J Ind Compet Trade (2018) 18:253–294
DOI 10.1007/s10842-017-0257-0

Forecasting European high-growth Firms - A Random Forest Approach

Jurij Weinblat[1]

Received: 6 February 2017 / Revised: 7 July 2017 /
Accepted: 18 July 2017 / Published online: 10 August 2017

Abstract High-growth firms (HGFs) have aroused considerable interest both by researchers and policymakers mainly because of their substantial contribution to job creation and to the advancement of the surrounding economy (Acs et al., Small Bus Res Summ (328):1–92 2008, Schreyer 2000). Any initiative to foster HGFs requires the ability to reliably anticipate them. There seems to be a consensus in previous mainly regression-based studies on the impossibility of such a prediction (Coad, Doc Trav Centre d'Econ Sorbonne 24:1–72 2007b). Using a novel random forest (RF) based approach and a recent data set (2004–2014) covering 179970 unique firms from nine European countries, we show the potential of a true out-of-sample prediction: depending on the country, we were able to determine up to 39% of all HGFs by selecting only ten percent of all firms. The RF algorithm is both used to determine relevant predictors and for the actual prediction and pattern analysis. Both the selection of the best RF and the cross-country comparisons are based on a Receiver Operating Characteristic analysis. We find that most accurate HGF predictions are possible in GB, France, and Italy and largely confirm this ranking using Venkatraman's unpaired test. Apart from the firm's size, age, and past growth, the sales per employee, the fixed assets ratio, and the debt ratio are quite important. Our "typical" HGFs determined using RF prototypes have been older and bigger than the remaining firms, which is counterintuitive and atypical in literature. Based on our finding, typical HGFs are not start-ups, which questions current political funding strategies. Apart from that, our results do not support and rather refute the existence of a survivorship bias. Moreover, approximately every fourth HGF remains to be a HGF in the next period.

Keywords High-growth firms · Random forest · Forecasting · Variable importance

JEL Classification M13 · L25 · O51 · O52 · C53

✉ Jurij Weinblat
jurij.weinblat@uni-due.de

[1] Faculty of Economics and Business, Chair of Statistics, University of Duisburg-Essen, Universitaetsstr. 12, 45117 Essen, Germany

1 Introduction

High-growth firms, which are also known as "gazelles" (Lopez-Garcia and Puente 2012, p. 1029), are an intensively researched topic which is closely observed by policymakers (Daunfeldt and Halvarsson 2015).[1] There are several explanations for this considerable research interest. From the perspective of various firms, growth is a crucial strategic priority and is often seen as an indicator of corporate success and market acceptance (Shin et al. 2005, p.6; Barringer et al. 2005, p. 665). Especially for new firms, rapid growth is essential for survival (Coad 2007b, p. 51). Concerning their impact on job creation, it is frequently mentioned that HGFs create a substantial number of long-term jobs (Coad 2007a, p. 81; Acs and Mueller 2008, p. 86; Coad et al. 2014b, p. 92; Lopez-Garcia and Puente 2012, p. 1030).[2] The hired people are oftentimes long-term unemployed or immigrants so that HGFs help these people to get a foothold in the job market (Coad et al. 2014b, p. 293). It is impressive that approximately three percent of all firms are responsible for the majority of private sector's revenue growth (Acs et al. 2008, p. 2). Oftentimes, this growth is sustainable since a considerable share of HGFs manages to stay a HGF in the following period (Acs et al. 2008, p. 8).

For private investors, HGFs are also an attractive investment opportunity: especially big firms are more likely to continue to grow after such a period than the remaining firms and therefore, continue to be a promising investment (Acs et al. 2008, p.8, p. 18, p. 27, p. 46).

Another motivation to study HGFs is put forward by Schreyer (2000) and also Acs et al. (2008) who regard these firms' potential to advance national economies. Especially in some European countries, unemployment rates are relatively high which is due to a lacking ability of the economies to adapt to ongoing changes. The authors argue that although governments can create a favorable environment, it is up to private companies to carry out the required implementations. HGFs are believed to be especially capable of adapting to change, are willing to take hazards, and show higher-than-average research and development (R&D) expenses to finance important innovations. Birch and Medoff (1994, p. 163) have also stressed the HGFs' innovative capacity. Companies like Apple and Cisco are frequently mentioned examples for such HGFs (Barringer et al. 2005, p. 664). Especially technology based HGFs are known to have positive spill-over effects and to have a higher propensity to invest in R&D (Schreyer 2000, p. 1; Fotopoulos and Louri 2004, p. 163). Consequently, HGFs foster an increase in productivity and competition. There is evidence that this competition creates more jobs than it destroys (Acs et al. 2008, p. 11 and p. 19).

To reduce unemployment and because of other previously mentioned reasons both the European Commission and the OECD have taken measures to foster HGFs (European Commission 2010, p. 14; Organisation for Economic Co-operation and Development 2010; Coad et al. 2014b, p. 93, Schreyer 2000, p. 6). A mechanism to anticipate future HGFs would allow to well-directedly support HGFs without wasting taxpayers' money and increasing inflation (Birch 1981, p. 4). Current HGF related initiatives invest money in start-ups without knowing if they will ever become HGFs which is criticized by Acs et al. (2008, p. 7). However, Birch (1981, p. 8), Schreyer (2000, p. 29), Coad et al. (2014a, p. 91), and Acs et al. (2008, p.45) point out that a priori, future HGFs are impossible to anticipate because of their high heterogeneity and their nonuniform growth development which hampers the

[1] Barringer et al. (2005, p. 666) believe the literature in this area to be "rich and mature". According to Coad et al. (2014b, p. 93), the number of studies in the area of HGFs has "exploded" in the recent past.

[2] Middle-sized HGFs with 20–500 employees seem to contribute most in this context (Acs et al. 2008, p. 25).

identification of common characteristics. Coad (2007b, p. 1 and p. 56) even believes that "Firm growth is characterized by a predominant stochastic element, making it difficult to predict" and that it is an "idiosyncratic and fundamentally random process". This is why Coad and also Acs et al. (2008, p. 40) believe that it is challenging to implement HGF related policies.

