CPB Netherlands Bureau for
Economic Policy Analysis

How Far Do Gazelles Run? Growth Patterns of Regular Firms, High Growth Firms and Startups

We investigate the general validity of an investor startup database in studying high growth firms and growth persistence. The vast majority of high growth firms in the Dutch economy does not appear in the database, but included firms have a higher probability of being a high growth firm. Second, in contrast to regular firms, startups show persistent growth patterns.

Commercial startup databases can complement more traditional data sources, but interpretation requires care due to unclear selection in the database.

CPB Discussion Paper

Ramy El-Dardiry, Benedikt Vogt

January 2022

Doi: https://doi.org/10.34932/tna2-rr44

How Far Do Gazelles Run? - Growth Patterns of Regular Firms, High Growth Firms and Startups

Ramy El-Dardiry Benedikt Vogt [*]

Abstract

High growth firms receive a lot attention from policy and academia. In this paper we investigate the general validity of an investor startup database in studying high growth firms and growth persistence. To do so, we match Dutch administrative firm level panel data with this startup database. We establish two main facts. First, although the vast majority of high growth firms in the Dutch economy does not appear in the database, included firms do have a higher probability of being a high growth firm. Second, in contrast to regular Dutch firms and previous findings in the literature, startups show a strikingly persistent growth pattern: their growth phases are prolonged. We conclude that commercial startup databases can complement more traditional data sources, but interpretation requires care due to unclear selection in the database.

JEL Classification Numbers: L25, L26
Keywords: start-ups, entrepreneurship, high growth firms

[*] Corresponding author: Vogt (b.vogt@cpb.nl): CPB - Netherlands Bureau for Economic Policy Analysis, 2594AV The Hague, The Netherlands & Department of Economics, Erasmus School of Economics, Burgemeester Oudlaan 50, 3062PA Rotterdam, The Netherlands; El-Dardiry (r.g.s.el-dardiry@cpb.nl): CPB-Netherlands Bureau for Economic Policy Analysis, 2594AV The Hague. Tim Oudenampsen provided excellent research assistance at the beginning of the project. We are especially thankful to Mark Kattenberg for his valuable comments on the estimation strategy. We also thank Rob Aalbers, Bastiaan Overvest, Erik Stam and seminar participants at the CPB for their input.

1 Introduction

To counter and understand declining business dynamism and productivity growth, both policy makers and academics have paid more attention to the role of high growth firms (henceforth HGFs) in mature economies (Decker et al. (2016); Brown et al. (2017); Haltiwanger et al. (2016); Calvino et al. (2015)). In many countries, policies to stimulate economic growth target startups – loosely defined as young firms that often rely on new technology to develop scaleable business models[1]. The above average potential of startups to become HGFs is one of the most important underlying premises for the special treatment startups receive. Young companies that classify as HGFs are often – almost affectionately – referred to as gazelles. Despite their prominence in policy, it turns out to be difficult to study these firms empirically as there is no clear way to isolate startups from the broader set of entrants in an economy.

Recently, the toolkit to study startups has been extended as privately owned databases about innovative firms have become available. Commercial investor databases such as Crunchbase and Dealroom are promising candidates to provide more insights on startups from an academic point of view. The database Crunch-

[1] Although the word startup is frequently used in academic and policy circles alike an exact definition of startups is missing. Cockayne (2019) notes that the term 'startup appears in both academic and popular literature as an unproblematic idea with an already agreed upon meaning that is therefore presumably unworthy of discussion or interrogation: there is surprisingly little discussion around what startup is or means.' In our work we consider startups to be new firms with ambitious growth plans. Such a definition is linked to the concept of productive entrepreneurship (see e.g. Leendertse et al. (2021)) where only a subset of new entrants are taken into account. Practically, we take a pragmatic approach and consider Dutch startups to be those firms that policy makers refer to as such. This usage coincides with the data sources at our disposal.

base, for example, served as input for at least 140 academic papers (Dalle et al., 2020). Studies and dashboards of major international policy institutions such as the OECD, and local governments make use of privately collected databases on firms (see for instance Breschi et al. (2019) and numerous dashboards for startup ecosystems [2]). These numbers fuel important policy recommendations and investor decisions. On the policy side, for example, subsidies and labor market policies for startups and SMEs are designed based on analyses flowing from private databases. On the investor side, market research and venture capital funding rounds are supported with information from these databases. In academic research the databases have been increasingly used to answer important questions on the decline in productivity, business dynamism and the process of creative destruction (see Autor et al. (2020) for a recent example).

These examples illustrate that startup databases offer a potentially rich and up to date source of information which complements data from national statistics agencies. However, evidence on how representative these type of databases are when it comes to HGFs or the startup landscape of a country is scarce. Little is known about how growth patterns of firms included in private startup databases actually compare to the residual firm population in a country. A more thorough understanding of the general validity of private startup databases is therefore much needed.

In this paper we answer three questions regarding the relationship between

[2] E.g. https://lafrenchtech.com/ (France) and https://technation.io (UK)

HGFs and the coverage of firms in startup databases. First, to what extent are HGFs in an economy present in such databases? Second, are firms within private startup databases more likely to be HGFs compared to regular firms? Third, how do the growth patterns of these companies compare to other firms in a country?

For this purpose, we make use of one of the major startup databases (from here we refer to this database simply as the startup database) and merge this dataset with the universe of limited liability firms in the Netherlands in the period from $2009 - 2018$. This allows us to shed light on which firms are selected by the database and to compare these firms to the whole firm population. Moreover, we are able to delve deeper into the differences between firms in the startup database and other firms. We investigate the incidence of being a HGF and analyze whether firms in the startup database are more likely to become a HGF. Then, we study how firms grow both in terms of employees and revenues.

We make three contributions to the literature. First of all, we add to the growing entrepreneurship literature that analyzes how private databases can be used for academic research in a rigorous way (Dalle et al., 2017, 2020; Leendertse et al., 2021). Such private databases are often built with the help of big data techniques. For instance, web scraping is applied to detect new funding rounds of start-ups. Although such techniques improve agility, they often jeopardize reproducibility. Maula and Stam (2020) recently urged scientists in the quantitative entrepreneurship research community to 'Understand the advantages and limitations of particular sources of data'. We believe startup databases are an important class of data that warrant such an understanding of the advantages and limitations. Do these databases help drawing a richer picture of entrants in an economy?

The second contribution relates to HGFs in an economy. In particular, we find that the proportion of HGFs in a startup database is above average, but the absolute number is still small. Thus, it remains important to focus on firms in the entire economy to foster productivity and job growth.

Third, we provide new insights on growth dynamics of firms by studying growth persistence. The literature finds mixed evidence on the growth persistence of firms. Firms experiencing strong growth seem to be rather (random) 'one hit wonders' that do not show persistent growth over time (Daunfeldt and Halvarsson, 2015; Coad and Hölzl, 2009). We show, that this finding is also true for the majority of firms in our data. However, we demonstrate that firms included in startup databases grow differently. This is very likely to stem from selection bias. A hypothesis that is supported by a survival rate analysis.

The paper is organized as follows. In section 2, we review the existing economic literature on business dynamism and entrepreneurship. We do not aim to provide a complete overview. Rather, we hope to connect different strands of literature and explain how our research relates to the existing literature. In section 3 we explain how the data in this study were constructed and discuss descriptive statistics. In section 4 we describe our empirical approach. Main results are discussed in section 5. We end with a discussion and conclusion in section 6.

2 Related Literature

A growing number of empirical papers in the economics and business literature makes use of private databases that contain information on firm performance, funding rounds and founders. Dalle et al. (2017, 2020) give an overview of recent papers in which the database Crunchbase is used as a key source. They conclude that the database provides useful information for economic research, especially if the data get linked with other information sources. In recent work Kalemli-Ozcan et al. (2015) analyze the van Dijk/ Orbis database and stress that firms in such databases do not form a representative set for the economy. Especially small and nationally operating firms are underrepresented. They put forward a detailed selection and cleaning procedure in order to regain a representative sample. Maula and Stam (2020) compiled a list of best practices concerning quantitative studies on entrepreneurship. They point out how the use of commercial databases can lead to selection bias as there is a risk that these databases focus on relatively successful firms. We complement their work, by directly comparing such a commercial database with archival data.

