whether up-or-out dynamics should be thought of as a process which sifts out firms with high ex-ante growth potential, or as one that reflects the idiosyncratic risk that firms face after they enter.

Finally, we evaluate aggregate outcomes, and in particular the aggregate gains that result from up-or-out dynamics.

Figure 5: Employment shares of different age/size bins: model versus data (firms)



Notes: Employment shares by firm age and size (employment). Values are expressed as percentages of total employment in firms between 0 to 19 year old firms, both in the data and the model. Data are obtained from the Business Dynamics Statistics, an aggregated and publicly available version of the LBD over the corresponding time period.

4.1 Cross-sectional dispersion in employment

We first revisit the importance of ex-ante heterogeneity for the cross-sectional dispersion in employment, conditional on age. Defining $\chi \equiv ((\eta-1)/\eta)^{\eta} w^{-\eta} Y$, the employment level of firm i can be expressed as:

$$n_i = \chi \varphi_i^{EXA} \varphi_i^{EXP}, \tag{4}$$

where $\varphi_i^{EXA} = e^{u_i+v_i}$ is the ex-ante component of demand and $\varphi_i^{EXP} = e^{w_i+z_i}$ is the ex-post component. As in the reduced-form exercise, we can now compute the contribution of ex-ante heterogeneity to the cross-sectional variance of employment by shutting down variation in φ_i^{EXP}. In contrast to the reduced-form model, however, the ex-ante and ex-post component are no longer orthogonal, due to endogenous selection which tends to induce a negative correlation between the two. This occurs because firms with relatively poor ex-ante conditions can survive only if they were exposed to favorable ex-post shocks and vice versa. Accounting for this correlation, we decompose the variance of logged employment as:

$$\begin{aligned} Var\left(\ln n_i\right) &= Var(\ln \varphi_i^{EXA}) + Var(\ln \varphi_i^{EXP}) + 2Cov(\ln \varphi_i^{EXA}, \ln \varphi_i^{EXP}), \\ &= Cov(\ln \varphi_i^{EXA}, \ln n_i) + Cov(\ln \varphi_i^{EXP}, \ln n_i). \end{aligned} \tag{5}$$

In the reduced-form model, the covariance term $Cov(\ln \varphi_i^{EXA}, \ln \varphi_i^{EXP})$ in the first equality is zero, due to the assumption of independently distributed shocks. However, in the structural model selection induces a non-zero covariance term which, as mentioned above, tends to be negative. We therefore decompose the variance according to the second equality in Equation (5).[13] Figure 6 depicts the contribution of ex-ante heterogeneity in the structural model (solid line), i.e. $Cov(\ln \varphi_i^{EXA}, \ln n_i)/Var\left(\ln n_i\right)$, together with the reduced-form decomposition (dashed line). Both decompositions attribute a similarly large fraction of size dispersion to ex-ante heterogeneity, at any age.

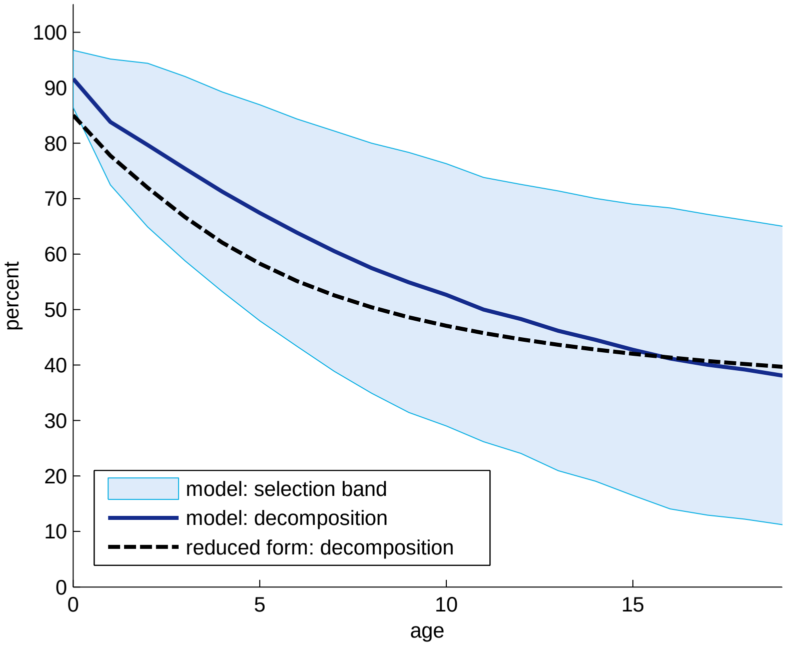
Figure 6 also plots a "selection band" based on the first equality in Equation (5). This band is constructed by attributing, in turn, the covariance term either fully to the ex-ante component or fully to the ex-post component. This gives us a sense of how much selection matters in the model. The widening band indicates that selection has an increasingly important impact on the cross-sectional dispersion of firm size as firms age. Overall, however, the various decompositions re-establish our earlier conclusion that ex-ante heterogeneity is a key source of size dispersion, in particular among younger firms.

4.2 Firm exit

Next, we study the importance of ex-ante heterogeneity for exit, the "out" part of up-or-out dynamics. One might think that exit is entirely triggered by unexpected ex-post

[13] Note that when $Cov(\ln \varphi_i^{EXA}, \ln \varphi_i^{EXP}) = 0$, it holds that $Var(\ln \varphi_i^{EXA}) = Cov(\ln \varphi_i^{EXA}, \ln n_i)$ and $Var(\ln \varphi_i^{EXP}) = Cov(\ln \varphi_i^{EXP}, \ln n_i)$. The decomposition then exactly coincides with the one we used in the reduced-form analysis.

Figure 6: Contribution of ex-ante heterogeneity to cross-sectional employment dispersion



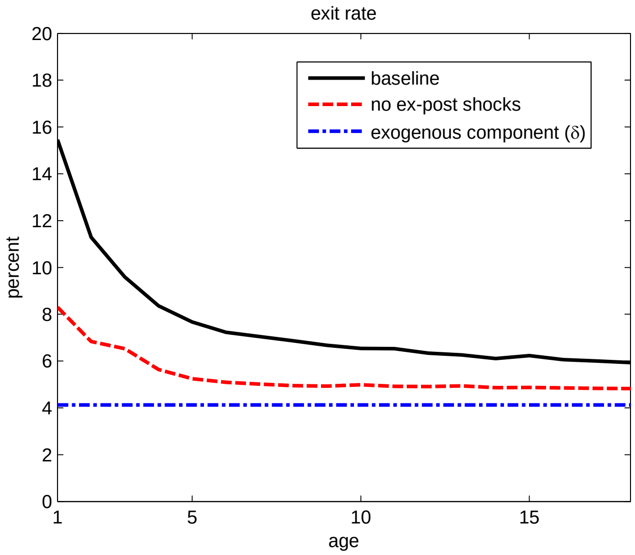
Note: Contributions to total cross-sectional variance by age. "Reduced-form" refers to the estimates from Figure 3, "model: covariance decomposition" is the decomposition based on the second line in Equation 5. The shaded areas ("model: selection band") is constructed based on the first equality in Equation 5 by attributing, in turn, the term $2Cov(\ln\varphi_i^{EXA}, \ln\varphi_i^{EXP})$ fully to the ex-ante component and to the ex-post component.

shocks. However, exit might also be the result of ex-ante heterogeneity. For example, if a firm has an ex-ante demand profile that is downward sloping in age, it may be economically viable in the initial years, but not later on. Hence, the firm would exit at some point even without ex-post shocks.

To quantify the importance of ex-ante heterogeneity for exit, we run a counterfactual simulation in which we use the firms' baseline decision rules but we completely shut down ex-post shocks to demand, while preserving exogenous exit shocks. Figure 7 shows the age profile of the exit rate in this counterfactual, together with the exit profile in the baseline model, and the exogenous component of the exit rate, δ. The difference between the latter two is the endogenous component of the exit rate, i.e. the part that is driven by selection.