In this context, this study aims to evaluate the capability of the random forest (RF) algorithm to predict future HGFs using publically available financial data. Therefore, we follow the appeal of Coad (2007b, p. 1) to apply novel statistical techniques in the area of HGFs. To provide a realistic impression, the out-of-sample prediction quality is evaluated on a more recent data set. Moreover, since the RF's byproducts allow to analyze the data's underlying patterns, the second contribution is to present and discuss these findings. For instance, an analysis of variable importance rankings can foster the identification of the previously mentioned common characteristics or patterns. According to Coad (2007b, p. 58), there was only a little progress in identifying the determinants of firm growth, making this analysis a valuable contribution. The analysis is based on nine countries, covering eleven years (2004–2014) and 179970 unique firms. Therefore, the results are robust and enable cross-national comparisons. In contrast to studies like (Schreyer 2000), the data originates from the same source so that differences cannot be attributed to different data collection and processing methods.[3] As it will become obvious in the literature review, previous empirical analysis either regard the influence of single variables on (high) firm growth or perform regression and panel data estimations. Although these studies were able to deliver various interesting insights into the determinants of HGFs, none of them evaluates whether these insights enable the prediction of future firm growth based on new data.

Our last contribution will be an analysis of the so-called "survivorship bias" (cf. Coad et al. 2014a, p. 96). Based on our data, such bias cannot be confirmed.

Hence, this study is supposed to introduce the predictive paradigm to the research area of HGFs. This paradigm is already common in other financial data applications. For instance, in the tradition of Altman (1968) and Ohlson (1980), scientists try to predict future corporate defaults using statistical techniques like the multiple discriminant analysis and the logistic regression. In the recent past, data mining approaches like support vector machines (Härdle W et al. 2005), artificial neural networks (ANNs) (Cross and Rarnchandani 1995) and RFs (Kartasheva and Traskin 2011; Behr and Weinblat 2017) are applied as well, usually outperforming the previously mentioned techniques. Another application of forecasting is the corporate financial performance prediction for instance in Lam (2004) using ANNs. In this study, ANNs enable to derive investment decisions which outperform the average market return.

The remaining article is structured as follows. In Section 2 we present important HGF related findings based on descriptive and regression studies. In Section 3 we introduce our data source, the structure of our train and test data and our used growth indicator. Section 4 describes the RF algorithm, how its prediction performance can be quantified, and the proceeding of our main analysis. The selected prediction variables following this proceeding are presented and described in Section 5 while our results can be found in Section 6. A summary and directions for further research are given in Section 7.

[3] Schreyer (2000, p. 7 and 39) acknowledges that his data originates from different sources like surveys, trade registers and commercial data bases, covers not identical periods and is, therefore, hard to compare. Moreover, Acs et al. (2008, p.8) admit that they do not know if their results are country-specific.

2 Literature Review

This section provides an overview of the body of literature on firm growth. Studies, which are mentioned here, are revisited in Section 5.1 to identify promising predictors for HGFs. Since there is currently no literature following the predictive paradigm in the area of high-growth firms, we will focus on studies which conducted panel model estimations or contributed otherwise to the research in this area.

Using data for western Germany, Boeri and Cramer (1992) stated that firm growth could usually be observed for young firms. About two years after foundation, these firms were found to have doubled their initial size. However, the growth rate decreased over time and was highly unstable. A higher-than-average growth rate was often followed by a lower-than-average rate in the consecutive year. The longer a company's number of employees remained in a certain size range, the more likely it became that it is going to stay in this range. Additionally, the authors identified a tough selection process for such young firms: the risk to fail was highest during the first years and lessened afterwards. Eight to nine years after birth, only approximately 40% of all firms founded in the same year still existed. The jobs created by the surviving firms still overcompensated the job losses of the closed firms. Boeri and Cramer (1992) could not determine a sector-effect.

The empirical results obtained by Schreyer (2000) are based on cross-national firm-level data from Germany, Italy, Netherlands, Spain, Sweden, and Canada. The data originate from different sources like surveys and commercial data bases and cover a maximal space of time from 1985 to 1996. The study's objective is the identification of distinctive features of HGFs. Similar to Boeri and Cramer (1992), Schreyer (2000) pointed out that small HGFs are responsible for a significant percentage of job creation (but not of total employment). The bigger the firm, the more its number of employees stabilized. However, relatively big HGFs were found to create considerable numbers of jobs, which questions the political focus on rather small firms. Concerning the sector effect, Schreyer (2000) observe that HGFs in all sectors but certain sectors contained a higher share of HGFs.[4] In this context, it is necessary to discriminate between growing firms and HGFs since both kinds of firms are concentrated in certain sectors but often not in the same ones. Regarding the firm's location, the number of HGFs was usually proportional to the overall number of firms in the geographical area. However, a few exceptions like Paris in the case of France with an above average share of HGFs were observed.

Moreover, HGFs were found to be younger than the other firms, more often owned by other companies and investing substantially in R&D. Country-specific differences became obvious regarding the relationship between age and growth. In Spain, older firms did not turn out to be less likely to grow intensively. In contrast to that, young German and Dutch firms rather tend to grow extensively than older ones. The finding that HGFs are oftentimes not economically independent is explained by an easier access to e.g. financing, staff and market knowledge.

A rather different approach was conducted by Barringer et al. (2005) who performed a textual analysis using 100 descriptions of the winners of the Ernst & Young LLP Entrepreneur of the Year award complemented with some financial data. Based on this data set, they intended to extract characteristics from the four areas "founder characteristics", "firm attributes", "business practices" and "human resource management practices" which

[4]The author explicitly mentioned "knowledge-intensive service industries", "education" and "health care". In Germany, the manufacturing industry contained a below average number of HGFs (Schreyer 2000, p. 22).

were frequently mentioned for rapidly growing firms and less so for slowly growing firms. Several characteristics from all four areas have been found to distinguish rapidly growing firms. For instance, such firm's founder had a high impact especially if she had a college education, was experienced in her industry and showed an exceptional motivation to succeed. Moreover, the firm's participation in interorganizational relationships and products, which creates a unique value, were beneficial. Concerning the staff, training and financial incentives were found to foster fast growth.

Analyzing North American firms from two data bases, Acs et al. (2008) divided their sample into three subsets: 1994–1998, 1998–2002 and 2002–2006. Firms, which expanded both revenues and the number of employees, have been identified in the second period; the remaining periods have been used to study the characteristics of these firms before and after this expansion.[5] Unlike Boeri and Cramer (1992), the authors differentiated their analysis by firm size. It was discovered that the majority of big expanding firms (500 employees and more) were expanding firms in the previous period and continued to grow afterwards. In contrast to this, expanding firms with less than 20 employees were found to be highly volatile since only less than ten percent of them expanded in the previous period and most of them declined after expansion. The authors could not find any considerable sectoral or regional effect. One interlocational finding was that most expanding firms were located 6 to 15 miles from the central business district. Moreover, it was confirmed that such expanding firms tend to be smaller and younger than the remaining ones. The age appears to have a higher impact than firm size. This does not mean that expanding firms are typically start-ups. In fact, only about six percent of all expanding firms were start-ups. This contradicts the findings of Birch (1981). The average age of an expanding firm increased with the firm's size and turned out to be 17 for small firms and 34 years for big firms. Expanding firms were also found to be 40% more efficient (measured in revenue per employee) than non-expanding firms.