The literature on firm growth dynamics and the impact of HGFs on the overall economy continues to evolve. Henrekson and Johansson (2010) conducted a meta-analysis on the characteristics of HGFs. They show that HGFs are on average younger and smaller, and that HGFs are strong net job creators. Empirical studies on the role of scale in company growth rates (Stanley et al. (1996)) have spurred research that aims to better understand the growth distribution of firms in an economy. We therefore contribute to a growing literature which provides evidence on heterogeneity in growth paths of firms (e.g. Garnsey et al. (2006)) and violations of Gibrat's law (Gibrat, 1931).[3] For instance, in two papers the growth dynamics of Austrian (Coad and Hölzl, 2009) and Swedish (Daunfeldt and Halvarsson, 2015) firms have been analyzed. Daunfeldt and Halvarsson (2015) find that HGFs are on average one hit wonders: few firms realize high growth in adjacent periods. Coad and Hölzl (2009) conclude that periods of high growth are often preceded by a period of low growth or decline. On the contrary, Capasso et al. (2014) find that next to 'bouncing' firms, also persistent 'outperformers' exist.

The literature on firm growth is related to the field of business dynamism. The key underlying mechanism here is that there is a healthy level of faster growing firms that push other firms out of the market. A decline in business dynamism has been extensively reported for the US (Decker et al. (2014)) and more recently for the Netherlands (Freeman et al., 2021) and Belgium (Bijnens and Konings (2018)). These studies show that entry rates have experienced a decades long decline and the share of employment from young firms is decreasing. Bijnens and Konings (2018) further argue that the strikingly similar observations for the US and Belgium indicate that *global* trends are the likely cause of the reduction in business dynamism. Digitalization and globalization are put forward as possible explanations. Bravo-Biosca et al. (2016) show that business dynamism varies across countries and that these differences are associated with a diverse set of factors, such as market regulation, R&D policies, and the structure of the local financial market. Although the decline in business dynamism in these studies is often associated with digitalization or technology in general, empirical evidence is limited. Partially because in these studies all entrants in an economy are considered

[3] An excellent overview of these competing theories is provided by Stam (2010).

equal – a new grocery story and a high tech startup typically both count as entrants – it remains difficult to grasp insights on the underlying reasons for a decline in business dynamism. A strand of literature considers how differences between firms determine differences in growth rates. In a recent paper, Sterk et al. (2021) show that for US firms employment growth is largely driven by firm heterogeneity at birth. It is an open question in how far startup databases, such as the one used in our study, are able to capture these type of heterogeneities.

3 Data & Descriptive Statistics

The empirical results of this paper are based on three main data sets: the general business register in the Netherlands (ABR), administrative data on the financial records of Dutch non-financial firms (NFO), and a commercial database listing Dutch startups (Dealroom). In section 3.1 we will discuss the three datasets in more detail and explain how we merge the different sets to create panel data. In section 3.2 we present descriptive statistics.

3.1 Full Population Panel Data

3.1.1 The general business register

The 'general business register' [4] forms the core datasource in our studies. This database, maintained by Statistics Netherlands, contains information on all limited liability firms in the Netherlands, including data on yearly employment and location. We focus our analysis on the 2009 − 2018 period. Firms appear on two different levels of aggregation: the *corporate* level and the *entity* level. In this

[4] Abbreviated in Dutch by 'ABR' which stands for Algemene Bedrijven Register

paper, we use the corporate level, the highest level of aggregation, to create firm panel data. By doing so, we avoid artificial inflation of business dynamics in the economy due to the establishment of new entities (e.g. a new branch) within the same firm.

3.1.2 Startup database

We augment the full population data on limited liability companies with information from a commercial database on start-ups and scale-ups. This database from Dealroom aims to cover the entire startup ecosystem in a country. Dealroom uses a variety of ways to collect and sanitize data. First, they harvest public data from the internet by e.g. web scraping and connecting to domain registries and job boards. Second, they partner with government agencies to share data on startups. Third, they manually verify new entries. The original database contains $6,536$ entries covering a period from 1994 until 2018.[5] For these companies, we then manually looked up the registration number within the Dutch Chamber of Commerce. The Chamber of Commerce registration number allows us to connect the startup database with the ABR. With this method we are able to retrieve $4,888$ of these companies in the business registry in the period of 2009 until 2018. For the remaining $1,646$ entries in the database we do not find matches in the company registry. In the final sample we analyze $18,685$ firm-year observations with a total of $3,935$ unique company groups.

The main reasons for the differences in the startup database and the matches with the register data are as follows. First, we only consider companies which are registered at the Netherlands Chamber of Commerce. Second, we discard all

[5] We received the database mid 2019.

companies aged ten years or older since we do not consider them to be young firms. Third, we only consider non-financial limited liability companies that file taxes in the Netherlands.

3.1.3 Financial records

We merge the general business register data with the financial records of non-financial organizations (NFO) to gain access to the balance sheets and profit & loss statements. In doing so we are able to create a measure of yearly revenue growth for each firm. This leaves us with $433,301$ unique firms over our sample period and a total of $2,055,863$ firm-year observations.

3.2 Descriptive Statistics

We start off with basic statistics over time and compare the firms in the startup database with remaining other firms in the Netherlands. Our focus lies on the presence of high growth and the comparison of some basic firm characteristics such as level of employment, total assets and firm revenue.

Table 1: Descriptive statistics - full sample

	Mean	Standard Dev.	Median
No. employees in company group	11.52	128.08	2
Revenue (in thsd. Euros)	5246	131913	395
Firm age	5.6	3.7	5
Observations		2,055,863	

Note: The table reports the descriptive statistics for the sample from 2009-2018.

3.2.1 Comparing descriptive statistics of the two databases

The number of employees and revenue form the basis for our analysis on growth dynamics. In Table 2, we compare the mean number of employees in the startup database with all other companies in the Netherlands. The entries in the startup database are on average larger, both in terms of employees and revenue.

Differences are also reflected in the size distribution of firms. Table 3 shows the percentage of firms that fall within a certain size category based on the number of employees. Almost 80% of all firms are classified as micro-firms that have less than 10 employees. In the startup database roughly 60% of companies fall within this smallest size category. Compared with all firms, the startup database has about twice as many firms in the small (10-49 employees) and medium (50-99 employees) size category. The percentage of large firms is lower in the population of non startup firms and the startup database. The majority of firms in our dataset is thus very small and the startup database is skewed towards slightly larger firms.

Table 2: Descriptive statistics - Startups and No Startups

Firms in Startup Database			
	Mean	Standard Dev.	Median
No. employees in company group	24.69	129.51	6
Revenue (in thsd. Euros)	29325	389890	773
Firm age	4.9	3.6	4
Observations		18,685	

Firms not in Startup Database			
	Mean	Standard Dev.	Median
No. employees in company group	11.40	128.06	2
Revenue (in thsd. Euros)	5025	127126	392
Firm age	5.6	3.7	5
Observations		2,037,178	

Note: The table reports the descriptive statistics for the sample from 2009-2018. The sample is split by companies which are linked to the investor database and those who are not.

Table 3: Size Distribution - Startups and No Startups

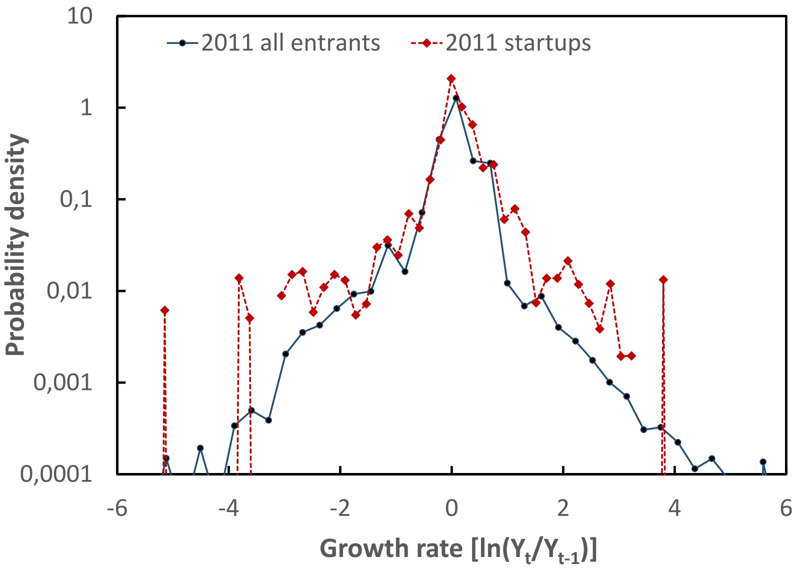
	Regular Firms	Startups
Share of micro firms (0-9) employees	81%	62.2%
Share of small firms (10-49) employees	15.8%	29.6%
Share of medium firms (50-249) employees	2.8 %	7%
Share of large firms (250+) employees	0.4%	1.2%

Note: The table reports the descriptive statistics for the sample from 2009-2018. The sample is split by companies which are linked to the investor database and those who are not. The numbers do not exactly add up to 100% due to rounding.