As expected, the exit rate is lower without ex-post demand shocks. However, there is

Figure 7: Exit rates



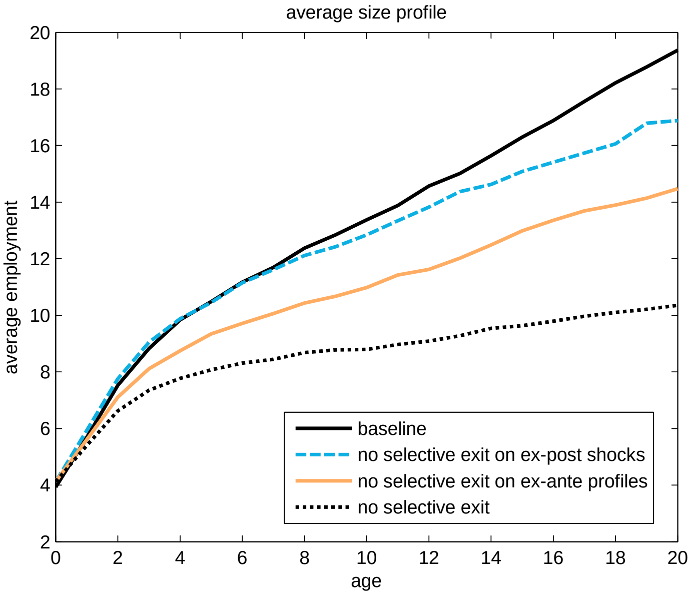
Note: exit rates by age in the baseline model, exit rates in and a counterfactual economy in with no ex-post demand shocks (but with exogenous exit), and the exogenous exit rate δ.

still selection and the exit rate in the counterfactual declines with age, as in the baseline. We can interpret the difference between the exit rate without ex-post shocks and the exogenous exit rate δ as the amount of endogenous exit that is driven by selection on ex-ante profiles. Figure 7 then implies that, depending on age, between 30 and 45 percent of overall endogenous exit is driven by selection on ex-ante profiles. Thus, we find that ex-ante heterogeneity is an important contributor to firm exit.

4.3 Average size

We now turn to the importance of ex-ante heterogeneity for the “up” part of up-or-out dynamics. In particular, we consider the age profile of average size. The impact of heterogeneity on this average size profile materializes via selection: if small and low-growth firms are more likely to exit the economy, then this increases the average size of the remaining firms. However, selection is a multifaceted process which is not only affected by a firm’s current fundamentals but also by expectations about how fundamentals will evolve in the future. This evolution is in turn driven by both the ex-ante profile and by ex-post shocks.

Figure 8: Average size and selection



Note: Average size, unbalanced panel and by age, in the baseline in three counterfactuals. See the main text for a description of these counterfactuals.

To examine the effect of selection on the age profile of average size, we conduct three counterfactuals based on Equation (4), which expresses firm-level employment as a function of the firm's demand fundamentals. In the first counterfactual, we shut down selective exit on the ex-post component φ_i^{EXP}. We do so by considering the stationary distribution of firms in the baseline model, but re-draw for each firm φ_i^{EXP} randomly, but conditional on age, from the distribution of ex-post components that would be obtained in the absence of exit. In the second counterfactual we shut down selection on the ex-ante component φ_i^{EXA}. Again, we consider the stationary distribution in the baseline, but now redraw the φ_i^{EXA} from the distribution without exit, leaving the baseline ex-post component intact. In the third counterfactual, we shut down all selective exit altogether by jointly redrawing both φ_i^{EXA} and φ_i^{EXP} from the distribution without exit.

Figure 8 shows the average size profile in the baseline and the three counterfactuals. As expected, shutting down margins of selection generally reduces average size. However, shutting down selective exit on the ex-post component has a relatively small impact on the average size profile, which differs from the baseline only after age seven. By contrast, selection on the ex-ante component has a much larger dampening effect on the average size profile. Moreover, the gap with the baseline arises already in the early years following entry. Finally, shutting down selective exit altogether has the largest impact on the average size profile, lowering average by almost 50 percent by age twenty. This gap indicates that there is a large interaction associated with joint selection on the ex-ante and the ex-post component.

We thus find that ex-ante heterogeneity is not only an important driver of dispersion in size, but also of the age profiles of exit and average size, especially among younger firms. Thus, up-or-out dynamics largely reflect the separation of firms with high and low long-run growth potential. An important driver of differences in up-or-out dynamics across countries or different time periods within a country might therefore reflect differences in the types of startups that enter the economy. We will return to this issue in the next section.

4.4 Aggregate output

We now explore some the aggregate implications of our findings. For this purpose we use the same counterfactuals, in combination with the following expression for aggregate output:

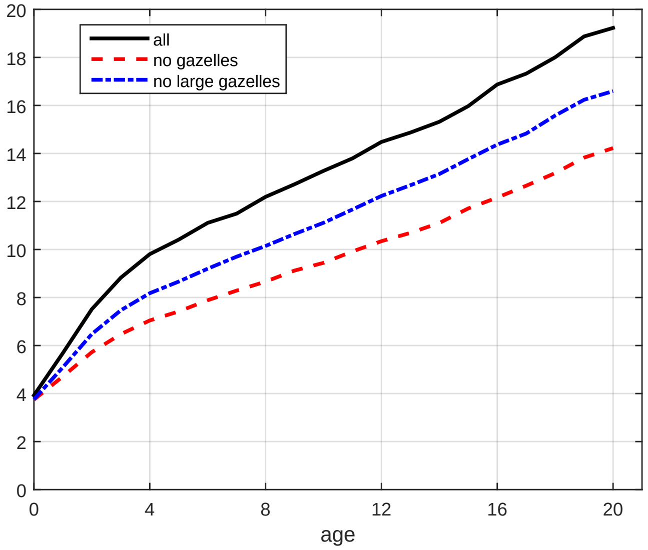
$$Y = \Omega^{\frac{\eta}{\eta-1}} \chi^{\frac{1}{1-\eta}} \overline{n}^{\frac{\eta}{\eta-1}}$$

where $\overline{n}$ is the average size across all firms; see Appendix B.1 for a derivation. We recompute average firm size in each of the three counterfactuals described in the previous subsection and then compute aggregate output based on the above equation.[14] We find that without selective exit on ex-post shocks, output is about 4 percent lower than in the baseline. In the second counterfactual, in which selection on the ex-ante component is shut down, output is about 15 percent lower than in the baseline. Shutting down selective exit altogether, output is 38 percent lower than in the baseline.

These results imply that up-or-out dynamics are indeed an important contributor

[14] These are partial-equilibrium counterfactuals since we do not recompute χ and Ω.

Figure 9: The importance of high-potential startups



Note: Average size, unbalanced panel and by age, in the baseline in two counterfactuals. See the main text for a description of these counterfactuals.

to aggregate output. Moreover, a key factor driving these dynamics is selection based on firms' ex-ante growth profiles, as well as interaction between ex-ante profiles and ex-post shocks. By contrast, ex-post shocks alone matter relatively little, especially at younger ages. Note further that our counterfactual exercises are based on distributions conditional on firm entry, i.e. based on demand fundamentals of firms which have already *decided* to begin operating. The impact of ex-ante heterogeneity would likely be even larger if selection before entry were to be included in the counterfactuals.

4.5 The importance of high-potential startups

Our estimates show a large amount of heterogeneity in ex-ante profiles: some high-potential startups are on steep ex-ante age profiles of demand growth, whereas others are on flat or even downward-sloping age profiles.

We now quantify the importance of high-growth startups. Such firms, labeled "gazelles" since Birch and Medoff (1994), have been emphasized in the literature as important engines of aggregate job creation. We classify firms according to their ex-ante growth profiles, i.e. the individual age profiles of size that firms would follow in the absence of ex-post shocks. We then define gazelles as those startups with an ex-ante projected growth rate of at least 15 percent annually, over the first five years, and an associated employment level that exceeds 10 workers at some point during their lifetimes.[15]

While our definition of gazelles is in line with the literature, we differ from existing studies in an important way: we classify firms according to their *ex-ante* profiles at startup. By contrast, the existing literature has classified firms based on ex-post realizations, since ex-ante profiles are not directly observed in the data. Thus, the gazelles as defined in the literature include firms which at startup were not expected to grow very much, but ex post were hit by positive shocks and grew as a result. It then follows almost by definition that gazelles contribute disproportionately to aggregate job creation. By contrast, in our definition firms that grow just because of favorable ex-post shocks are not counted as gazelles. A priori, it then becomes less clear that gazelles will contribute disproportionately to job creation.

Having classified firms on an ex-ante basis, we re-compute the average size profile leaving out the gazelles, see Figure 9. Without gazelles, average size is considerably smaller and the difference remains large up to at least age twenty. At that age, average size is more than 25 percent lower than in the baseline. This difference is striking when compared to the roughly 8 percent share of gazelles among startups. In a second counterfactual we leave out only "large gazelles", which are defined as gazelles with a startup size of at least 10 workers. In this counterfactual, average size is about 15 percent lower at age twenty than in the baseline, even though large gazelles account for about 1 percent of all startups.

These counterfactuals make clear that high-potential startups are indeed important contributors to aggregate output and employment. Moreover, it follows that seemingly small shifts in the distribution of ex-ante profiles of startups may have large consequences, as suggested also by Sedláček and Sterk (2017). Our results further provide a perspective on the findings of Hsieh and Klenow (2014), who report that average size

[15] Defining gazelles using not only growth rates but also size excludes firms which grow quickly but nevertheless always stay small.

profiles are much flatter in India and Mexico than in the United States. A flat profile can indicate that there are few startups that operate a high-potential business model, or that high-potential startups have relatively low chances of survival.