Lopez-Garcia and Puente (2012, p. 1029) pointed out that the previously mentioned univariate studies may be misleading due to multivariate dependencies. The following studies considered interdependencies of several variables simultaneously and therefore have a higher explanatory power.

Apart from frequently used variables like firm size, age and sector, Harhoff et al. (1998) have also included the firm's legal form in their OLS and two-stage Heckman regression. The data set stems from the Creditreform[6] database covering 8,068 firms from West Germany starting with the year 1989. The data set was supplemented by telephone interviews. On average, private limited liability firms showed both a 4.5% higher growth rate than a single proprietorship but also a higher rate of bankruptcy. These results are explained by different tax liabilities, financial accountabilities, and ownership structures, which come with various legal forms. For instance, owners of proprietorships are liable without limitation whereas limited liability firm owners only risk the value of their equity. Public limited companies did not show any clear differences to proprietorships as far as growth is concerned. The previously mentioned firm's independence status only had a significantly positive effect in construction and trade industries and turned out to be insignificant for manufacturing and service.

[5]These firms will be denoted as "expanding firms" here to distinguish them from HGFs which only have to increase the number of employees.

[6]Creditreform is a commercial credit reporting agency covering firms from mainly European countries. The database covers a substantial amount of financial data for German firms.

Becchetti and Trovato (2002) regard 4,000 Italian small and medium-sized enterprises (SMEs) between 1989 and 1997. Referring to the study by Harhoff et al. (1998), the authors also studied the effect of the ownership structure on firm growth using a multivariate approach. In contrast to the previous study, a significant effect of the ownership structure could not be discovered. Instead, an effect of the accessibility to credit capital was shown for the first time. Not granted loans impede future growth and government subsidies foster it. Whether a firm exports their product also had a significant positive influence.

To analyze the determinants of the probability of high-growth, Lopez-Garcia and Puente (2012) applied a dynamic probit analysis on panel data taken from the local National Institute of Statistics containing 1,411 Spanish trading companies between 1996 and 2003. Growth was observed over a period of one year. Hence, controlling for other potential determinants is inevitable. Besides that, the used random effect model controlled for past growth to avoid interpretation problems caused by autocorrelation. This autocorrelation has been, indeed, found in their study: high-growth firms were 14% more likely to also be a high-growth firm in the following year than a non-high-growth firm. Additionally, it was discovered that a paid wage premium also increased a firm's probability to be a high-growth firm stressing the importance of human capital. Similar to Schreyer (2000), the authors found it necessary to distinguish between average growth and high growth since the amount of financial debt seemed to only restrain average growth but not high-growth firms.

Autocorrelation has also been studied by Coad (2007a) by analyzing 10000 French manufacturing firms from 1996 until 2002 using quantile regression. In contrast to Lopez-Garcia and Puente (2012), it was analyzed whether the autocorrelation depended on the firm's size and growth rate. Coad (2007a) showed for his sample that above-average growth was usually followed by a poor growth in the following period and vice versa. When disaggregating for ten different firm sizes, it was observed that big firms had a slightly positive autocorrelation while small firms showed the previously mentioned negative autocorrelation. Hence, bigger firms were often able to grow over several periods. The results were justified by the long-term planning horizons of big firms, which might lead to long-lasting growth and are in line with the results of Jovanovic (1982).

In a survey, Coad (2007a) compared different autocorrelation studies concluding that "there does not appear to be an emerging consensus" as far as the strength and lag are concerned (Coad 2007b, p. 16). A more robust finding is the decrease of both the growth rate and its variance with increasing firm's age. Concerning innovations, Coad (2007b) identified contradictions in the regarded literature but recognized that product innovations usually lead to greater employment while process innovations have an unclear effect. No strong influence is found for a firm's financial performance. The same was found for relative productivity, which is either explained by firms which downsize to increase productivity or by a lack of competition. Coad (2007b) found evidence for the existence of the previously mentioned industry effect. For instance, overall growth in a certain industry was also beneficial for this sector's firms. However, the sector's explanatory power is rather small. Besides these factors, the author mentioned other determinants like country-specific legal regulations and different taxation requirements for certain sizes. Moreover, a highly uncertain future demand also seemed to affect the firm's efforts to grow.

The ability of HGFs to remain a HGF in the next periods continues to be a highly researched area. Recently, Daunfeldt and Halvarsson (2015) regarded an almost complete sample of Swedish firms between 1997 and 2008 split in three consecutive periods. A quantile autocorrelation model was applied. They found only a negligible probability for this

recurrence. This is why they believe HGFs to be "one-hit wonders". Hence, they concluded that policy initiatives to foster HGFs are likely to be useless.

Levratto et al. (2010) analyze high-growth firms separately from average-growth firms. The authors regarded 12,811 French manufacturing small and medium-sized enterprises (SMEs) from 1997 through 2007. The data originated from Bureau Van Dijk and the French National Institute of Statistics and Economic Studies. They identified that high firm growth is more likely to be found in Paris than in other French regions. Being an exporting firm had no significant effect on high-growth firms but a positive and significant effect on the remaining firms. One of the most important variables appeared to be the corporate structure: main firms of a group grew faster than controlled subsidiaries and independent firms. Contrary to Gibrat (1931) who assumes that firm growth is random, Levratto et al. (2010) believe that growth is influenced by structural variables such as the firm's sector and age and strategic choices like the firm's financial structure. The underlying analysis was carried out using a pooled multinomial logit model and a hybrid multinomial logit model which are able to capture non-linear interdependencies. For all firms, it was found that growing firms in general were rather young and small regarding the number of employees. In contrast to Boeri and Cramer (1992), a sector effect could be discovered. As far as the strategic choices are concerned, labor productivity and the share of obligations to the supplier to total liabilities were found to be positively correlated with growth.

In summary, the research of HGFs remains to be a much-publicized research topic. The determinants of HGFs and their impact on the economy as well as their persistence are often contradictory. While most authors seem to agree on the importance of age and firm size, the influences of the sector, legal form, and financial variables are controversial. Moreover, the models are not evaluated on out-of-sample data so that overfitting is likely leading to results, which are strongly influenced by the peculiarities of the used data sets and are not generalizable. Furthermore, cross-country studies, which are based on data sets from the same source, are rare. This is why we estimate country-specific RF models for nine countries and determine their out-of-sample performance on more recent data. The RF's variable importance ranking enables us to compare the determinants of HGFs leading to results which are more robust.

3 Data Base and Identification of High-growth Firms

We use accounting records from the Amadeus data base. In this section, we provide a short description of this data source and explain how we define a high-growth firm.