3.2.2 Growth Distribution & Dynamics

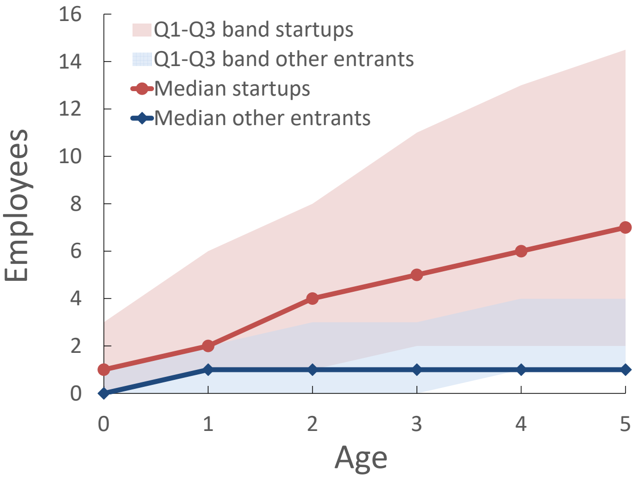
To get a first glimpse of a potential difference in growth dynamics, we compare year on year (YoY) employee growth. Figure 1 shows the kernel density plots for log employee growth in 2011 for all firms (blue dots) and startups (red diamonds). This figure is not cleaned for size or sector dependence. Similar to Coad and Hölzl (2009) we note that the growth distribution is fat-tailed. For most years we observe that the growth distribution for startups is skewed towards stronger growth.

Figure 1: Kernel density plots for employee growth in 2011 for startups (red diamonds) and all firms (blue dots)



Similar patterns emerge when we study the trajectory companies follow over time in terms of number of employees and revenue. Figures 2 and 3 show the median number of employees and median revenue versus age. Here, we limit our analysis to firms that entered the economy between 2010 and 2013 and that report financial data in their birth year. The grow trajectory for startups is different compared to other new entrants: startups in our dataset grow in terms of employees during the first five years of existence, while other entrants remain stagnant. In terms of revenue, startups and other entrants start at roughly the same level, between 100 and 200 k€. But again startups grow faster: median revenue approaches one million euros after five years while for other entrants revenue grows to around 300 k€. Interestingly, the pattern for pretax profits looks different, as shown in Fig. 4. In contrast to regular entrants, the median startup makes a loss in the first year and only starts making a profit in year 3. The strong rise in revenue combined with low profits suggests that startups are on average more focused on reinvesting to fuel further growth in revenue and employees.

Figure 2: Employee growth trajectory for startups (red) and other entrants (blue) with entry years 2010-2013



Finally, before turning our attention to HGFs and our regression analyses, we study the survival rate of entrants. Figure 5 shows for every cohort (given by entry year) which percentage of firms remain as time progresses. About 20% of firms do not survive the first year. After the first year the survival curve flattens. More importantly, on average entrants from our startup database survive longer than other entrants. This trend is present for every cohort in our panel data. We believe that this observation is a clear consequence of the fact that firms in the startup database are selected after birth. Firms that are more visible in later years, e.g. due to investments rounds, end up having a higher chance of appearing in this

Figure 3: Revenue growth trajectory for startups (red) and other entrants (blue) with entry years 2010-2013

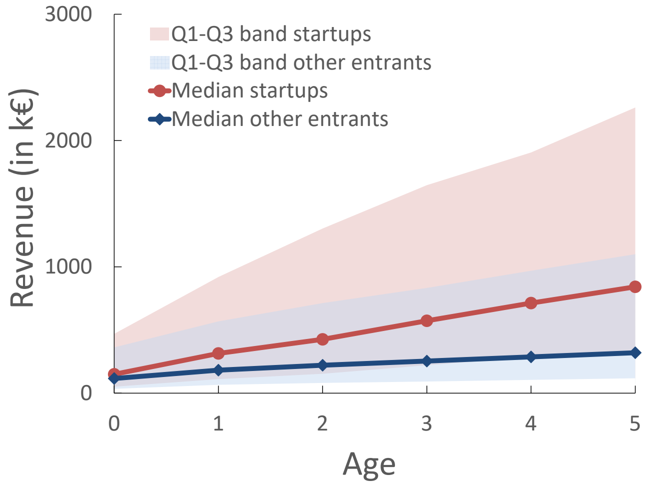


Figure 4: Profit growth trajectory for startups (red) and other entrants (blue) with entry years 2010-2013

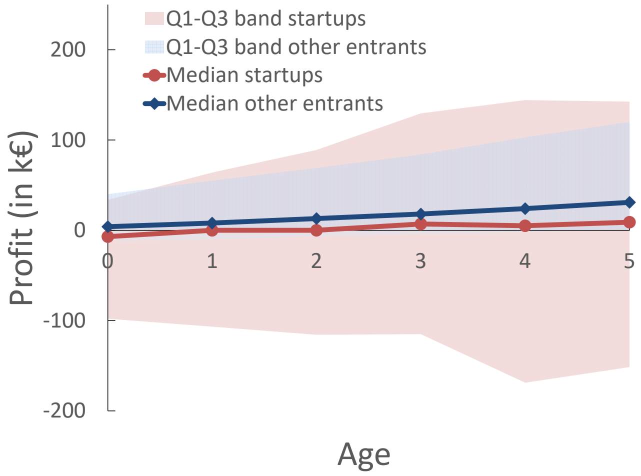
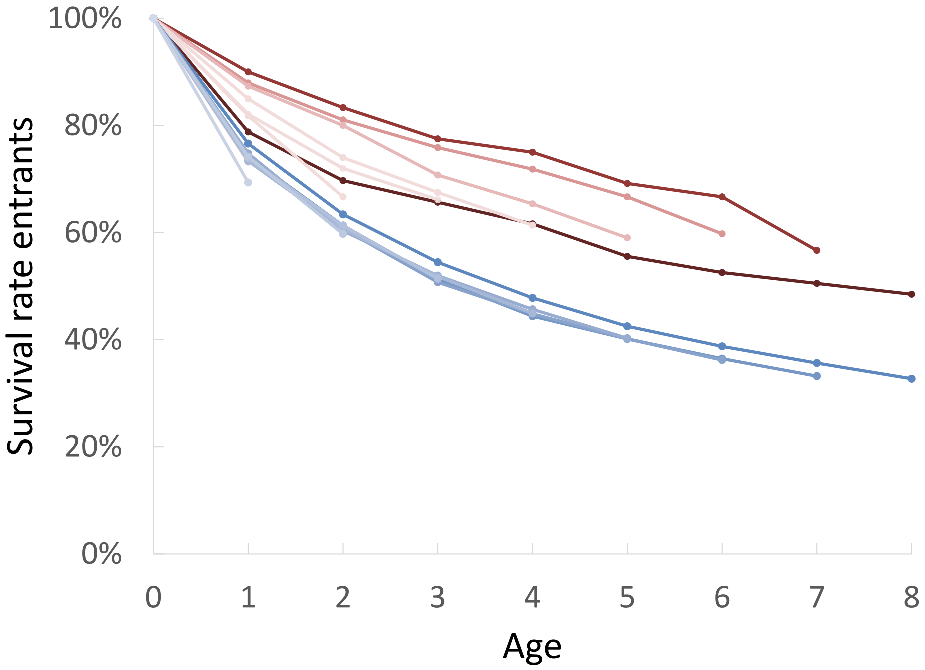


Figure 5: Survival rate by cohort for startups (red) and other entrants (blue)



database.

3.2.3 HGFs

We investigate the incidence of HGFs in the two samples over time. Here, we follow the OECD definition of HGFs[6]. A firm has to have more than 10 employees and a yearly average growth rate of at least 20% over a three year period. We also conduct all analyses with the growth of revenue. The results are presented in

[6] Euostat - OECD manual on business demography statistics (http://www.oecd.org/sdd/39974588.pdf)

Appendix B.[7]

Table 4 shows the presence of HGFs in 2018 for all firms and the presence of HGFs in the startup database. The table shows the number of firms which classified as an HGF for at least one year. In total, we find that in 2018 $6,031$ firms classified as an HGF at some point in their lifecycle of which only 280 firms were covered by the startup database. This pattern is very similar for other years as shown by tables in Appendix A.