5 Changes in the firm dynamics process

Finally, we investigate how firm dynamics have changed over time. Such changes have attracted attention in the light of the disappointing evolution of employment and productivity growth in the US over the last ten to fifteen years. A disconcerting trend that has been witnessed over the same period is that the skewness of firm growth rates has declined, suggesting that high-growth firms are becoming increasingly rare, see Decker, Haltiwanger, Jarmin, and Miranda (2016).

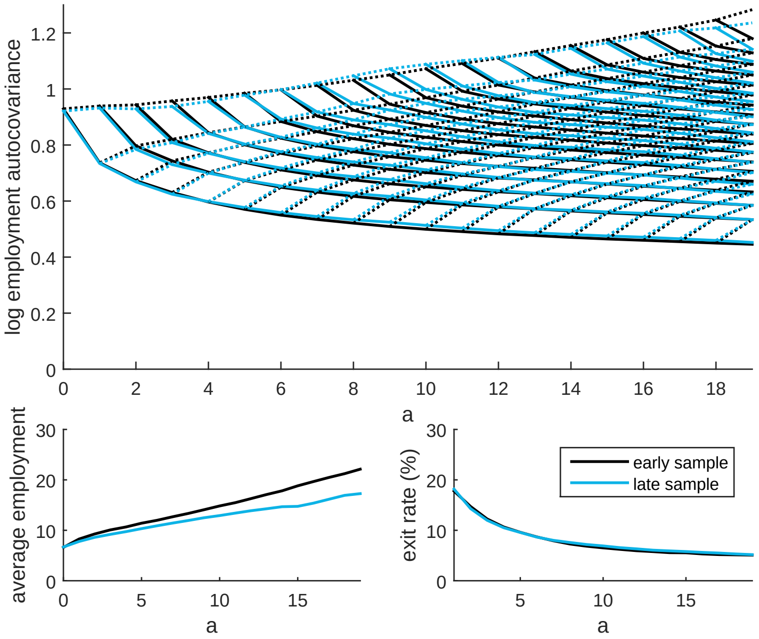
We analyze the changes in firm dynamics by splitting our data into an early sample, including firms born between 19XX and 19XX, and a late sample with firms born between 19XX and 19XX. We first document changes in the three sets of key moments, the autocovariance function, the average size profile, and the exit profile. Next, we re-estimate our model on the split sample and interpret the changes in the data through the lens of our model. In particular, we study whether these patterns were driven by a change in ex-post shocks or whether the fraction of ex-ante "gazelles" among startups has changed over time and how this affected aggregate outcomes.

5.1 Changes in the data

Figure 10 plots the three sets of key moments in the two samples. The top panel shows that the autocovariance function of logged employment of firms (balanced panel), which has remained remarkably stable over time. This suggests that the relative importance of ex-ante and ex-post heterogeneity has not change much over time. The bottom right panel shows that exit rates have also remained stable across the two samples, see also Pugsley and Şahin (2016).

What has changed, however, is the profile of average size by age, which is shown in the bottom left panel of Figure 10. Over time, this profile has flattened. At startup,

Figure 10: Split-sample data moments



Notes: Top panel: Autocovariances of log employment between age $a = h + j$ and age $h \leq a$ in the early and the late sample, for a balanced panel of firms surviving up to at least age $a = 19$. Bottom left panel: Average employment by age a (unbalanced panel). Bottom right panel: exit rate by age a.

average size is about 7 employees in both the early and the late sample. However, by age nineteen, average employment has declined by almost 25 percent from an average 22 workers in the early sample to 17 employees in the late sample. In addition, this divergence in size profiles seems to set in gradually and not occur only for old firms.

Finally, recall that that the exit profile is predominantly driven by firms at the bottom of the distribution, i.e. those with low growth potential. The fact that the age profile of average size has decline but that of exit rates remained stable across the two subsamples suggests that changes in the firm distribution have taken place not at the bottom but at the top of the growth profile distribution, where the gazelles are located.

To better understand the flattening of the age profile, we consider in more detail how it occurred. Figure 11 plots average firm size in different five-year age bins.[16] The left panel plots these by year of observation. The figure clearly shows the decline in

[16] The figure uses the Business Dynamics Statistics data, which is a publicly available aggregated version of the underlying LBD data set used in our estimations.

average size among older firms. However, it also shows that the decline occurred in a staggered way, taking place later in older age bins. In particular, the average size time paths of the three oldest age categories clearly move in lock-step with five year gaps between them. This also makes clear that the flattening of the average size profile was set into motion before the Great Recession. Finally, note that average size declined also for firms $0-5$ and $6-10$ years of age by 3 and 15 percent, respectively.[17]

The right panel of Figure 11 plots the same data but now by cohort defined by the birth year of the youngest firms in each age category. The figure shows very clearly that the flattening occurred by cohort. In addition, this change was not gradual, but it happened rather abruptly around the mid 1980's. Specifically, cohorts born since the late 1980's had a much flatter average size profile compared to cohorts of firms born earlier. These changes gradually fed through the economy as more cohorts with lower growth potential came into existence. This link to different cohorts of firms suggests that the flattening of the (aggregate) average size profile was an ex-ante phenomenon, rather than the result of changes in the character of ex-post shocks that would affect *all* firms.

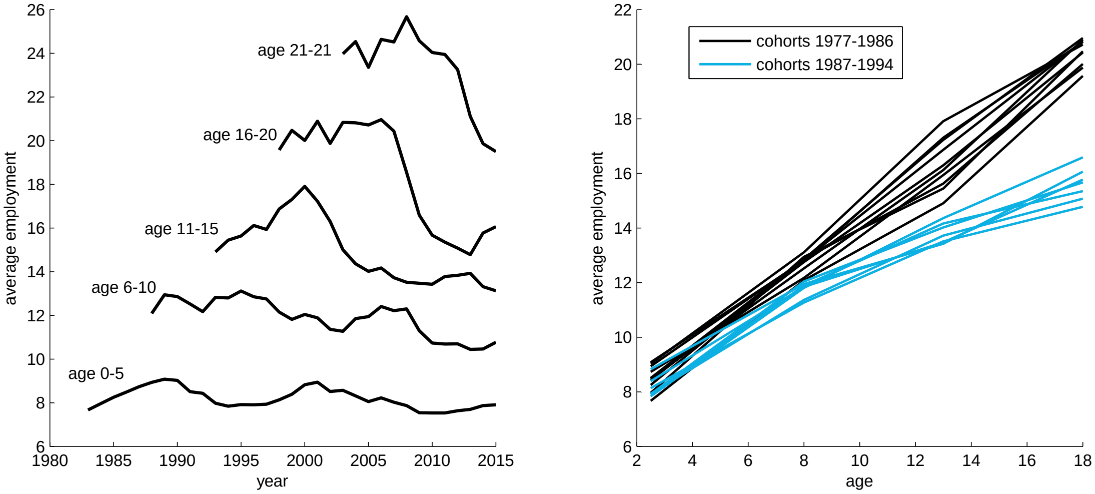
5.2 Are gazelles dying out?

Our previous analysis suggests that the flattening of the average size profile might be related to ex-ante characteristics of startups. To investigate the underlying changes more directly, we re-estimate the model on the two subsamples. The parameter values and model fit are shown in Appendix B.2.

Within the two estimated models, we compute the fraction of gazelles in the population of firms, by age. This is shown in the left top panel of Figure 12. Among startups, the fraction of gazelles has declined by 17 percent from a share of 6.4 percent in the early sample to 5.3 percent in the late sample. As firms age, the fraction of gazelles increases because gazelles are relatively unlikely to shut down compared to other firms with lower growth potential. Therefore, the gap in the share of gazelles widens with age

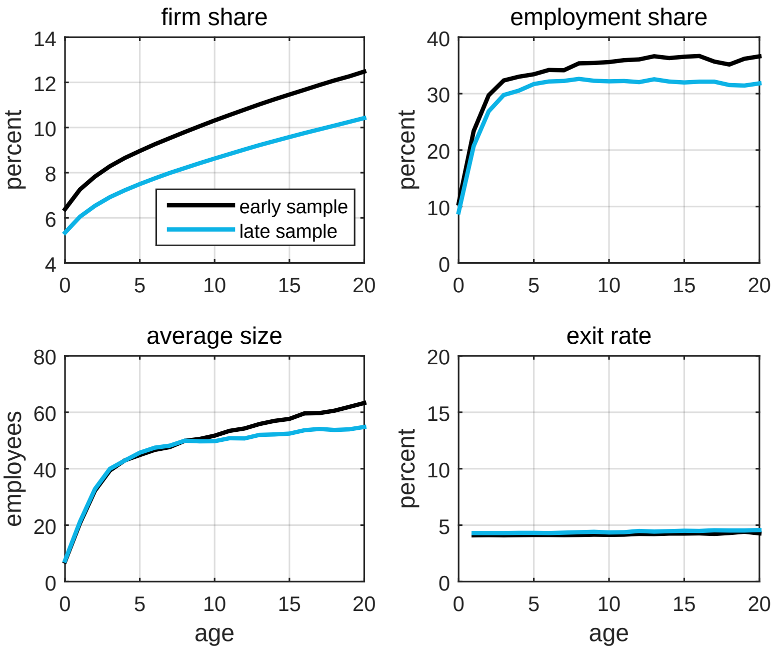
[17] These values are based on averages in the first and last 15 years of the sample period.