3.1 The Amadeus Data Base

The Amadeus data base is generated by 'Bureau van Dijk'(BvD). Amadeus comprises information for both large firms and SMEs and contains data about East and West European countries with a focus on nonincorporated firms facilitating international comparisons (Van Dijk Electronic Publishing GmbH B 2015).[7]

For our analysis, we use the finance data and the master file data for Finland (FI), France (FR), Germany (DE), Italy (IT), Portugal (PT), Spain (ES), Great Britain (GB), Poland (PL)

[7]There are still differences regarding the country-specific reporting procedures, which complicate cross-national comparisons.

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Train data	$t-2$	$t-1$	t	$t+1$	$t+2$	$t+3$	$t+4$				
	Observation period			Growth or bankruptcy period							
Test data					t^*-2	t^*-1	t^*	t^*+1	t^*+2	t^*+3	t^*+4
					Observation period			Growth or bankruptcy period			

Fig. 1 Overview of the analyzed years in the train and test data sets

and Sweden (SE). We converted figures from British Pound, Swedish Krona, and Polish Zloty to Euro using the 2014 exchange rates. The selection of countries is due to the lack of data for other countries.

For every country, we created two data sets: a train data set and a test data set. The train data sets are used to identify the most promising corresponding country-specific RFs. As shown in Fig. 1, each train data set contains predictive variables from 2004 to 2006. These variables are presented in Section 5.1. Growth and a potential bankruptcy is observed between 2007 and 2010.[8] The more recent test data sets contain the same predictive variables from 2008 to 2010 and growth or bankruptcy indicator variables for the period between 2011 and 2014. The test data sets are only used to evaluate the out-of-sample prediction performance of the models. Such a set-up simulates a prediction at the end of year 2010 based on all the necessary data to both generate the RFs using the train data and apply them to the (already available) financial data of the corresponding test data set to predict which firms are likely to grow in the upcoming four years.[9] An overview of the numbers of firms per country and the shares of HGFs and insolvent firms is given in Table 1.[10]

Apart from removing incomplete cases, we applied additional selection criteria to improve the data quality. For instance, we ensured that every regarded balance sheet covers exactly 12 months. Furthermore, we defined sensible intervals for the predictor variables to prevent implausible observations biasing the results. E.g., we checked that each firm's turnover, short-term and long-term debts and interest payments were positive. While RFs are known to be rather robust towards the unavoidable arbitrariness of the data cleaning process, erroneous data are still disadvantageous (Williams 2011, 246). Furthermore and similar to Coad (2007a), this study only regards "organic growth" which is not caused by merger or take-over. Therefore, such firms have also been removed.

3.2 Identification of High-growth Firms

To estimate prediction models and to evaluate their performance, the information on firm's growth status is crucial. This information is generated from the data contained in the Amadeus data base. A firm is regarded to be a HGF if its Birch-Schreyer growth indicator[11]

[8] As explained in Sections 5.2 and 6.3, one research question of this article is to analyze whether a higher propensity to be a HGF is associated with a higher propensity to default as often claimed in literature. In this study, both growth and bankruptcy is observed within the same periods.

[9] Of course, to make a general statement of the predictability of HGFs in these countries, an analysis of more data sets is inevitable. Still, this study can be considered to be a proof of concept and initiates further analysis.

[10] The relative frequency in the HGF column is usually less than 10%, because the Total column includes both solvent and bankrupt firms. The 90% quantile is determined based on the solvent firms.

[11] In earlier articles, this index was also called "Mustar index" (Schreyer 2000, p. 40).

Table 1 Total, HGF and bankruptcy figures per country and data set

	Train			Test		
Country	Total	HGF	BANKR	Total	HGF	BANKR
DE	1906	182 (9.55%)	95 (4.98%)	3253	307 (9.44%)	198 (6.09%)
ES	55596	5313 (9.56%)	2669 (4.80%)	40172	3526 (8.78%)	4921 (12.25%)
FI	1905	191 (10.03%)	0 (0.00%)	1722	155 (9.00%)	180 (10.45%)
FR	13440	1292 (9.61%)	524 (3.90%)	4687	337 (7.19%)	1325 (28.27%)
GB	4124	394 (9.55%)	190 (4.61%)	3748	359 (9.58%)	161 (4.30%)
IT	31729	3017 (9.51%)	1703 (5.37%)	36206	3350 (9.25%)	2800 (7.73%)
PL	4294	427 (9.94%)	50 (1.16%)	416	42 (10.10%)	2 (0.48%)
PT	306	28 (9.15%)	35 (11.44%)	4736	354 (7.47%)	1199 (25.32%)
SE	17941	1792 (9.99%)	25 (0.14%)	12081	1167 (9.66%)	411 (3.40%)

value belongs to the top 10% of all firms from the same country and data set over a period of three years. Therefore, this study follows the proceeding of Schreyer (2000), Acs et al. (2008), and Lopez-Garcia and Puente (2012). The index is defined as follows in our study:

$$growth = (e_{t+4} - e_{t+1}) \cdot \frac{e_{t+4}}{e_{t+1}}.$$

e_t is the firm's number of employees in year t. For the train data set, growth is observed between $t + 1 = 2007$ and $t + 4 = 2010$. The corresponding interval thresholds for the test data are 2011 and 2014. Such a timeframe is common in the literature (Coad et al. 2014a, p. 95).

The target variable Y for a given data set ($ds \in \{train, test\}$) and a particular country c is, therefore, defined as follows:[12]

$$Y_{ds,c} = \begin{cases} 1, & \text{if } growth_{ds,c} \geq q_{0.9,ds,c} \\ 0, & \text{if } growth_{ds,c} < q_{0.9,ds,c} \end{cases}$$

$q_{0.9,ds,c}$ is the 90% quantile of all $growth$ values in the country c from data set ds. Firms with a growth value equal or above the 90% quantile are labeled as HGFs whereas the remaining firms are treated as low-growth firms (LGFs). It is important to point out that firms, which went bankrupt, are neither considered to be LGF nor HGF and are not considered for model estimation. Apart from that, firms of every age can become a HGF so that the study is not limited to start-ups. The shortcoming of our proceeding is that the value of $q_{0.9,ds,c}$ depends on the analyzed country and period which complicates cross-country comparisons. Another possibility would be to require that every HGF grows by a certain percentage (Hölzl 2014, p. 204). However, this would lead to different shares of HGFs for each country and period. Since the extent of a HGF's growth is affected by the sectoral and overall economic growth, this does not seem to facilitate comparisons either (Audretsch and Mahmood 1994, p. 243; Bravo Biosca 2010, p. 2).