Table 4: HGFs and startup database in 2018

	Not in database	In database	Total
No HGF	210,485	2,337	212,822
Ever HGF	6,707	304	7,011
Total	217,192	2,641	219,833

4 Empirical Approach

In this section we first present our empirical approach to better understand characteristics of HGFs and their appearance in the startup database. In a second step we explain how we analyze the growth dynamics of Dutch firms.

[7] More formally, our definition of HGFS is as follows:

$$HGF_{it} = \begin{cases} 1, & \text{if } \Delta Y_{it} > 0.2 \wedge employees_{it-3} > 10 \\ 0, & \text{otherwise} \end{cases}$$

$$\Delta Y_{it} = \left(\frac{Y_{it}}{Y_{it-3}} \right)^{\frac{1}{3}} - 1$$

$$Y_{it} \in \{employees_{it}, revenue_{it}\}$$

4.1 Characteristics of HGFs

To answer the question whether HGFs are more likely to appear in the startup database, we conduct a simple regression analysis. We estimate a linear probability model by using OLS in which we aim to explain the likelihood of high growth using a set of explanatory variables. The dummy D_i indicates whether a firm is captured in the startup database. Our main estimate of interest is β which captures the correlation of being in the startup database and classifying as HGF.
We add a rich set of control variables in the vector $\mathbf{F_{it}}$: firm age in four categories, size measured by the number of employees, cohort dummies and sector-year dummies.[8] Equation 1 shows the specification of the linear probability model, where we denote α as the intercept:

$$P(HGF_{it} = 1 | D_i, \mathbf{F_{it}}) = \alpha + D_i\beta + \mathbf{F_{it}}'\boldsymbol{\gamma} \tag{1}$$

4.2 Revealing Growth Dynamics with Quantile Regressions

An important question is what the growth processes of firms look like and whether they are different for companies in the startup database. While correlates of the incidence of high growth are an essential first step, the dynamics of growth are important for a more thorough understanding and the design of data-informed policies. Is a period of growth followed by another period of growth or, is a period of growth followed by a period of contraction? If growth is persistent then statements

[8] The sector year dummies are indicator variables on the 2-digit industry level interacted with the year

such as HGFs are ‘job motors’ in the economy can be justified. However, if periods of growth are followed by periods of poorer performance, HGFs might contribute less to job creation than initially thought (Daunfeldt and Halvarsson, 2015).

In order to investigate the growth dynamics of firms we follow a similar quantile regression approach as Coad and Hölzl (2009). Essentially, we analyze whether growth dynamics differ with regard to the position of a firm in the growth distribution curve. Such an approach allows us to understand whether firms follow smooth or jumpy growth paths.

We define year on year growth in log changes $\Delta Y_{it} = \ln Y_{it} - \ln Y_{it-1}$, for employees (E) and revenue (R): $Y_{it} \in \{E_{it}, R_{it}\}$.

The quantile regressions (Koenker and Hallock, 2001) of the following form are used to study growth dynamics. We estimate the year on year growth of the growth quantile τ for each firm i in year t:

$$Q_{\tau}(\Delta Y_{it}) = \iota_s + \upsilon_t + \sum_{l=1}^{L} \delta_l \Delta Y_{it-l} + \zeta Y_{it-1} + \epsilon_{it} \qquad \text{For the quantiles } \tau \in \{0.1, 0.2, ..., 0.9\} \quad (2)$$

L refers to the number of lags, ι_s, υ_t are industry and year fixed effects. We follow the literature and set the number of lags equal to 2. Equation 2 also contains a control for the total *level* of employment or revenue in the previous period. This is an important variable because it allows us to control for regression to the mean (Davies and Geroski, 1997).[9] Our main coefficients of interests are δ_1 and δ_2.

One potential concern in the estimation strategy is a bias due to regression to the mean. In equation 1, we also control for the *level* of Y in the previous year. Hence the estimates of ζ capture regression to the mean. In section **??** we show, that in simulated data in which Y_{it} is generated by a random process, the estimates of δ_1 and δ_2 are not statistically significantly different from zero in any quantile. This suggests, that conditional on the level in the previous period, the estimates of δ_1 and δ_2 do not pick up regression to the mean.

Second, we investigate whether the growth dynamics of the firms in the startup database differ from the rest of the firm population. Therefore, we estimate an augmented version of equation 2 where we add an interaction term with an indicator variable D_i if a firm is included in the startup database and the lagged growth.

$$\begin{aligned} Q_\tau(\Delta Y_{it}) = \iota_s + \upsilon_t + \beta \mathbb{1}[D_i] + \sum_{l=1}^{L} \delta_l \Delta Y_{it-l} + \\ \sum_{l=1}^{L} \kappa_l \Delta Y_{it-l} \times \mathbb{1}[D_i] + \zeta Y_{it-1} + \epsilon_{it} \\ \text{For the quantiles } \tau \in \{0.1, 0.2, ..., 0.9\} \end{aligned} \tag{3}$$

We are interested in whether the growth dynamics of the firms in the database are significantly different from those which do not show up in the database. Our coefficients of interest are δ_1 and δ_2 for the residual firms, and κ_1 and κ_2 for the

[9] We discuss potential concerns about regression to the mean in section 5.3

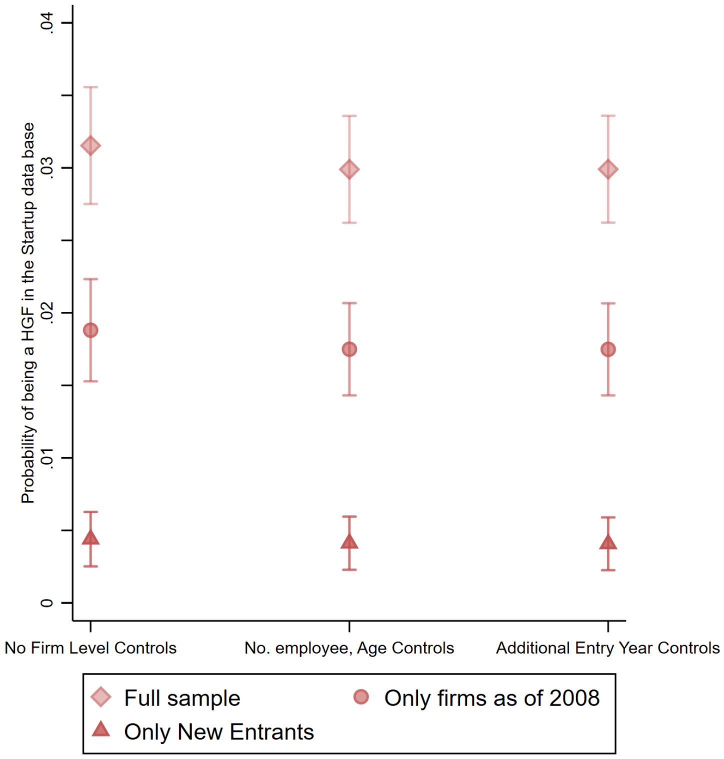
firms in the startup database.

5 Results

In this section we present our main results. We start off with a description of the regression analyses on HGFs. In a second step we show the growth dynamics of all Dutch firms and compare it to dynamics of firms in the startup database.

5.1 HGFs in the Startup Database

Figure 6: Incidence of HGFs in the startup database



Note. The figure shows regression results of linear probability models from equation 1. Point estimates of the startup base indicator variable (β) from three samples are shown: 1) Full sample 2) Only firms as of 2009 3) Only new entrants. The vertical lines are the 95% confidence intervals.

Are companies in the startup database more likely to be high growth firms also when we control for other observable firm characteristics? The results of the regression analyses below provide an answer to this question: Companies in the startup database are significantly positively associated with high growth. The estimation results of $\hat{\beta}$ in equation 1 are shown in Figure 6. The figure shows the point estimates of the 'startup database' indicator variable for three specifications and three samples. Diamonds show the results of the full sample. Circles show the results for firm with entries in 2008 and later. Triangles show the specification for firms less than four years old.

First, the startup indicator variable is independent of other observable characteristics such as firm size, age and the cohort.[10] This results from the observation that the estimates are very stable if we include control variables on the firm level.