Figure 11: Flattening of the average size profile in the data



Note: The left panel plots, by year, average firm employment in different age bins: 0-5 years, 6-10 years, 11-15 years, 16-20 years, and 21-25 years. The right panel plots the same data, but now by overlapping 5-year cohorts, grouped by the birth year of the youngest firm in each cohort. Source: Business Dynamics Statistics.

Figure 12: Characteristics of gazelles in the early and late sample



Note: Top panels: share of gazelles in the total number of firms and in total employment. Bottom panels: average size and exit rate profile of gazelles. Gazelles are classified on an ex-ante basis, as those startups with an ex-ante growth rate of at least 15 percent annually, over the first five years, and an associated employment level that exceeds 10 at some point during this period.

between the two samples. At age twenty, the fraction of gazelles is 12.5 percent in the early sample but only 10.4 in the late sample.

A similar picture is painted by the top right panel which shows the employment shares, by age. Among startups at age zero, gazelles account for around 9 percent of employment in both the early and the late samples. Again, however, a gap emerges between the two samples as firms age and start fulfilling their ex-ante growth potential.

The bottom left panel shows the average size profile of gazelles. In both sub-samples, gazelles start with around 7 employees, but grow quickly to reach on average about 46 employees by age five. Around age 10, however, the two sub-samples diverge, and a reduction in the average size between the two sub-samples becomes apparent. Thus, in the late sample gazelles on average do not grow as large as in the early sample. Finally, the exit profile, plotted in the bottom right panel, is essentially the same in both samples, as gazelles exit practically only for exogenous reasons.

Our findings thus confirm the concerns that high-growth firms are becoming increasingly rare. While Decker, Haltiwanger, Jarmin, and Miranda (2016) document that the decline in the skewness of firm growth rates occurred around 2000 and primarily in the Services, Information and High-tech sectors, the sources of these secular changes remain to be identified. While our framework does not provide a definitive answer to this question, it does offer additional new insights. First, we document that the disappearance of gazelles is related to ex-ante factors, suggesting that high-growth firms are in fact dying out. Second, not only are there fewer gazelles, but those that nevertheless start up tend to expand less than high-growth firms of the past. Third, our results suggest that the decline in average size and the disappearing skewness of growth rates was set in motion already in the late 1980's, as opposed to the early 2000's when the patterns became apparent.

The above insights point to potential future avenues of research attempting to identify the reasons behind the disappearance of gazelles. For instance, while many existing studies focus on how firms operate in the economy, much less is known about which individuals start businesses and what type of firms are founded.[18] Alternatively, an intriguing connection may be made between the demise of gazelle startups and the decline in the aggregate labor share of income, which also started in the late 1980's. For example, Autor, Dorn, Katz, Patterson, and Reenen (2017) suggest that the decline in the labor share was due to an increase in product market concentration, giving rise to "superstar firms". Increased domination of incumbent superstar firms might have made it more difficult for high-potential startups to enter the economy. Or vice versa, a lack of competitive pressure from gazelle startups might have contributed to the increase in market concentration. These and other research questions may contribute to our understanding of why high-growth startups are becoming increasingly rare.

5.3 Aggregate implications

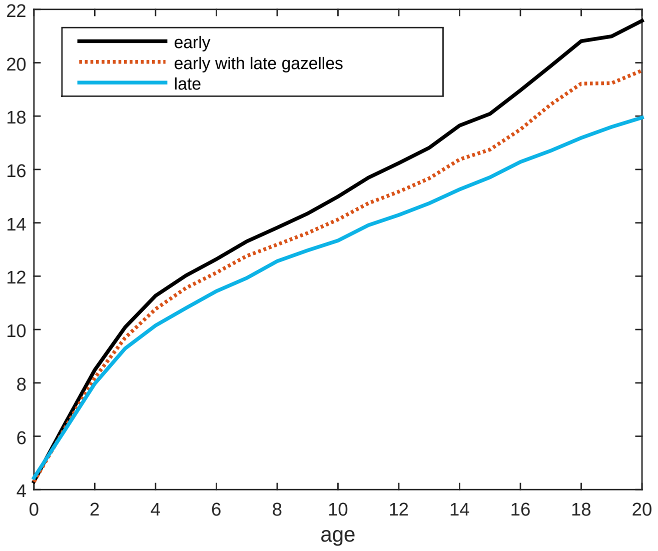
We now explore some aggregate implications of our findings. Figure 13 plots the average size profile, in the estimated model over the two sub-samples. As noted before, this profile has flattened. To assess the contribution of disappearing gazelles to this shift, we conduct a simple counterfactual exercise. In particular, we note that at any age, the average size among all firms is the mean of the average size of gazelles and non-gazelles, weighted by their respective firm share. We then construct a counterfactual in which we re-compute the average size in the early sample, but with the average size and firm share profiles of the gazelles in the late sample.

The dashed line in Figure 12 plots this counterfactual. It shows that the change in the fraction of gazelles and their average size profile accounts for roughly half of the decline in the average size profile. This is remarkable, given that gazelles account for only about five percent of the startups.

Finally, we evaluate the aggregate implications of the overall shift in the firm growth process. We find that between the two samples, aggregate output declines by 4.5 percent. Thus, seemingly small changes in the distribution of firms, such as the decline in the (already low) share of high-potential startups as well as a reduction in their growth

[18] See Guzman and Stern (2015) and Belenzon, Chatterji, and Daley (2017) for evidence on how startup characteristics are informative about future firm growth.

Figure 13: The impact of disappearing gazelles



Note: The figure plots the average size profile among all firms in the early sample and the late sample. It also plots a counterfactual average size profile for the early sample, computed by replacing firm share and average size profile of gazelles by their counterparts from the late sample.

potential, emerge as important drivers of *aggregate* changes.[19,20]

6 Conclusions

We have used data on the population of U.S. firms over several decades to better understand why startups grow rapidly whereas others remain stagnant or exit quickly. To this end, we documented the autocovariance structure of employment and exploited this structure to estimate firm dynamics models, which allowed us to disentangle heterogeneous ex-ante profiles from ex-post shocks. We found a dominant role for heterogeneous ex-ante profiles, which capture unrealized potential present at the moment of startup. Most of the dispersion in firm size, at a given age, is driven by such ex-ante potential. Moreover, we found that that ex-ante heterogeneity also drives much of the "up-or-out" dynamics observed in the data: high-potential firms, "gazelles", grow quickly and sur-

[19] Within the model, this decline is entirely driven by a change in output per worker, i.e. labor productivity, since we keep labor supply fixed. In a model version with endogenous labor supply, there could be an associated decline in aggregate employment as well.

[20] Shifts in the number of startups may also have important macroeconomic consequences, see Sedláček (2015).

vive at high rates, whereas low-potential firms tend to exit quickly. These dynamics lead to substantial gains in aggregate output.

We have also investigated potential changes in the firm dynamics process, following up on recent concerns that high-growth firms are disappearing. We documented a dramatic flattening of the age profile of average size, among cohorts of firms born since the late 1980's. Re-estimating the model using this information, we found a decline in the presence of high-potential "gazelles" in the population of startups, with important repercussions for aggregate output.

Our results highlight the need for future research on which individuals become entrepreneurs and what decisions such aspiring entrepreneurs make *before or at* startup, as opposed to their behavior *after* the firm has become operational. While the macroeconomic implications of the latter have been studied extensively in the literature, much less is known about how institutional conditions change who becomes an entrepreneur and what types of firms are being created. Our results show that such changes can be of first-order importance for macroeconomic outcomes.

References

ABBRING, J., AND J. CAMPBELL (2005): "A Firm's First Year," Tinbergen Institute Discussion Paper 05-046/3.

ABOWD, J., AND D. CARD (1989): "On the Covariance Structure of Earnings and Hours Changes," *Econometrica*, 57(2), 4111–445.

AKCIGIT, U., H. ALP, AND M. PETERS (2017): "Lack of Selection and Limits to Delegation: Firm Dynamics in Developing Countries," NBER Working Paper 21905.

ARELLANO, M., AND S. BOND (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and Application to Employment Equations," *Review of Economic Studies*, 58(2), 277–297.

Arkolakis, C. (2016): “A Unified Theory of Firm Selection and Growth,” *Quarterly Journal of Economics*, 131(1), 89–155.

Arkolakis, C., T. Papageorgiou, and O. Timoshenko (forthcoming): “Firm Learning and Growth,” .

Autor, D., D. Dorn, L. F. Katz, C. Patterson, and J. V. Reenen (2017): “The Fall of the Labor Share and the Rise of Superstar Firms,” Working paper.