The Birch-Schreyer growth indicator combines absolute and relative employment growth. This makes the indicator less dependent on the firm size than its components. Large firms usually have larger absolute changes in the number of employees than smaller firms

[12] We are no longer going to distinguish between t and t^* since this is obvious when considering whether the train or the test data set is used.

(absolute growth) whereas smaller firms rather can increase their staff by a given percentage than bigger firms (Schreyer 2000, p.14; Lopez-Garcia and Puente 2012, p. 1030). Because of these favorable properties, this is the most frequently used indicator (Coad et al. 2014a, p. 94). However, the Birch-Schreyer growth indicator is known to prefer absolute over relative growth (Hölzl 2014, p. 226). This is not necessarily a disadvantage because politicians will rather be interested in high absolute numbers of new jobs than in high relative employment increases in certain firms with possibly low impacts on the country's unemployment rates. However, studies like National Commission on Entrepreneurship (2011), Coad (2007b), and Harhoff et al. (1998) purely focus on relative growth.

Focusing purely on employment growth is common but has been criticized by Aiginger (2006) since similar policies might foster inefficient behavior. Furthermore, employment is not a parameter, which firms try to maximize (Levratto et al. 2010, p. 9). In fact, an employment growth based definition is not the only possible approach. Another frequently used measure is the firm's turnover growth[13] which has several deficiencies and which is the reason why we focus on employment growth (Daunfeldt et al. 2014). For instance, the definition of turnover depends on the observed country's accounting practice, which hampers cross-country comparisons. Since consecutive years are observed, country-specific deflation becomes necessary which leads to controversial assumptions (Lopez-Garcia and Puente 2012, p. 1035); Levratto et al. 2010, p. 9). Acs et al. (2008) even analyzed both employee and revenue growth at the same time. Daunfeldt and Halvarsson (2015) followed a similar approach. Further growth indicators are occasionally created based on the firm's value-added, total assets or total profit (Coad 2007b, p. 3).

According to Coad et al. (2014a, p. 99 – 104) and Hölzl (2014, p. 199), the choice of the HGF indicator is extremely important and has a strong influence on the obtained results. However, irrespective of the applied definition, HGFs achieve higher growth rates after their high growth period than a control group. This is why the dependence of the indicator does not question the HGF concept as a whole.

4 Methodological Considerations

In this section, the used prediction method, the measure, which quantifies the appropriateness of the estimated HGF propensities, and our proceeding are explained.

We use the RF algorithm to estimate HGF propensities for a given firm in a given country. We have chosen this particular algorithm because of several reasons. RFs were found to have a very promising prediction performance for default prediction which will become relevant in Section 6.3 (Behr and Weinblat 2017). Furthermore, the RF provides several byproducts which can assist the user in different stages of the data analysis process. As shown in the further course of this section, a RF helps to identify promising predictors, can assess their contribution to predict a HGF and can be used to derive "typical" country-specific HGFs and LGFs. Thus, the RF offers a powerful framework for HGF prediction.

A RF consists of a user-defined number B of classification trees, which are generated using a slightly modified classification and regression trees (CART) algorithm (Verikas et al. 2010, p. 331). As the generation of individual classification trees is the core of the RF and such a tree is presented in Section 6.2, we provide a brief description of the CART algorithm first.

[13] Such a HGF definition can be found in Barringer et al. (2005).

4.1 Classification and Regression Trees Algorithm

CART is a tree-based approach to generate qualitative (classification) or quantitative (regression) predictions for given observations (Breiman et al. 1984). Due to our specific context of HGF prediction, we only regard the classification application. CART starts with the complete data set and recursively splits it into two sub data sets, which should be as different as possible in their shares of HGFs and LGFs (Kumar and Ravi 2007, p. 3). Based on the value of a single variable, the observation is assigned to one of the two subjacent sub data sets. The variable used for splitting the data set and the specific threshold value are identified by considering all potential variables and all their observed values. To be more specific, the decision of which variable and which threshold value should be used, is determined by comparing the impurity of the class distribution inside the original node and the two subjacent nodes. The split that enables the highest impurity reduction is chosen. The impurity for a node t is quantified using the Gini coefficient I. In this context, $p(j|t)$ is the share of the members of the class j in this node (Breiman et al. 1984, p. 20–26)

$$I(t) = 1 - \sum_{j} p^2(j|t). \tag{1}$$

Leaves are the lowermost nodes of the tree. The majority class (binary prediction) in the reached leaf is predicted for a given observation (Breiman et al. 1984, p. 33; Frydman et al. 1985, p. 272).

The main advantage of such trees is that they are easy to comprehend and do not depend on any distributional assumptions. Furthermore, both quantitative and qualitative variables are suited as predictors (Breiman et al. 1984, p. 56; Chandra et al. 2009, p. 4832; Shirata 1998, p. 3). An important downside is that even minor changes in the data can lead to different decision trees. This is why CART models are considered to be weak learners having a small bias but a high variance.

4.2 Random Forest Algorithm

The basic idea of a RF is to reduce the variance without increasing the low bias. This is achieved by averaging over a large number of decision trees, which usually leads to a better forecasting performance (Breiman 1996, p. 123–140).. The concept of creating a user-defined number B of sub-models (decision trees in case of a RF) is called bagging which stands for "bootstrap aggregation". Each of the RF's trees is estimated based on its own data set, which is obtained by drawing a bootstrapped sample from the original data set. Each tree's data contain as many observations as the original data set. This means that about two-thirds of all observations will be considered by a tree once or more, and others will not be considered at all (Breiman 2001, p. 5; Zighed et al. 2000, p. 117–118; Hastie et al. 2009, p. 283, 587; Rokach 2007, p. 106–107)

Each of the B trees is grown using a slightly modified version of the CART algorithm. The modification affects the number of variables that are analyzed at each node to select the splitting variable and its threshold value. Instead of analyzing all available variables, the RF algorithm first randomly selects m variables for each node and only regards these selected variables when determining the best split. This is called "random subspace". Once again, the condition that enables the highest Gini value is chosen (Breiman 2001, p. 11; Han et al. 2011, p. 377–378). Similar to B, m is also defined by the user.

The estimated RF provides B individual binary predictions based on the B classification trees for a new observation in a first step. These predictions are called votes and are represented by $v_1 \ldots v_{100}$. The most frequently occurring class v is returned in the second step. A propensity can be obtained by returning the share of a certain class instead.

A model (forest) that embodies B sub-models (trees), each of which is estimated based on a sub data set and which combines their individual predictions towards an overall prediction is called "ensemble". Such an ensemble can usually better adapt to the data patterns than a single sub-model can.