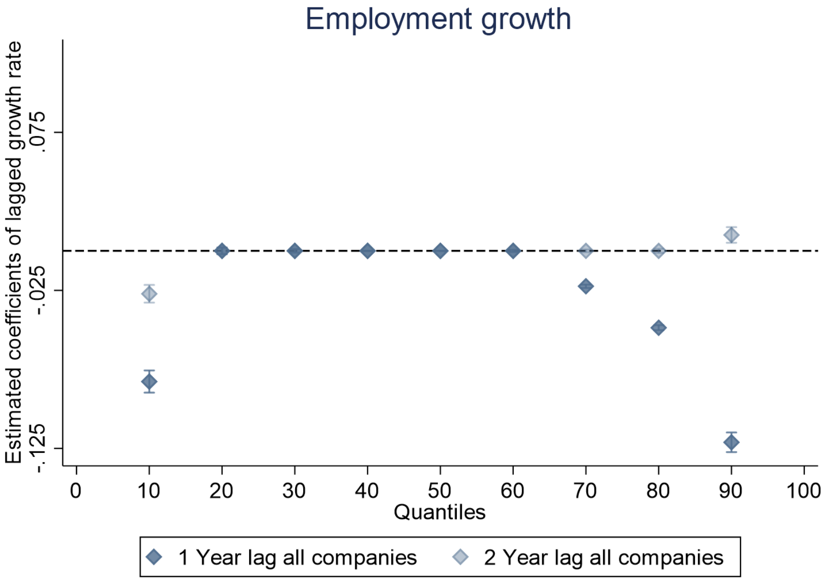
The estimates for β change in magnitude when we study different samples, but remain stable if we include control variables. The full sample estimates are the greatest in magnitude, followed by the estimates in the sample where we exclude firms with an entry year before 2008. We obtain the smallest estimates in the sample in which we focus only on young firms, i.e. a sample where all firms older than 4 years of age are excluded. This decrease in magnitude is intuitive as older firms are more likely to have experienced a period of high growth somewhere in their lifecycle.

[10] Please note that the age variables are still identified in our case because we analyze an unbalanced panel of companies which generates variation in company age and cohort.

The results for HGFs using revenue growth are very similar and are provided in Appendix B. This suggests that the database does not only capture HGFs in employment but also in terms of revenue.

5.2 Growth dynamics

Figure 7: Dynamics of growth rates for all firms



Note. The figure shows results of the quantile regressions which are described in equation 2. Dark blue diamonds refer to the estimated coefficient of the first lag of the change in the growth rate $\hat{\delta_1}$. Light blue diamonds show the estimates of the second lag of the change in the growth rate $\hat{\delta_2}$.

Growth is not persistent for the majority of Dutch firms. In Figure 7 we show the first results from our quantile regression analysis. This figure shows the estimates of the first lag $\hat{\delta_1}$ (dark blue) and second lag $\hat{\delta_2}$ (light blue) growth autocorrelations for each quantile of the full sample.

Our first observation is that firms at the extremes of the growth rate distribution at time t show a negative autocorrelation with growth in the previous period. Thus, these results show that firms experiencing the highest growth probably did not perform as well in the years before. By the same token, firms performing relatively poorly at time t probably did somewhat better in previous periods.[11]

Growth two years back $(t-2)$ is only weakly associated with the growth in the last year for all firms. This is represented in the pattern of the light blue quantile estimates in Figure 7 which show the coefficients of the quantile regressions of the second lag.

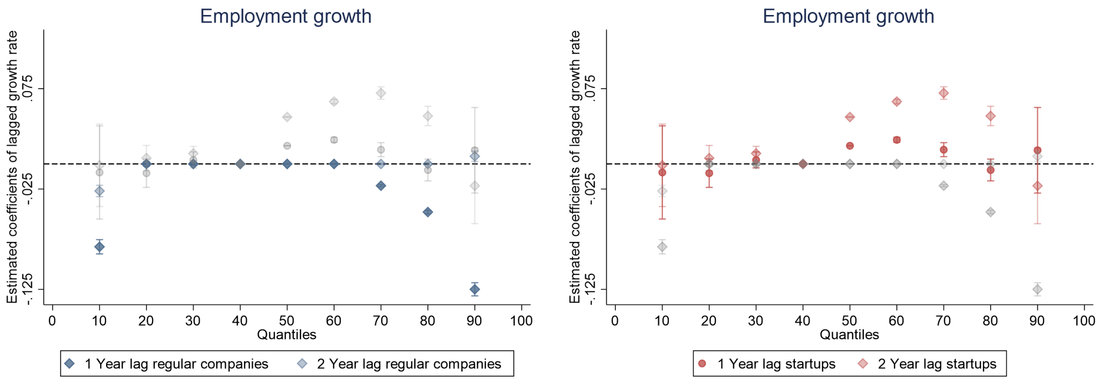
Second, the majority of firms does not show any systematic growth dynamics. The growth autocorrelation coefficients for the 30^{th} until the 90^{th} quantile are not statistically significantly different from zero.

The growth dynamics for the companies in the startup database are strikingly different from 'regular' firms. Our results indicate that firms in the startup database grow in a more persistent way. Figure 8a shows the growth patterns of 'regular' firms once we add an interaction term with the startup database in the regressions as described in equation 3. As expected, this figure looks very similar to Figure 7 as the vast majority of firms falls outside the scope of the database.

In Figure 8b we show the estimates of the estimates $\hat{\kappa_1}$ (dark red) and $\hat{\kappa_2}$ (light red). This allows us to show if companies in the startup database exhibit a statis-

[11] Note that ultimately, the effect of the negative auto-correlation depends on whether a firm grew or shrank in the periods before. If a firm shrank, a negative autocorrelation coefficient implies a less stronger decline in the current period. Conversely, if a firm grew in the periods before a negative autocorrelation coefficient implies an even stronger decline in the current period.

Figure 8: Dynamics of growth rates of non - startups compared to startups



(a) Non - Startups

(b) Startups

Note. The figure shows the estimated coffecients of 3. Panel 8a shows the estimated lags of each quantile for 'regular' firms. Panel 8b shows the following estimates: $\widehat{\kappa_1}$ (dark red) and $\widehat{\kappa_2}$ (light red)

tically significantly different growth pattern from 'regular' firms. For a comparison we plot the coefficients of the other groups in gray.

Startups in the upper quantiles of the growth distribution display positive growth autocorrelation coefficients. Hence, startups that show strong growth have likely experienced strong growth in earlier years.

5.3 Robustness: Regression to the Mean?

The methodology which we apply in section 4.2 to analyze growth persistence has been frequently used in the empirical literature on firm growth. A set of control variables and quantile regression is used to explain annual firm growth with lagged firm growth from the previous years (see for instance Coad et al. (2014); Coad and Hölzl (2009); Coad (2007)). One potential explanation of a significant relationship could be regression towards the mean (RTTM). RTTM in our sample means, that firms which experience a period of high growth are very likely to have experienced a period of low growth before and vice versa. This is especially true if firm growth is a random process (Davies and Geroski, 1997).

One way to control for RTTM is to include the level of the dependent variable in the previous period (Coad and Hölzl, 2009; Davies and Geroski, 1997). A negative correlation with this variable indicates that periods of high (low) growth a preceded by low (high) levels of the dependent variable. However, since the lagged differences in equation 2 are also functions of the levels in the previous periods we cannot exclude that the estimates of the lagged growth rates ($\hat{\delta}_1$, $\hat{\delta}_2$) could potentially be driven by regression to the mean.

In order to test if our estimates $\hat{\delta}_1$, $\hat{\delta}_2$, $\hat{\kappa_1}$ and $\hat{\kappa_2}$ are driven by regression to the mean, we run some simple Monte Carlo simulations. A description of the procedure can be found in Appendix C. The main conclusion from the simulations is that our results are unlikely to be driven by regression to the mean.

6 Discussion & Conclusion

In this paper we investigate whether growth patterns of startups differ from other firms. For the whole Dutch firm population we show that growth is not persistent: the one and two years lagged growth rates show an inverse U-shaped autocorrelation pattern across the growth distribution. One year of high growth is preceded by a relatively poor performance in previous periods. Our results are unlikely to be driven by regression to the mean.

The picture changes when we focus on firms that are classified as startups by a commercial party. Here, the growth patterns become more persistent: Firms in the startup database display strong performance over a longer period of time.

From these observations, it is tempting to conclude that a startup database can be used as predictor for success or an early warning indicator for persistent growth. However, such a conclusion would be premature. We know too little about the timing and the exact selection criteria for companies in these type of databases. It is likely that firms are selected ex-post on success and it is a priori unclear whether they are indeed informative about business dynamism in a country.