Barseghyan, L., and R. Dicecio (2011): “Entry Costs, Industry Structure, and Cross-Country Income and TFP Differences,” *Journal of Economic Theory*, 146, 1828–1851.

Belenzon, S., A. Chatterji, and B. Daley (2017): “Eponymous Entrepreneurs,” *American Economic Review*, 107(6), 1638–1655.

Birch, D., and J. Medoff (1994): “Gazelles,” in *Labor Markets, Employment Policy and Job Creation*, ed. by L. Solmon, and A. Levenson, pp. 159–168. Boulder: Westview Press.

Cabral, L., and J. Mata (2003): “On the Evolution of the Firm Size Distribution: Facts and Theory,” *American Economic Review*, 93(4), 1075–1090.

Campbell, J., and M. De Nardi (2009): “A Conversation with 590 Nascent Entrepreneurs,” *Annals of Finance*, 5(3), 313–340.

Chamberlain, G. (1984): “Panel Data,” in *Handbook of Econometrics, Volume II*, ed. by Z. Griliches, chap. 22, pp. 1274–1318. Elsevier Science, Amsterdam, North-Holland.

Decker, R., J. Haltiwanger, R. Jarmin, and J. Miranda (2016): “Where Has All the Skewness Gone? The Decline in High-Growth (Young) Firms in the U.S.,” *European Economic Review*, 86, 4–23.

Drozd, L., and J. Nosal (2012): “Understanding International Prices: Customers as Capital,” *American Economic Review*, 102(1), 364–395.

Foster, L., J. Haltiwanger, and C. Syverson (2016): "The Slow Growth of New Plants: Learning about Demand?," *Economica*, 83(329), 91–129.

Gourio, F., and L. Rudanko (2014): "Customer Capital," *Review of Economic Studies*, 81(3), 1102–1136.

Guvenen, F. (2009): "An Empirical Investigation of Labor Income Processes," *Review of Economic Dynamics*, 12(1), 58–79.

Guvenen, F., and A. Smith (2014): "Inferring Labor Income Risk and Partial Insurance from Economic Choices," *Econometrica*, 82(6), 2085–2129.

Guzman, J., and S. Stern (2015): "Where is Silicon Valley?," *Science*, 347(6222), 606–609.

Haltiwanger, J., R. Jarmin, R. Kulick, and J. Miranda (2016): "High Growth Young Firms: Contribution to Job, Output and Productivity Growth," US Census Bureau Center for Economic Studies Paper No. CES-WP-16-49.

Haltiwanger, J., R. Jarmin, and J. Miranda (2013): "Who Creates Jobs? Small vs. Large vs. Young," *The Review of Economics and Statistics*, 45(2), 347–361.

Hopenhayn, H. (1992): "Entry, Exit and Firm Dynamics Long Run Equilibrium," *Econometrica*, 60(5), 1127–1150.

Hopenhayn, H., and R. Rogerson (1993): "Job Turnover and Policy Evaluation: A General Equilibrium Analysis," *Journal of Political Economy*, 101(5), 915–938.

Hottman, C., S. Redding, and D. Weinstein (2016): "Quantifying the Sources of Firm Heterogeneity," *Quarterly Journal of Economics*, 131(3), 1291–1364.

Hsieh, C.-T., and P. J. Klenow (2009): "Misallocation and Manufacturing TFP in China and India," *Quarterly Journal of Economics*, 4, 1403–1448.

——— (2014): "The Life Cycle of Plants in India and Mexico," *Quarterly Journal of Economics*, 129, 1035–1084.

Hurst, E., and B. Pugsley (2011): “What Do Small Businesses Do?,” *Brookings Papers on Economic Activity*, 2, 73–118.

Klette, T. J., and S. Kortum (2004): “Innovating Firms and Aggregate Innovation,” *Journal of Political Economy*, 112.

Lee, Y., and T. Mukoyama (2015): “Productivity and Employment Dynamics of US Manufacturing Plants,” *Economics Letters*, 136, 190–193.

Luttmer, E. (2007): “Selection, Growth, And the Size Distribution of Firms,” *Quarterly Journal of Economics*, 122(3), 1103–1144.

——— (2011): “On the Mechanics of Firm Growth,” *Review of Economic Studies*, 78, 1042–1068.

MaCurdy, T. (1982): “The Use of Time-series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis,” *Journal of Econometrics*, 18, 83–114.

Melitz, M. J. (2003): “The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity,” *Econometrica*, 71(6), 1695–1725.

Perla, J. (2015): “Product Awareness, Industry Life Cycles, and Aggregate Profits,” mimeo.

Pugsley, B. W., and A. Şahin (2016): “Grown-up Business Cycles,” working paper.

Rouwenhorst, K. G. (1995): “Asset Pricing Implications of Equilibrium Business Cycle Models,” in *Frontiers of Business Cycle Research*, ed. by T. Cooley, pp. 294–330. Princeton University Press.

Sedláček, P. (2015): “Lost Generations of Firms and Aggregate Labor Market Dynamics,” Working paper.

Sedláček, P., and V. Sterk (2017): “The Growth Potential of Startups Over the Business Cycle,” *American Economic Review*, 107(10), 3182–3210.

Appendix

A Reduced-form model

A.1 Derivation of the Autocovariance formula

Consider the employment process given in Section 2.3 in the main text. Note that we can write the components as:

$$
\begin{aligned}
u_{i,a} &= \rho_u^{a+1} u_{i,-1} + \sum_{k=0}^{a} \rho_u^k \theta_i \\
v_{i,a} &= \rho_v^{a+1} v_{i,-1} \\
w_{i,a} &= \sum_{k=0}^{a} \rho_w^k \varepsilon_{i,a-k} = \rho_w^j w_{i,a-j} + \sum_{k=a-j+1}^{a} \rho_w^{a-k} \varepsilon_{i,k}
\end{aligned}
$$

Using this, the level of employment of firm i at age a can be written as:

$$
\begin{aligned}
\ln n_{i,a} &= \rho_u^{a+1} u_{i,-1} + \sum_{k=0}^{a} \rho_u^k \theta_i + \rho_v^{a+1} v_{i,-1} + \sum_{k=0}^{a} \rho_w^k \varepsilon_{i,a-k} + z_{i,a} \\
&= \rho_u^{a+1} u_{i,-1} + \sum_{k=0}^{a} \rho_u^k \theta_i + \rho_v^{a+1} v_{i,-1} + \rho_w^j w_{i,a-j} + \sum_{k=a-j+1}^{a} \rho_w^{a-k} \varepsilon_{i,k} + z_{i,a}
\end{aligned}
$$

We can now write the autocovariance as:

$$
\begin{aligned}
& Cov\left(\ln n_{i,a}, \ln n_{i,a-j}\right) \\
= \; & \rho_u^{2(a+1)-j} \sigma_{\tilde{u}}^2 + \rho_v^{2(a+1)-j} \sigma_{\tilde{v}}^2 + \left(\sum_{k=0}^{a} \rho_u^k\right)\left(\sum_{k=0}^{a-j} \rho_u^k\right) \sigma_\theta^2 + \rho_w^j Var\left(w_{i,a-j}\right) + 0^j \\
= \; & \rho_u^{2(a+1)-j} \sigma_{\tilde{u}}^2 + \rho_v^{2(a+1)-j} \sigma_{\tilde{v}}^2 + \left(\sum_{k=0}^{a} \rho_u^k\right)\left(\sum_{k=0}^{a-j} \rho_u^k\right) \sigma_\theta^2 + \rho_w^j \sum_{k=0}^{a-j} \rho_w^{2k} \sigma_\varepsilon^2 + 0^j \\
= \; & \rho_u^{2(a+1)-j} \sigma_{\tilde{u}}^2 + \rho_v^{2(a+1)-j} \sigma_{\tilde{v}}^2 + \left(1-\rho_u^{a+1}\right)\left(1-\rho_u^{a-j+1}\right) \frac{\sigma_\theta^2}{\left(1-\rho_u\right)^2} + \rho_w^j \frac{1-\rho_w^{-2(a-j+1)}}{1-\rho_w^2} \sigma_\varepsilon^2 + 0^j
\end{aligned}
$$

This gives Equation (2) in the main text.