Another of RF's beneficial features is that it is found to perform well in the case of highly imbalanced data as shown by Brown and Mues (2012). This is highly relevant for our study because, by definition, only a small fraction of all firms are HGFs. Yeh et al. (2014, p. 101) find the RF to be less affected by overfitting than the other considered models. Moreover, random subspacing speeds up the computation process considerably (Verikas et al. 2010, p. 331). Unfortunately, RFs are less transparent compared with a single classification tree due to simultaneously analyzing B classification trees (Olson et al. 2012, p. 464). However, this can be compensated by analyzing the RF's byproducts (global variable importance rankings and prototypes).

Based on the classification accuracy and the unused observations of the single trees (out-of-bag observations), a RF allows the estimation of a global variable importance ranking (Breiman 2001, p. 23–25; Verikas et al. 2010, p. 331–334; Strobl et al. 2007, p. 17–18). The underlying idea is that a permutation[14] of the value of an important variable should lead to a high increase in the error rate whereas permutating unimportant variables should not cause a considerable effect (Breiman 2001, p. 11; Breiman and Cutler 2004; Hastie et al. 2009, p. 593). The higher the significance value, the higher the importance of the current variable (Breiman 2001, p. 23–24; Verikas et al. 2010, p. 331). According to Verikas et al. (2010, p. 340), the RF's variable importance ranking has certain drawbacks like its inability to capture variable dependencies.

Prototypes are representative observations of every class. A prototype of a given class c can be derived by identifying an observation o of this class, which has the highest number of other class c observations among its w nearest neighbors. The distance between o and another observation l can be determined from the proximity matrix of the RF: $prox_{o,l}$. The further proceeding depends on the type of the variable. For a quantitative variable, the prototype's corresponding value is the median of the values of the w neighbors. For a qualitative variable, the prototype's value is the most common value of its w neighbors (Breiman and Cutler 2004).

4.3 Evaluation of the Classification Performance

In this study, it is necessary to assess the appropriateness of our RFs in an out-of-sample prediction setting. Predictive performance measures quantify this appropriateness based on the values from the so-called 'confusion matrix (Hart et al. 2005, 362). The components of this matrix are explained in Hassan et al. (2010). Based on the matrix components, performance evaluation measures like the accuracy, precision, true positive rate (TPR) and false positive rate (FPR) can be obtained. It is crucial to point out in this context that the accuracy

[14] In this context, permutation means to randomly redistribute the existing variable values of the analyzed variable among all the observations.

and precision measures are highly misleading for imbalanced data (cf. Maimon and Rokach 2006, p. 666 and Han et al. 2011, p. 365–367).[15]

To obtain informative indicators even in settings of highly imbalanced class distributions, it is important to only rely on measures from the same row of the confusion matrix. Two such metrics are the TPR and the FPR. Ideally, a RF should have a TPR of one and a FPR of zero which is usually unrealistic: a model which tends to classify a new firm as a HGF in case of doubt will usually achieve a high TPR (close to one) because it predicts many HGFs correctly. However, the FPR rate will also be quite high because many actual LGFs will also be treated as HGFs. A model which rather tends to classify firms as LGFs will have a low TPR rate but also a low FPR rate. This is why Pagans (2015, p. 156) mentions a trade-off between these two indicators. Another problem in this context is the arbitrary cutoff point up to which a firm is treated as a LGF. The choice influences all the previously mentioned indicators so that it is not clear whether unsatisfying indicator values represent a weak model performance or just a disadvantageous selection of threshold (Fawcett 2006, p. 861–864).

The Receiver Operating Characteristic (ROC) curve is a graphical tool to assess and select prediction models based on their prediction performance. It is adapted from the field of signal detection and addresses the two previously mentioned problems (Gorunescu 2011, p. 323). As it is the case for our RFs, the ROC requires a prediction model which can quantify how certain a given firm belongs to a particular class in a two-class-setting. The advantages of a ROC are that it does not require calibrated class probabilities and it is independent of uneven class distributions because it only relies on the TPR and FPR, which do not depend on class distributions either. Furthermore, it regards all possible thresholds instead of only one. However, it does not lead to a numerical value facilitating comparisons of several curves, which possibly lie close to each other or even intersect (Fawcett 2006, p. 861–867).

To obtain such a numerical value, the area under the ROC curve (AUC) can be calculated. In the case of an ideal model, this area will be one (area of the unit square) and in the case of the useless model, it is going to be 0.5 (half of area of the unit square). The closer the area of a model is to one, the rather does the model assign higher propensities to HGFs than to LGFs. More precisely, the AUC is the probability that the model assigns a higher propensity to a HGF than to a LGF (Fawcett 2006, p. 868).

In this study, one important challenge is to compare the performance of nine country-specific RFs on their test data sets. Unfortunately, it is not appropriate to just compare the corresponding AUC values because the differences might be caused by chance. A statistical test can be applied to ensure that the observed superiority of one country is unlikely to be random. The test must be applied in an unpaired setting because firms are taken from different data sets and are, therefore, uncorrelated (Krzanowski and Hand 2009, p. 107; Zhou et al. 2014). Such a two-sided test has been developed by Venkatraman (2000) extending the work of Venkatraman and Begg (1996). Both methods are based on permutation tests and compare the ROC curves instead of just the AUC values. The null hypothesis states that two analyzed ROC curves are equal.

4.4 Model Specification Based on Cross-validation and Variable Importance

After having explained the used prediction methods and a technique to evaluate the obtained results, the proceeding of the main analysis is presented in this section.

[15] We still report it because it is very popular in literature.

To conduct the analysis, several obstacles have to be overcome. The first one originates from the fact that the HGF status is only assigned to every tenth firm so that most of the firms belong to the low growing class. This is why it is necessary to cope with the class imbalance problem in this analysis. A whole branch of data mining literature deals with this topic, which goes beyond the scope of this article. Contributions have been made, for instance, by Chawla et al. (2002), Batista et al. (2004) and Chawla (2005). In this study, the Synthetic Minority Over-sampling Technique (SMOTE) by Chawla et al. (2002) is used[16]. Its main idea is to generate synthetic instances of the minority class, which lie between two existing adjacent minority class instances. This is achieved by randomly placing this instance on the connecting line of these two neighbors in the feature space. In contrast to rather primitive alternatives like random oversampling, SMOTE prevents the RF from overfitting by widening the regions of the minority class (Chawla et al. 2002, p. 328). In a comparison study of different approaches to the class imbalance problem, Batista et al. (2004) find that SMOTE delivers the best or competitive results for data sets with class distributions which are similar to our data sets' distributions. SMOTE is also attractive from an economical point of view because its created virtual firms can at least be seen as imaginable because their structural and financial information are derived from existing firms.