Conversely, we also conclude that a too stringent focus on startup databases by policy makers increases the probability of missing firms with strong growth potential. Future research could investigate the feasibility of setting up early warning indicators for *persistent* growth. Next steps can be the use of complementary data sources, e.g. on R&D expenditure and patents, to reveal explanatory variables for business dynamism and strong growth.

References

Autor, D., D. Dorn, L. F. Katz, C. Patterson, and J. Van Reenen (2020). The fall of the labor share and the rise of superstar firms. *The Quarterly Journal of Economics 135*(2), 645–709.

Bijnens, G. and J. Konings (2018, November). Declining business dynamism in Belgium. *Small Business Economics*.

Bravo-Biosca, A., C. Criscuolo, and C. Menon (2016, October). What drives the dynamics of business growth? *Economic Policy 31*(88), 703–742.

Breschi, S., J. Lassébie, A. C. Lembcke, C. Menon, and C. Paunov (2019). Public research and innovative entrepreneurship. (64).

Brown, R., S. Mawson, and C. Mason (2017). Myth-busting and entrepreneurship policy: the case of high growth firms. *Entrepreneurship & Regional Development 29*, 414–443.

Calvino, F., C. Criscuolo, and C. Menon (2015, July). Cross-country evidence on start-up dynamics. *OECD*.

Capasso, M., E. Cefis, and K. Frenken (2014). On the existence of persistently outperforming firms. *Industrial and Corporate Change 23*(4), 997–1036.

Coad, A. (2007). A closer look at serial growth rate correlation. *Review of Industrial Organization 31*(1), 69–82.

Coad, A., S.-O. Daunfeldt, W. Hölzl, D. Johansson, and P. Nightingale (2014). High-growth firms: introduction to the special section. *Industrial and Corporate Change 23*(1), 91–112.

Coad, A. and W. Hölzl (2009). On the autocorrelation of growth rates. *Journal of Industry, Competition and Trade 9*(2), 139–166.

Cockayne, D. (2019). What is a startup firm? a methodological and epistemological investigation into research objects in economic geography. *Geoforum 107*, 77–87.

Dalle, J.-M., M. Den Besten, and C. Menon (2017). Using crunchbase for economic and managerial research.

Dalle, J.-M., M. Den Besten, and C. Menon (2020). Crunchbase research: Monitoring entrepreneurship in the age of big data.

Daunfeldt, S.-O. and D. Halvarsson (2015). Are high-growth firms one-hit wonders? evidence from sweden. *Small Business Economics 44*(2), 361–383.

Davies, S. W. and P. A. Geroski (1997). Changes in concentration, turbulence, and the dynamics of market shares. *Review of Economics and Statistics 79*(3), 383–391.

Decker, R., J. Haltiwanger, R. Jarmin, and J. Miranda (2014, September). The Role of Entrepreneurship in US Job Creation and Economic Dynamism. *Journal of Economic Perspectives 28*(3), 3–24.

Decker, R. A., J. Haltiwanger, R. S. 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. The Economics of Entrepreneurship.

Freeman, D., L. Bettendorf, H. van Heuvelen, and G. Meijerink (2021). The contribution of business dynamics to productivity growth in the netherlands. Technical report, CPB Netherlands Bureau for Economic Policy Analysis.

Garnsey, E., E. Stam, and P. Heffernan (2006). New firm growth: Exploring processes and paths. *Industry and Innovation 13*(1), 1–20.

Gibrat, R. (1931). Les inégalits économiques. *Sirey*.

Haltiwanger, J., R. Jarmin, R. Kulick, and J. Miranda (2016). High growth young firms: Contribution to job, output, and productivity growth. In *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*, pp. 11–62. National Bureau of Economic Research, Inc.

Henrekson, M. and D. Johansson (2010, September). Gazelles as job creators: a survey and interpretation of the evidence. *Small Business Economics 35*(2), 227–244.

Kalemli-Ozcan, S., B. Sorensen, C. Villegas-Sanchez, V. Volosovych, and S. Yesiltas (2015). How to construct nationally representative firm level data from the orbis global database: New facts and aggregate implications. Technical report, National Bureau of Economic Research.

Koenker, R. and K. F. Hallock (2001). Quantile regression. *Journal of economic perspectives 15*(4), 143–156.

Leendertse, J., M. Schrijvers, and E. Stam (2021). Measure twice, cut once: Entrepreneurial ecosystem metrics. *Research Policy*, 104336.

Maula, M. and W. Stam (2020). Enhancing rigor in quantitative entrepreneurship research. *Entrepreneurship Theory and Practice 44*(6), 1059–1090.

Stam, E. (2010). Growth beyond gibrat: firm growth processes and strategies. *Small Business Economics 35*(2), 129–135.

Stanley, M. H., L. A. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, P. Maass, M. A. Salinger, and H. E. Stanley (1996). Scaling behaviour in the growth of companies. *Nature 379*(6568), 804–806.

Sterk, V., P. Sedláček, and B. Pugsley (2021, February). The nature of firm growth. *American Economic Review 111*(2), 547–79.

A Additional Descriptive Statistics

Table 5: HGFs and startup database in 2012

	Not in database	In database	Total
No HGF	192,826	1,203	194,029
Ever HGF	7,056	318	7,374
Total	199,882	1,521	201,403

Table 6: HGFs and startup database in 2014

	Not in database	In database	Total
No HGF	198,965	1,639	200,604
Ever HGF	7,436	354	7,790
Total	206,401	1,993	208,394

Table 7: HGFs and startup database in 2016

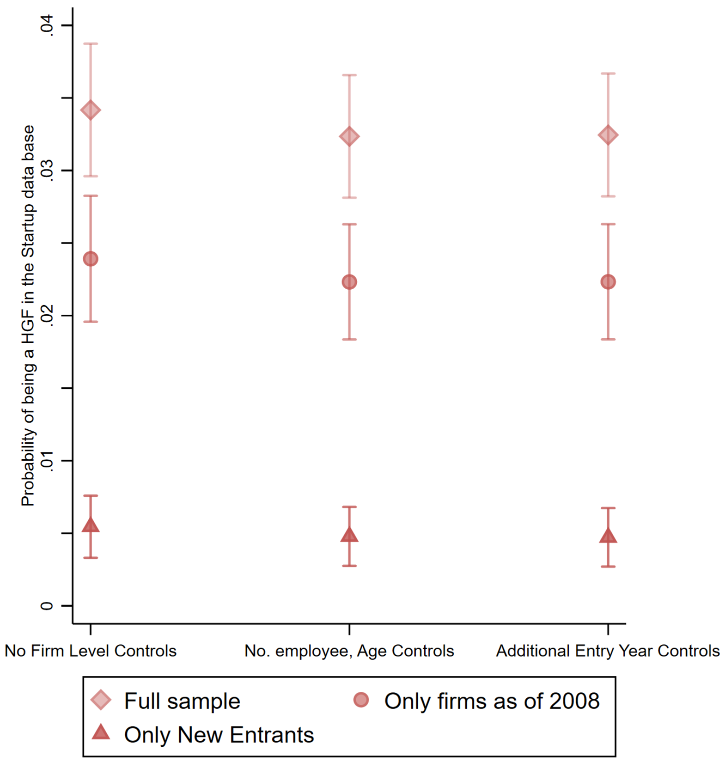
	Not in database	In database	Total
No HGF	210,585	2,145	212,730
Ever HGF	7,235	347	7,582
Total	217,820	2,492	220,312

B Complementary Regression results

In this section we present complementary regression results to section 5. We show all regression results for the alternative variable making use of revenue growth instead of employment growth.