A.2 Estimation details

The reduced-form model is estimated using a minimum distance procedure, following Chamberlain (1984). Let ϑ be an arbitrary parameter vector in compact parameter space. Since we use ages $a = 0, \ldots, 20$, we define the $K = \frac{21\cdot(21+1)}{2}$-length vector function, for any arbitrary observation i:

$$f\left(n_i, \vartheta\right) = \left[\left(\ln n_{i,a} - E\left[\ln n_{i,a}\right]\right)\left(\ln n_{i,a-j} - E\left[\ln n_{i,a-j}\right]\right) - Cov(\ln n_{i,a}, \ln n_{i,a-j}; \vartheta)\right],$$

where $j \leq a$ and where $Cov(\ln n_{i,a}, \ln n_{i,a-j}; \vartheta)$ is computed from Equation (2) in the main text, for a parameter vector ϑ. The moment condition we exploit in the estimation is $E\left[f\left(n_i; \vartheta\right)\right] = 0$. To operationalize the estimator we define

$$\tilde{f}\left(n_i, \vartheta\right) \equiv (\ln n_{i,a} - \frac{1}{N}\sum_i \ln n_{i,a})(\ln n_{i,a-j} - \frac{1}{N}\sum_i \ln n_{i,a-j}) - Cov(\ln n_{i,a}, \ln n_{i,a-j}; \vartheta),$$

and

$$\tilde{g}_N\left(\vartheta\right) \equiv \frac{1}{N}\sum_i \tilde{f}\left(n_i, \vartheta\right).$$

The minimum distance estimator solves $\min_\vartheta \tilde{g}_N\left(\vartheta\right)' A \tilde{g}_N\left(\vartheta\right)$, where A is a $K \times K$ weighting matrix. Following Guvenen (2009) and many others, we choose A to reflect only differences in the number of data observations underlying the various moments. The estimator $\widehat{\vartheta}$ follows, asymptotically, a normal distribution with a mean equal to the true value of ϑ and a covariance matrix given by $\Sigma = \left(D'D\right) D'\Omega D \left(D'D\right)^{-1}$, where $D = E[\frac{\partial f(n_i,\vartheta)}{\partial\vartheta}]$ is the Jacobian of the moment vector and $\Omega = E[f\left(n_i, \vartheta\right) f\left(n_i, \vartheta\right)']$. The sample analogues of the latter two are $\widetilde{D} = \frac{1}{N}\sum_i \frac{\partial \widetilde{f}(n_i,\vartheta)}{\partial\vartheta}$, and $\widetilde{\Omega} = \frac{1}{N}\sum_i \widetilde{f}\left(n_i, \vartheta\right)' \widetilde{f}\left(n_i, \vartheta\right)$, where we take numerical derivatives to compute $\widetilde{D}$.

B Structural model

B.1 Numerical solution of the structural model

Let us define $\widehat{\mu}(\mathbf{S}) \equiv \frac{\mu(\mathbf{S})}{M^e}$, which evolves as:

$$\widehat{\mu}(\mathbf{S}') = \int \left((1 - x(\mathbf{s})) F(\mathbf{S}'|\mathbf{s}) (\widehat{\mu}(d\mathbf{s}) + G(d\mathbf{s}))\right).$$

and note that in the stationary equilibrium $\mu(\mathbf{S}') = \mu(\mathbf{S})$. The labor market clearing condition in the stationary equilibrium can now be written as:

$$\bar{N} = M^e \left(\frac{\eta}{\eta - 1}\right)^{-\eta} w^{-\eta} Y \widetilde{\varphi} + M^e \widetilde{f} + M^e f^e,$$

where $\widetilde{\varphi} \equiv \int \varphi(\mathbf{s}) \widehat{\mu}(d\mathbf{s})$ and $\widetilde{f} \equiv \int f(1 - x(\mathbf{s}))(\widehat{\mu}(d\mathbf{s}) + G(d\mathbf{s}))$. Note further that $p_i = \frac{\eta}{\eta-1}$ and that the wage is given as

$$w = P^{-1} = \frac{\eta - 1}{\eta} (M^e \widetilde{\varphi})^{\frac{1}{\eta-1}}$$

We solve the model using the following algorithm (following Hopenhayn and Rogerson, 1993):

1. Solve for $Q \equiv w^{-\eta} Y$ from the free entry condition (i.e. guess Q, solve for the firm value functions, evaluate the free-entry condition, update the guess for Q and iterate until the condition holds with equality).

2. Normalize $M^e = 1$, simulate the model and compute $\widehat{\mu}(S)$, $\widetilde{\varphi}$ and $\widetilde{f}$.

3. Solve for M^e from the labor market clearing condition. Compute w, Y, and $\frac{Y}{N}$.

To derive Equation (4.4), note that aggregate output can be written as:

$$Y = Qw^{\eta} = \chi (M^e \widetilde{\varphi})^{\frac{\eta}{\eta-1}} = \chi \left(\int \varphi(\mathbf{s}) \mu(d\mathbf{s})\right)^{\frac{\eta}{\eta-1}} = \Omega^{\frac{\eta}{\eta-1}} \chi \left(\int \varphi(\mathbf{s}) \widetilde{\mu}(d\mathbf{s})\right)^{\frac{\eta}{\eta-1}}$$

where $\widetilde{\mu}(d\mathbf{s}) \equiv \mu(d\mathbf{s})/\Omega$ is the density of firms at state $\mathbf{s}$. It now follows that

$$Y = \Omega^{\frac{\eta}{\eta-1}} \chi^{\frac{1}{1-\eta}} \overline{n}^{\frac{\eta}{\eta-1}}$$

where $\overline{n} = \int n(\mathbf{s})\, \widetilde{\mu}(d\mathbf{s})$ is average firm size and where we have used that $n_i = \chi \varphi_i$.

The state variables for an individual firm consist of the separate components of its demand fundamental: $u_{i,t}$, $v_{i,t}$ and $w_{i,t}$.[21] As mentioned in the main text, we restrict $\rho_v = \rho_w$, in which case the firm only needs to keep track of the sum $v_{i,t} + w_{i,t}$, rather than the two terms separately.

We allow for 31 grid points (equally spaced between -3 and 4) for the permanent component of the demand fundamental, θ. Similarly, we allow for 31 grid points (equally spaced between -5 and 7) for the initial condition $\widetilde{u}$. Finally, the transitory, AR(1) process $w_{i,t}$, is discretized using the method of Rouwenhorst (1995) allowing for 31 grid points. We use value function iteration to solve the firm's maximization problem on the grid specified above.

In simulating the economy, we use $100,000$ startups (i.e. firms which endogenously decide to remain in operation in the first period) and we follow these until the age of 20, consistent with the autocovariance data. Aggregate model variables are constructed using all surviving firms in the model.

B.2 Details on split-sample results

This Appendix presents details on the parametrization and model fit of the model in the split-sample analysis presented in Section 5 of the main text. Tables 3 and 4 show the parameter values for the two subsamples and Figures 14 and 15 document the model's fit across the two subsamples. From the two tables, it is apparent that most of the parameters remain relatively stable across the two sub-samples. However, the distribution of the permanent component θ, a key determinant of long-run size, is estimated to have changed. In particular, both the mean and the dispersion have

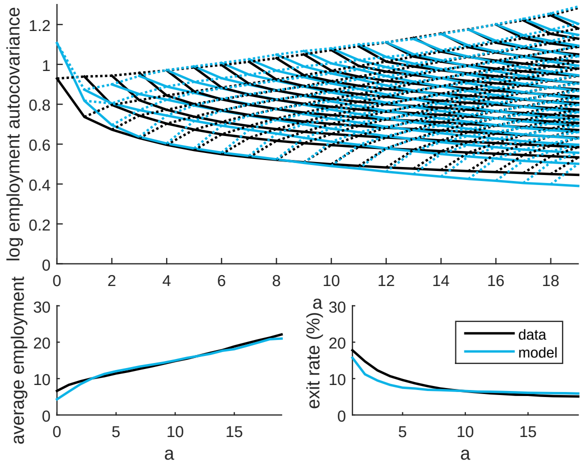
[21] Note that $z_{i,t}$ is purely transitory and therefore its past values do not affect the decision of the firm.

Table 3: Parameter values (early sample)

parameter		value
		set a priori
β	discount factor	0.96
η	elasticity of substitution	6.00
f^e	entry cost	0.447
		estimated
f	fixed cost of operation	0.545
δ	exogenous exit rate	0.042
μ_θ	permanent component θ, mean	-1.770
σ_θ	permanent component θ, st. dev.	1.322
$\sigma_{\widetilde{u}}$	initial condition u_{-1}, st. dev.	1.540
$\sigma_{\widetilde{v}}$	initial condition v_{-1}, st. dev.	1.208
σ_ϵ	transitory shock ϵ, st. dev.	0.304
σ_z	noise shock z, st. dev.	0.153
ρ_u	permanent component, persistence	0.394
ρ_v	transitory component, persistence	0.987

Notes: parameter values. Top three parameters are calibrated as discussed in the main text. The remaining parameters are set such that the model matches the empirical autocovariance of employment and the age profiles of average size and exit rates from age 0 to 19.

Figure 14: Targeted moments: data and structural model (early sample)



Notes: Top panel: Autocovariances of log employment between age $a = h + j$ and age $h \leq a$ in the data and the model, for a balanced panel of firms surviving up to at least age $a = 19$. Bottom left panel: Average employment by age a (unbalanced panel). Bottom right panel: exit rate by age a.