The second obstacle of this analysis is related to the problem that there is no consensus in the literature on the variables, which contribute to anticipating future high growth. Apart from variables like firm size, previous growth, age and sector (c.f. Section 2), the findings are oftentimes contradicting. Levratto et al. (2010, p. 2) add that financial, environmental, productive and technical aspects all contribute to future growth. In order not to overlook valuable predictors, it is necessary to evaluate a relatively wide range of potential predictors. Thus, the analysis starts with a total number of 30 potential predictors. Although the RF is known to be able to deal with a high number of predictors some of which are almost unrelated to the class membership (Han et al. 2011, p. 383), Breiman and Cutler (2004) recommend estimating a preliminary RF using all predictors. This RF is only used to estimate a first variable importance ranking. We chose the 15 most important variables and therefore follow Yeh et al. (2014) who applied a similar approach in the area of going-concern prediction. This means that separately for each of the nine countries, the training data is balanced using SMOTE. Afterwards, nine country-specific RFs are estimated based on this data, and the variable importance rankings are obtained. To achieve that country-specific differences are not caused by different sets of variables, the same 15 variables are selected for all countries although their rankings are not identical.[17] This is done by averaging each variable's country-specific ranks and selecting the variables with the 15 highest averages. The 30 initially chosen variables and a detailed presentation of the 15 selected variables can be found in Section 5.1.

An intuitive reason why not all 30 initial variables should be considered for the final analysis is the RF's mode of operation. The probability that the analyzed subset contains no

[16] An alternative would be to modify RF's class weights. We chose SMOTE instead because it can also be applied for predictions using other algorithms than RF.

[17] Since the country-specific rankings show similar tendencies (c.f. Section 6.2), such a proceeding should not worsen the results all too much. However, for practical applications, it is reasonable to use country-specific results to enable best prediction performance.

Table 2 Best m and $maxnodes$ values of the country-specific RFs

Parameter	DE	ES	FI	FR	GB	IT	PL	PT	SE
m	2	7	2	4	2	2	2	4	2
$maxnodes$	Unlim	Unlim	Unlim	19	Unlim	Unlim	16	6	Unlim

Unlim stands for a RF where the grid search determined not to restrict the number of maximum nodes

useful predictor for a given split of a given tree with $m = 3$ and 16 out of 30 variables with almost no use for HGF prediction is:

$$\frac{\binom{15}{0}\binom{15}{3}}{\binom{30}{3}} \approx 11.2\%$$

This means that without preliminary variable selection, approximately every eighth split will be counterproductive.

After having extracted 15 promising variables and applying SMOTE to the country-specific data sets to establish an even distribution of HGFs and LGFs, the final country-specific RFs can be estimated. As explained in Section 4.2, the RF's two main tuning parameters are the number of trees (B) and the m-value. In Behr and Weinblat (2017), we have also found the maximum number of nodes per tree ($maxnodes$) to be a valuable tuning parameter to limit overfitting. Based on observations of the OOB error, which stabilizes for approximately 100 trees, the number of trees is set to 100 according to Breiman and Cutler (2004) to limit the computational overhead. The m and $maxnodes$-value for each country are determined in a grid-search by performing three three-fold stratified cross-validations (CVs).[18] Inside of each CV, the AUC measure is calculated and averaged so that the parameter combination with the highest average AUC values is chosen. The optimal parameters according to CV are presented in Table 2.

It turned out that for six of the nine countries, it is not necessary to vary $maxnodes$. These parameter combinations were used to estimate the final RF based on the country-specific training data. In the main analysis, these final RFs are used to predict the HGF status of the test data, to estimate the prototypical HGF and LGF as well as to estimate the variable importance ranking of the remaining 15 variables. Please note that SMOTE is not applied to the test data to simulate a realistic setting in which all reputedly highly growing companies contain a huge number of false positives, which will also be the case in practical applications.

Becchetti and Trovato (2002) and Coad et al. (2014a) believe that a risky corporate strategy might succeed and lead to high growth or fail and end up in bankruptcy. To evaluate this belief, we estimated nine further RFs (one RF for each country) which predict bankruptcy instead of high growth. These RF's predictions are compared with the prediction of the

[18] In k-fold CV, the data is randomly separated into k folds of almost equal size. In each of the k iterations, $k - 1$ folds are combined again and used to estimate the RF. The observations of the remaining fold are then classified by the RF to estimate the RF's prediction performance. This model generation and testing is repeated k times so that each of the folds will be used for testing once. The mean of the error estimates of the k iterations is the overall result of the CV (Ablameyko 2003, p. 67, Alpaydin 2004, p. 331). A stratified CV ensures that the relative frequencies of the two classes (LGF/HGF) are identical in each fold. To further improve reliability of a CV, the whole CV is repeated several times and average over all the results of the single CVs (Witten et al. 2011, p. 152–154). to limit the computational overhead, we decided to perform three three-fold CVs instead of a ten-fold CV, which is usually used in literature.

previous nine RFs to evaluate whether high growth and bankruptcy are "two sides of one coin".

5 Predicting Variables and Descriptive Evidence

In this section, we discuss the potential predicting variables and provide some descriptive statistics for the train data set.

5.1 Predictive Variables

Pytlik (1995, p. 233) states the following conditions a key figure has to meet to be regarded as a promising predictor: a key figure turned out to be useful in previous studies, theoretical considerations suggest that it is useful or it has a high importance in practical applications.

Their values (except of firm age and sector) are both regarded one year before the first year of the growth or bankruptcy period X_t, in first differences covering both one year $\Delta_1 X = X_t - X_{t-1}$ and two years $\Delta_2 X = X_t - X_{t-2}$. The only exception is the employment-based variables. We analyze the employment (emp) in year X_t. Instead of regarding its first differences, we have calculated two Birch - Schreyer growth indicators $\Delta_1 growth$ and $\Delta_2 growth$ from Section 3.2 covering a time lag of one and two years respectively. As Coad (2007b, p. 16) pointed out, there is no agreement in the literature concerning the time lag so that we decided to consider two different ones simultaneously and to delegate the choice to the RFs. Besides that, we also consider the three non-accounting related key figures age, sector and legal form.

As explained in Section 4.4, country-specific preliminary RFs have been used to estimate a variable importance ranking out of 30 considered variables. Both the initial 30 variables and the chosen ones can be seen in Table 3.

Only the chosen variables are discussed in this section.