B.1 Incidence of HGFs (revenue growth)

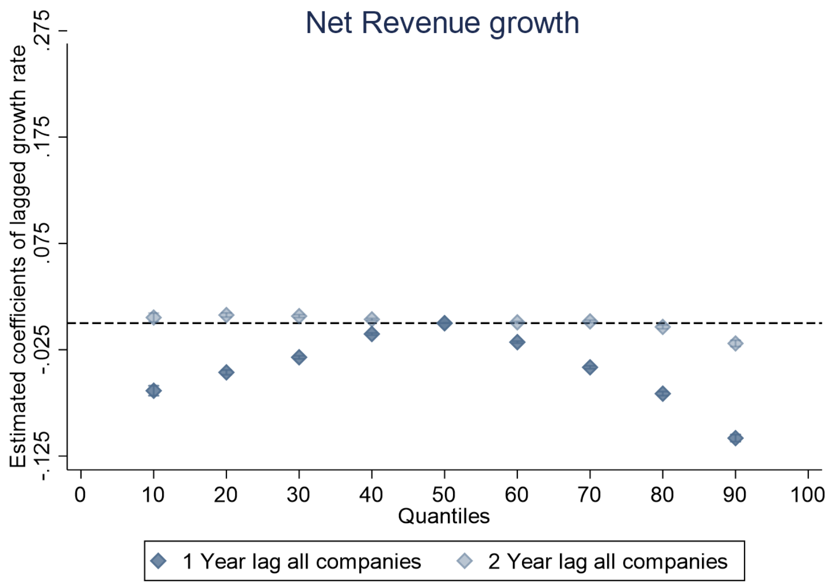
Figure 9: Characteristics of HGFs



Note. The figure shows regression results of linear probability models from equation 1. Point estimates of the startup base indicator variable ($\hat{\beta}$) from three samples are shown: 1) Full sample 2) Only firms as of 2008 3) Only new entrants. The vertical lines are the 95% confidence intervals.

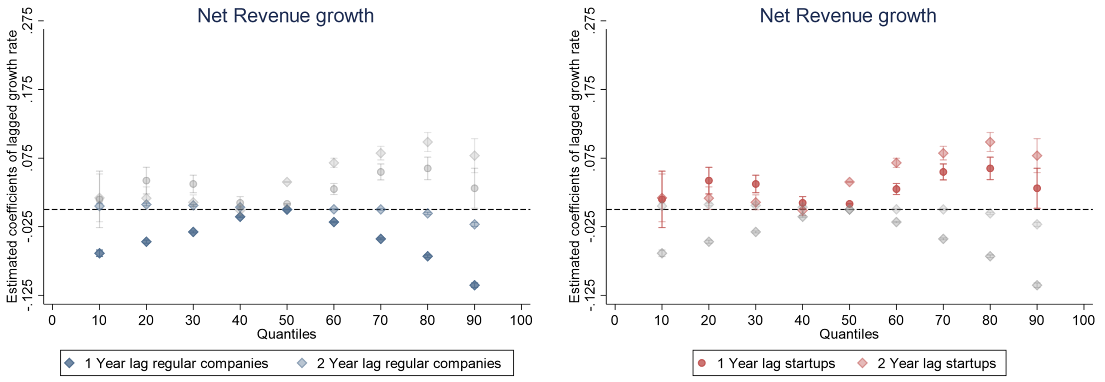
B.2 Growth Dynamics (revenue growth)

Figure 10: Dynamics of growth rates for the full sample



Note. The figure shows results of the quantile regressions which are described in equation 2. Dark blue diamonds refer to the estimated coefficient of the first lag of the change in the growth rate $\hat{\delta_1}$. Light blue diamonds show the estimates of the second lag of the change in the growth rate $\hat{\delta_2}$.

Figure 11: Dynamics of growth rates of non - startups compared to startups (revenue growth)



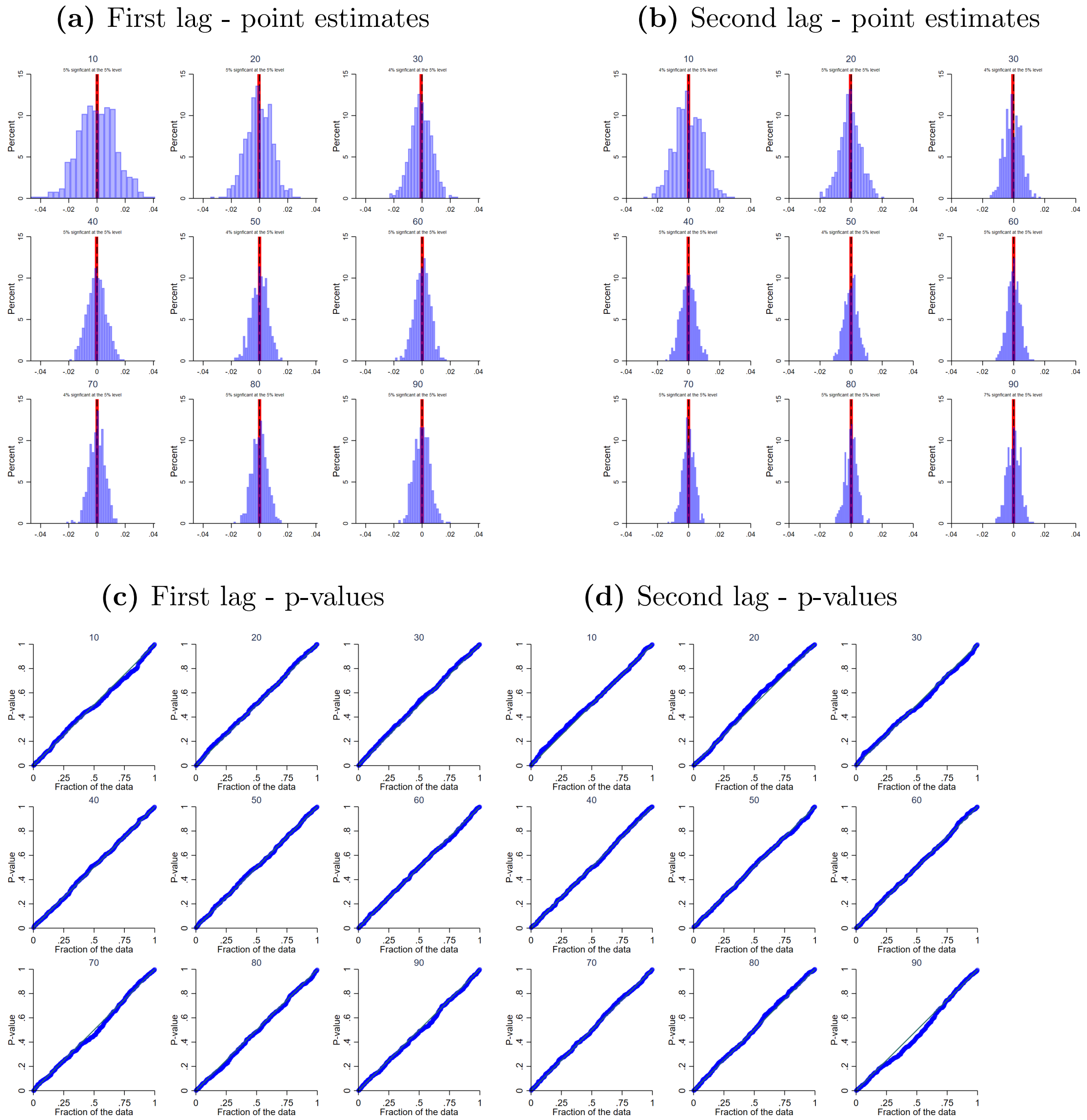
(a) Non-Startups

(b) Startups

Note. The figure shows the estimated coffecients of 3. Panel 11a shows the estimated lags of each quantile for 'regular' firms. Panel 11b shows the following estimates: $\hat{\kappa_1}$ (dark red) and $\hat{\kappa_2}$ (light red)