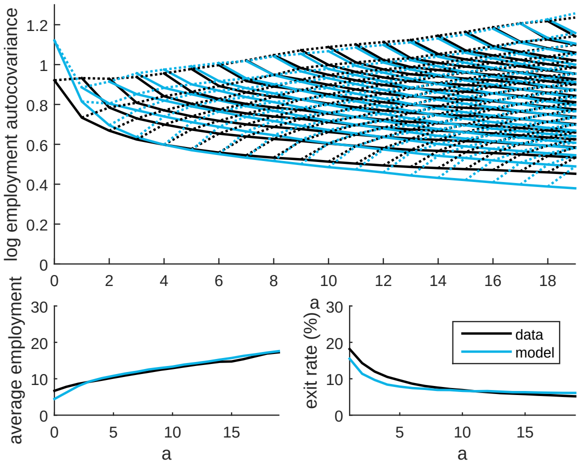
declined going from the early to the late sample.

Table 4: Parameter values (late sample)

parameter		value
		set a priori
β	discount factor	0.96
η	elasticity of substitution	6.00
f^e	entry cost	0.434
		estimated
f	fixed cost of operation	0.530
δ	exogenous exit rate	0.043
μ_θ	permanent component θ, mean	-1.846
σ_θ	permanent component θ, st. dev.	1.303
$\sigma_{\widetilde{u}}$	initial condition u_{-1}, st. dev.	1.563
$\sigma_{\widetilde{v}}$	initial condition v_{-1}, st. dev.	1.209
σ_ϵ	transitory shock ϵ, st. dev.	0.301
σ_z	noise shock z, st. dev.	0.195
ρ_u	permanent component, persistence	0.393
ρ_v	transitory component, persistence	0.987

Notes: parameter values. Top three parameters are calibrated as discussed in the main text. The remaining parameters are set such that the model matches the empirical autocovariance of employment and the age profiles of average size and exit rates from age 0 to 19.

Figure 15: Targeted moments: data and structural model (late sample)



Notes: Top panel: Autocovariances of log employment between age $a = h + j$ and age $h \leq a$ in the data and the model, for a balanced panel of firms surviving up to at least age $a = 19$. Bottom left panel: Average employment by age a (unbalanced panel). Bottom right panel: exit rate by age a.

B.3 Adjustment costs

This Appnedix introduces adjustment costs in the accumulation of the "permanent" ex-ante component u. One interpretation of this specification are costs of accumulating customers (as in e.g. Arkolakis, 2016; Sedláček and Sterk, 2017).

Formally, the process for the demand fundamental can now be written as

$$\begin{array}{lcllll}
\ln \varphi_{i,t} & = & u_{i,t} + v_{i,t} + w_{i,t} + z_{i,t} & & & \\
u_{i,t} & = & \rho_u u_{i,t-1} + \theta_i \left[\lambda + (1-\lambda) q_{i,t}\right], & u_{i,-1} \sim iid(\mu_{\widetilde{u}}, \sigma^2_{\widetilde{u}}) & \theta_i \sim iid(\mu_\theta, \sigma^2_\theta) & \rho_u \in [0,1), \quad \lambda[0,1] \\
v_{i,t} & = & \rho_v v_{i,t-1}, & v_{i,-1} \sim iid(\mu_{\widetilde{v}}, \sigma^2_{\widetilde{v}}) & & \rho_v \in [0,1) \\
w_{i,t} & = & \rho_w w_{i,t} + \varepsilon_{i,t}, & w_{i,-1} = 0 & \varepsilon_{i,a} \sim iid(0, \sigma^2_\varepsilon) & \rho_w \in [0,1) \\
z_{i,t} & \sim & iid(0, \sigma^2_z), & & &
\end{array}$$

The above therefore generalizes the baseline specification in that the permanent component of the demand fundamental accumulates exogenously at the rate of λ and endogenously at the rate of $1-\lambda$, where $g_{i,t}$ is a choice variable for incumbent firms. The parameter λ therefore introduces the the distinction between "active" and "passive" demand accumulation or "demand accumulation by being" as in Foster, Haltiwanger, and Syverson (2016). Importantly, active investment in to demand accumulation comes at a cost, $\frac{\kappa}{2} g^2_{i,t}$. Firms, therefore, maximize firm value not only by choosing prices, employment, output as in the baseline model, but also by choosing demand accumulation (all subject to remaining in the economy). The rest of the model is identical to that in the main text.

The parametrization also follows the same principles as described in Section 3.2. That is, we estimate the majority of the parameters (including the adjustment cost level, κ) by matching the model predicted autocovariance matrix, average size and exit rates by age to their empirical counterparts. The rest of the parameters are set a priori.[22] Table 5 shows the model parameters and Figure 16 displays the model fit.

Intuitively, the dispersion of the permanent component of the demand fundamental, σ_θ is somewhat narrower than in the benchmark model. This is because part of the cross-sectional dispersion in firm sizes at old ages is now also driven by adjustment costs and not only firm types. Nevertheless, the variance decomposition of the cross-sectional variation in firm size in Figure 17 shows that the importance of ex-ante heterogeneity

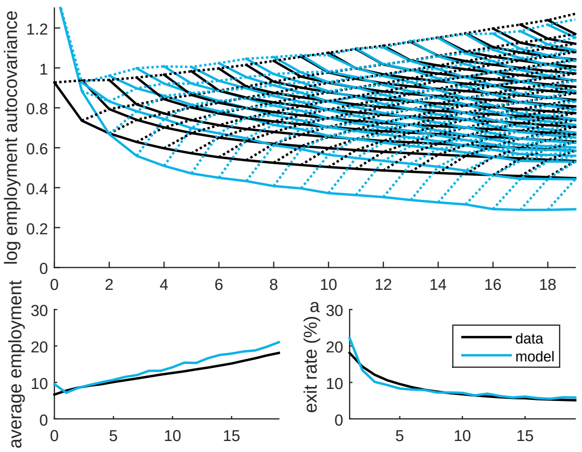
[22] The share of active accumulation, λ, is set to 0.5.

Table 5: Parameter values (model with adjustment costs)

parameter		value
		set a priori
β	discount factor	0.96
η	elasticity of substitution	6.00
f^e	entry cost	0.82
λ	share of active demand accumulation	0.5
		estimated
f	fixed cost of operation	1.000
δ	exogenous exit rate	0.0252
μ_θ	permanent component θ, mean	-1.447
σ_θ	permanent component θ, st. dev.	0.768
$\sigma_{\widetilde{u}}$	initial condition u_{-1}, st. dev.	2.586
$\sigma_{\widetilde{v}}$	initial condition v_{-1}, st. dev.	1.000
σ_ϵ	transitory shock ϵ, st. dev.	0.214
σ_z	noise shock z, st. dev.	0.203
ρ_u	permanent component, persistence	0.388
ρ_v	transitory component, persistence	0.986
κ	adjustment cost level	6.152

Notes: parameter values. Top three parameters are calibrated as discussed in the main text. The remaining parameters are set such that the model matches the empirical autocovariance of employment and the age profiles of average size and exit rates from age 0 to 19.

Figure 16: Targeted moments: data and structural model (model with adjustment costs)



Notes: Top panel: Autocovariances of log employment between age $a = h + j$ and age $h \leq a$ in the data and the model, for a balanced panel of firms surviving up to at least age $a = 19$. Bottom left panel: Average employment by age a (unbalanced panel). Bottom right panel: exit rate by age a.

is at least as important as in the benchmark specification. This holds true also for the importance of ex-ante heterogeneity for firm selection, as depicted in Figure 18. Therefore, while adjustment costs introduce an additional margin of adjustment, they do not alter the main qualitative or quantitative conclusions regarding the relative importance of ex-ante heterogeneity and ex-post shocks.

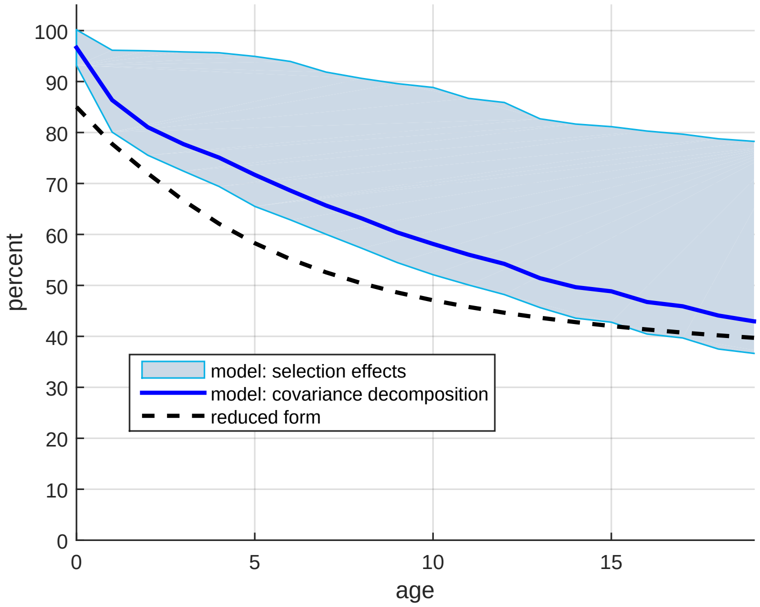
B.4 Results for establishments

While the main text uses both firm and establishment data in the reduced form analysis of Section 2, for simplicity the structural model sections focus only on firms. This Appnedix provides results for the structural model using establishment-level data. Table 6 shows the parameter estimates and Figure 19 depicts the model fit.