C Simulation Results

Figure 12: Simulated lagged log changes



(a) First lag - point estimates

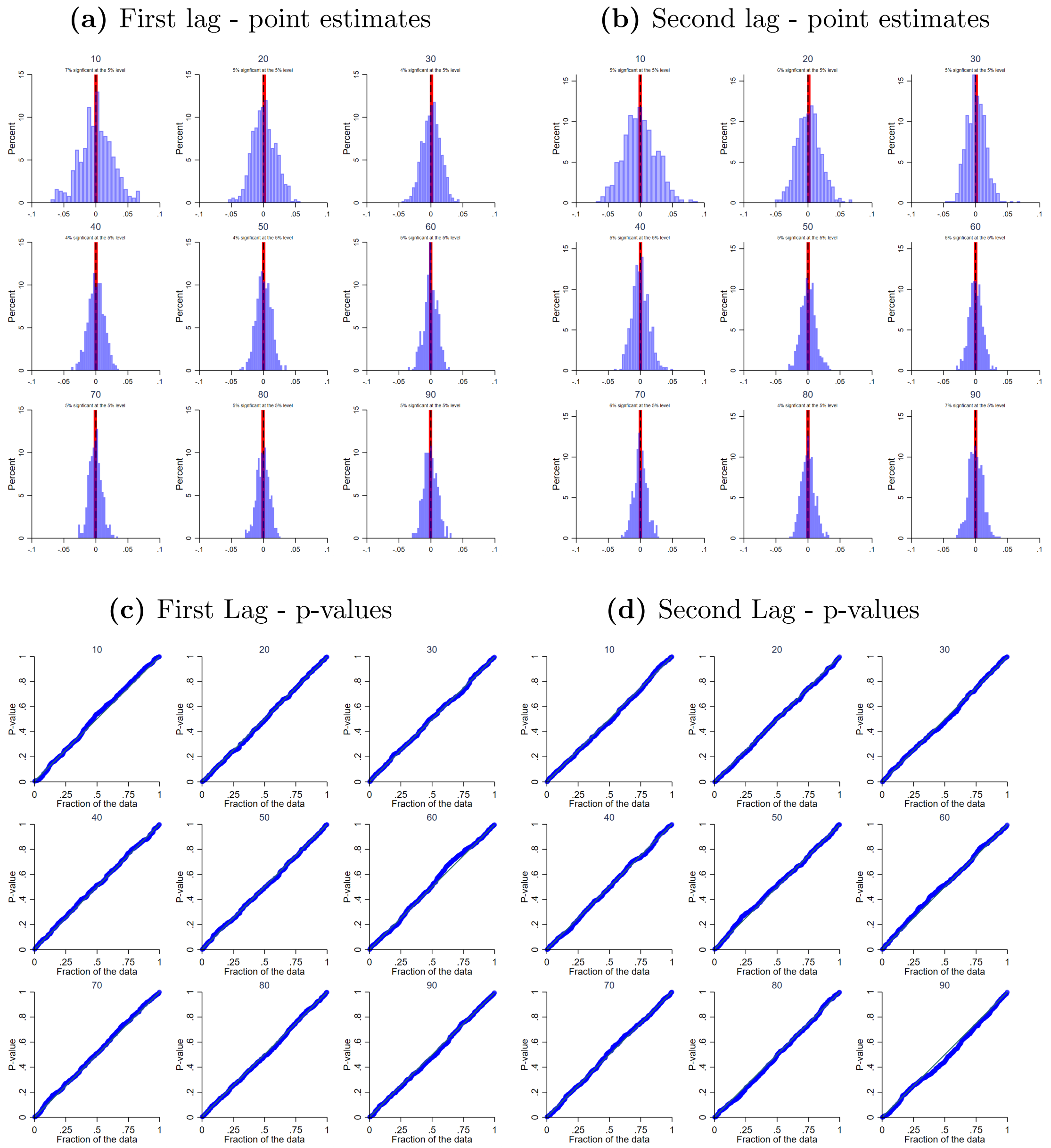
(b) Second lag - point estimates

(c) First lag - p-values

(d) Second lag - p-values

Notes. The figure shows regression results per quantile based on 500 simulated data sets of the random variable y_{it}. The variable is drawn from a Weibull distribution with scale 1 and shape 10,000. For each dataset we simulate 10,000 'firms' and 10 'periods'. We then run the following quantile regressions: $Q_\tau(\Delta Y_{it}) = \beta \mathbb{1}[D_i] + \sum_{l=1}^{L} \delta_l \Delta Y_{it-l} + \sum_{l=1}^{L} \kappa_l \Delta Y_{it-l} \times \mathbb{1}[D_i] + \zeta Y_{it-1} + \epsilon_{it}$ For the quantiles $\tau \in \{0.1, 0.2, ..., 0.9\}$, $L = 2$ and $\Delta Y_{it} = \ln y_{it} - \ln y_{it-1}$. In Panel 12a we show the distribution of $\hat{\delta_1}$. In Panel12b we show the distribution of $\hat{\delta_2}$. The red vertical line indicates the mean of the estimates and the dashed line is positioned at 0. Panel 12c and 12d show the distributions of the p-values in the form of quantile plots.

Figure 13: Simulated lagged log changes for random startups



(a) First lag - point estimates

(b) Second lag - point estimates

(c) First Lag - p-values

(d) Second Lag - p-values

Notes. The figure shows regression results per quantile based on 500 simulated data sets of the random variable y_{it}. The variable is drawn from a Weibull distribution with scale 1 and shape 10, 000. For each dataset we simulate 10, 000 'firms' and 10 'periods'. We then run the following quantile regressions: $Q_\tau(\Delta Y_{it}) = \beta \mathbb{1}[D_i] + \sum_{l=1}^{L} \delta_l \Delta Y_{it-l} + \sum_{l=1}^{L} \kappa_l \Delta Y_{it-l} \times \mathbb{1}[D_i] + \zeta Y_{it-1} + \epsilon_{it}$ For the quantiles $\tau \in \{0.1, 0.2, ..., 0.9\}$, $L = 2$ and $\Delta Y_{it} = \ln y_{it} - \ln y_{it-1}$. In Panel 13a we show the distribution $\hat{\kappa_1}$. In Panel 13b we show the distribution of $\hat{\kappa_2}$. The red vertical line indicates the mean of the estimates and the dashed line is positioned at 0. Panel 13c and 13d show the distributions of the p-values in the form of quantile plots.

We simulate panel data with one random variable y which follows a Weibull distribution with shape parameter 1 and scale parameter $10,000$.[12] We chose the Weibull distribution because it resembles the distribution of employees and revenue in our sample: it is strongly right-tailed and jumps within the distribution are more likely to resemble the nature of our growth distribution. We simulate 500 data sets with $10,000$ firms and 10 time periods. For each simulation, all observations a drawn independently from the above mentioned Weibull distribution. We then construct the exact same variables as in section 4.2 and run quantile regressions as in equation 3 on each of the 500 simulated datasets.

Figure 12 shows the results of the simulations. Panel 12a shows the distributions of the estimates of the first lag ($\hat{\delta_1}$) and Panel 12b of the second lag ($\hat{\delta_1}$) for each quantile. Above each panel we indicate the percentage of estimates which are significant at the 5% level. Panel 12c and 12d shows the distribution of the respective p-values in form of quantile plots.

Our simulations reveal no systematic relationship between the first and second lagged differences of the dependent variable. The estimates are concentrated around zero. As expected about 5% of the estimates are significantly different from zero. The distribution of the p-values tends to be uniformly distributed between 0 and 1 (Panels 12c and 12d)

[12] More formally, the function looks as follows, for each draw of a random variable y_i:

$$f(y_i) = \begin{cases} \frac{1}{10000} e^{-\frac{y_i}{10000}}, y_i \geq 0, \\ 0, y_i < 0, \end{cases}$$

The conclusions do not change if we alter the shape of the distribution or take a normal distribution.

In a last step we interact the lags with a random indicator variable, which takes the value of one for 10% of our sample. We then interact this random variable with the lagged growth. The intuition behind this approach is, that if a random variable was spuriously picking-up similar growth patterns as our startup dummy variable in the real data, the simulations would reveal that.

Figure 13 shows that there is no systematic pattern in the interactions with a random variable and the lagged growth paths. The point estimates are centered around zero (Figure 13a and 13b) and the p-values of the estimates uniformly distributed in an interval between 0 and 1 (Figures 13c and 13d).

D Construction of the Dataset

In this section we provide a description on how we created the database.The procedure consists of 4 major steps in which we combine existing microdatasets from Statistics Netherlands (CBS) with our (partly) hand collected data on startups. For the startup database, we first merged the information on the companies with the information available in the database of the Dutch Chamber of Commerce (KVK). Not all companies could automatically be matched and we manually searched all remaining trade registry numbers online. In total we were able to merge $3,935$ unique companies with our data base.

Table 8: Construction of the Dataset

	Data Description	Files used
1	Merge information of events (Exits, entries, mergers & acquisitions) on company group and company-group-entity level per year (2009-2018).	General company registry (henceforth ABR) and all its sub-files: `OG_ABR`, `BE_ABR`, `BE_OG_ABR`, `BE_persoon`, `BE_eventbijdragen`, `OG_eventbijdrage`.
2	Append years 2009 - 2018 of all company-group-entity combinations. Consolidate the information on number of employees, industry, age on the level of the company group.	ABR
3	Merge list of encrypted trade registry numbers of startups with company register. We obtain an unbalanced annual panel on the company-group level with information on: an indicator if the company group is matched with the startup database, number of employees, industry, age.	Startup data base with encrypted trade registry numbers and ABR (identifier variable is *vep_kvkdossiernummer*)
4	Consolidate balance sheets and earnings and loss statements of all limited liability companies in the Netherlands on the level of the company group ($rog\_identificatie$) and merge with panel of the company groups.	NFO 2009-2018, ABR 2009-2018