Figures 20 and 21 then establish that, also for establishments, ex-ante factors are a dominant force when it comes to the cross-sectional variation in employment and the establishment selection by age, respectively.

Intuitively, the dispersion of the permanent component of the demand fundamental, $\sigma_t heta$ is somewhat narrower than in the benchmark model. This is because part of the cross-sectional dispersion in firm sizes at old ages is now also driven by adjustment costs

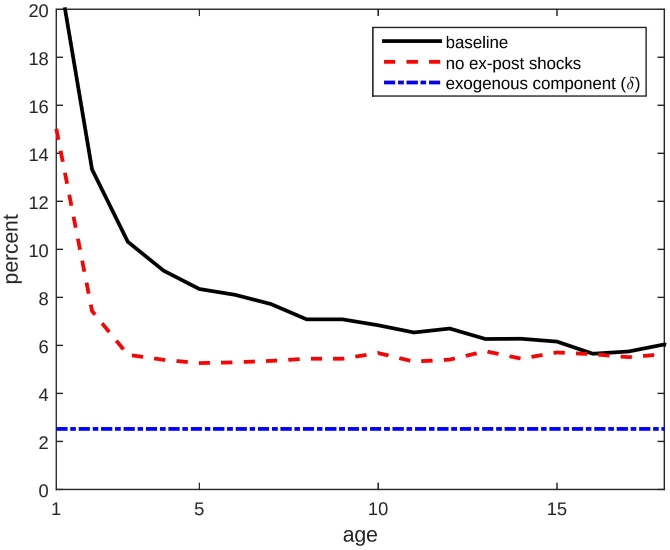
Figure 17: Contribution of ex-ante heterogeneity to cross-sectional employment dispersion (model with adjustment costs)



Note: Contributions to total cross-sectional variance by age. "Reduced-form" refers to the estimates from Figure 3, "model: covariance decomposition" is the decomposition based on the second line in Equation 5. The shaded areas ("model: selection band") is constructed based on the first equality in Equation 5 by attributing, in turn, the term $2Cov(\ln \varphi_i^{EXA}, \ln \varphi_i^{EXP})$ fully to the ex-ante component and to the ex-post component.

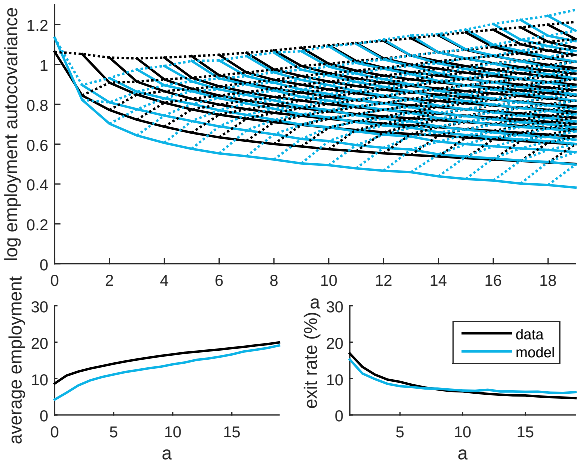
and not only firm types. Nevertheless, the variance decomposition of the cross-sectional variation in firm size in Figure 17 shows that the importance of ex-ante heterogeneity is at least as important as in the benchmark specification. This holds true also for the importance of ex-ante heterogeneity for firm selection, as depicted in Figure 18. Therefore, while adjustment costs introduce an additional margin of adjustment, they do not alter the main qualitative or quantitative conclusions regarding the relative importance of ex-ante heterogeneity and ex-post shocks.

Figure 18: Exit rates (model with adjustment costs)



Note: exit rates by age in the baseline model, exit rates in and a counterfactual economy in with no ex-post demand shocks (but with exogenous exit), and the exogenous exit rate δ.

Figure 19: Targeted moments: data and structural model (establishments)



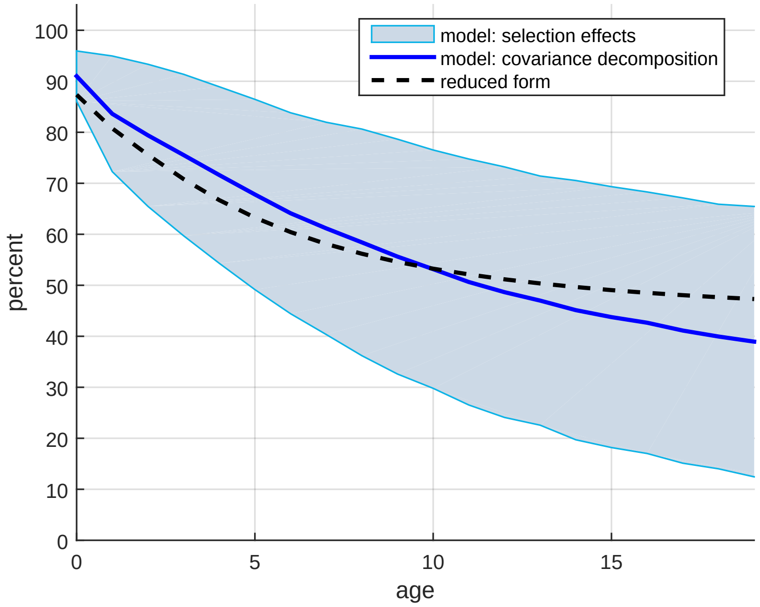
Notes: Top panel: Autocovariances of log employment between age $a = h + j$ and age $h \leq a$ in the data and the model, for a balanced panel of firms surviving up to at least age $a = 19$. Bottom left panel: Average employment by age a (unbalanced panel). Bottom right panel: exit rate by age a.

Table 6: Parameter values (establishments)

parameter		value
		set a priori
β	discount factor	0.96
η	elasticity of substitution	6.00
f^e	entry cost	0.448
		estimated
f	fixed cost of operation	547
δ	exogenous exit rate	0.044
μ_θ	permanent component θ, mean	-1.758
σ_θ	permanent component θ, st. dev.	1.309
$\sigma_{\widetilde{u}}$	initial condition u_{-1}, st. dev.	1.541
$\sigma_{\widetilde{v}}$	initial condition v_{-1}, st. dev.	1.206
σ_ϵ	transitory shock ϵ, st. dev.	0.303
σ_z	noise shock z, st. dev.	0.211
ρ_u	permanent component, persistence	0.393
ρ_v	transitory component, persistence	0.987

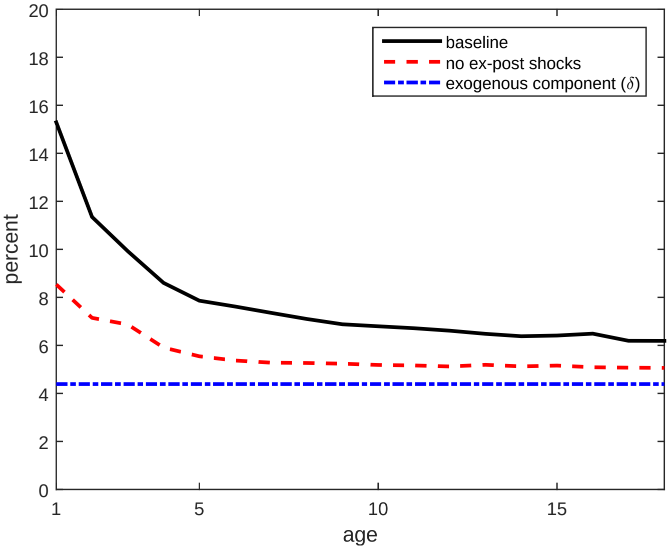
Notes: parameter values. Top three parameters are calibrated as discussed in the main text. The remaining parameters are set such that the model matches the empirical autocovariance of employment and the age profiles of average size and exit rates from age 0 to 19.

Figure 20: Contribution of ex-ante heterogeneity to cross-sectional employment dispersion (establishments)



Note: Contributions to total cross-sectional variance by age. “Reduced-form” refers to the estimates from Figure 3, “model: covariance decomposition” is the decomposition based on the second line in Equation 5. The shaded areas (“model: selection band”) is constructed based on the first equality in Equation 5 by attributing, in turn, the term $2Cov(\ln \varphi_i^{EXA}, \ln \varphi_i^{EXP})$ fully to the ex-ante component and to the ex-post component.

Figure 21: Exit rates (establishments)



Note: exit rates by age in the baseline model, exit rates in and a counterfactual economy in with no ex-post demand shocks (but with exogenous exit), and the exogenous exit rate δ.