- Age of the firm: age
  The firm's age is regarded in this study because several studies like Harhoff et al. (1998) showed that young firms tend to grow faster than older ones. The growth process of younger firms is often characterized to have a high variance, e.g. by Coad (2007b), Dunne et al. (1989) and Boeri and Cramer (1992). Furthermore, the HGF literature found that a small group of young firms is responsible for a disproportionately high number of new jobs (Acs et al. 2008, p. 11 and p. 19; Schreyer 2000, p. 6). Henrekson and Johansson (2010) did not confirm that idea. One possible explanation is that young firms have to overcome an intensive selection process by quickly growing to be capable of competing (Boeri and Cramer 1992, p. 555).
- Size of the firm: $size$
  age and $size$ are considered to be important in almost all HGF related studies. Since bigger firms are usually older than smaller ones, there is possible a correlation between these two predictors (Coad 2007b, p. 18 and p. 51). Furthermore, small firms are under constant pressure to grow to reduce costs to the same level as other (bigger) competitors (Coad 2007b, p. 51). A certain minimal size is important in fields with high fixed costs whereas a relatively small size increases the firm's flexibility (Schreyer 2000, p. 13). Consequently, large firms have longer planning horizons and more long-term investments, which pay-off over several years and lead to a higher autocorrelation of growth (Coad 2007a, p. 74). Besides that, national legislation is often firm's size-specific. For

Table 3 All analyzed predictors

Variable	t	Δ_1	Δ_2
$emp = Number\ of\ employees$	✓	✗	✗
$growth = Birch\text{-}Schreyer\ growth\ indicator$	✗	✓	✓
$dr = \frac{total\ debt}{total\ assets}$	✓	(✓)	(✓)
$roa = \frac{net\ profit + interests\ on\ borrowed\ capital}{total\ assets}$	✓	(✓)	(✓)
$ros = \frac{net\ profit}{sales}$	(✓)	(✓)	(✓)
$far = \frac{fixed\ assets}{total\ assets}$	✓	(✓)	(✓)
$liq = \frac{short-term\ assets}{short\text{-}term\ debts}$	(✓)	(✓)	(✓)
$efar = \frac{equity}{fixed\ assets}$	✓	(✓)	(✓)
$spe = \frac{sales}{number\ of\ employees}$	✓	✓	✓
$size = total\ assets$	✓	✓	✓
$lf = legal\ form$	(✓)	✗	✗
$sec = sector$	✓	✗	✗
$age = age\ of\ the\ firm$	✓	✗	✗

t is the absolute value one year before HGF status is determined. Δ_1 is the first difference to the previous year. Δ_2 is the first difference to the next to last year. ✓ means that the corresponding variable was analyzed by the final country-specific models. (✓) indicates that the predictor is not used because it was not considered to be useful by the preliminary RF. Predictors with an ✗ have not been considered at all

instance, bigger firms in certain countries have higher firing costs and have to pay higher taxes. On the other hand, they also have a higher lobbying power, which might facilitate growth (Coad 2007b, p. 27 and p. 78–79; Schreyer 2000, p. 7). There seems to be no consensus in the literature whether a small or a big size facilitates high growth. For instance, Harhoff et al. (1998) find that a bigger size leads to lower growth rates whereas Henrekson and Johansson (2010, p. 1) disagree on that. We also consider relative size changes using $\Delta_1 lsize$ and $\Delta_2 lsize$.

– Number of employees: emp
  emp is a different indicator of the firm's size (Coad 2007b, p. 3). Lopez-Garcia and Puente (2012, p. 1038) regard emp as a proxy for human capital which is an important determinant of HGFs.
– Birch-Schreyer growth indicator: $growth$
  Following Lopez-Garcia and Puente (2012, p. 1031), we considered past growth to allow for autocorrelation. Coad (2007a, p. 78) also reported a positive autocorrelation for big firms and a negative one for small ones. Coad (2007b, p. 15–16) points out that findings on autocorrelation are highly contradictory so that neither the existence of autocorrelation nor its influence and the length of the time lags are certain.
– Sector of the firm: sec
  We define six different sectors based on the NACE taxonomy ensuring that every sector contains a sufficient number of firms in every country. The regarded six sectors including their NACE code intervals are presented in Table 4. Although it is known that HGFs can be found in all industries, there are still reasons to consider this variable as a possible predictor (Schreyer 2000, p. 3; Acs et al. 2008, p. 2). For instance, Lopez-Garcia and Puente (2012, p. 1037) find a disproportionately high number of HGFs in some sectors. Other studies as Boeri and Cramer (1992, p. 546) only find a small sectoral effect.

Table 4 Sector classification and the corresponding NACE regions

Range of first two NACE digits	Sector designation	Acronym
01-33	Manufacturing	s_ma
35-43	Energy	s_en
45-56	Trade	s_tr
58-68	Finance	s_fi
69-82	Services	s_se
84-99	Social	s_so

Another reason to include sec is that growth within a sector also influences the growth patterns of individual firms (Levratto et al. 2010, Audretsch and Mahmood 1994).

– Debt ratio: dr
This is a measure of a firm's leverage indicating the borrowed share of a firm's funding. The remaining funding originates from past retained profits or has been introduced by the shareholders (Albrecht et al. 2007, 476; Penner 2004, 218). One important motivation to include dr is put forward by Lopez-Garcia and Puente (2012, p. 1036–1039). They argue that firm growth requires financing and that a high dr might lead to future financing constraints and the inability to realize all reputedly profitable projects. In their study, dr was found to have a significant non-linear influence when not controlling for firm-specific time-invariant factors. The importance of dr for firm growth was also confirmed by Fagiolo and Luzzi (2006, p.33), Becchetti and Trovato (2002, p. 294) and Levratto et al. (2010, p. 10). Lopez-Garcia and Puente (2012, p. 1031) found no effect of dr. One possible reason for doubt is that especially in case of start-ups, not bank credits but risk capital and internal finance are important funding sources (Lopez-Garcia and Puente 2012, p. 1038).
– Return on assets: roa
roa measures how efficient a firm uses its assets and is an indicator of profitability (Stickney et al. 2009, 245). It is a widely used financial ratio and might be a predictor of future growth because investors consider firms with high returns as secure investments (Chen et al. 1985, p. 202; Levratto et al. 2010, p. 8). However, according to Coad (2007b, p. 24–25) roa can be expected to have a significant but rather small explanatory power.
– Sales per employee ratio: spe
According to Vause (2009, p. 176) and Puri (2012, p. 131) both spe and Δspe quantify the (average) firm's employee's ability to generate sales and therefore their productivity. High spe values might incentivize certain firms to hire additional staff to increase sales. In opposition to this reasoning, Coad and Broekel (2012) determined a negative correlation of spe and growth. The contribution of roa and spe is questioned by Coad (2007b, p. 25) and Baily et al. (1996, p. 259) because both figures can be increased by growth but also by downsizing so that its explanatory power might be inconclusive. For instance, Levratto et al. (2010, p. 18) did not find a significant influence on growth rates for HGFs.
– Fixed assets ratio: far
far indicates the degree of capital commitment. High values imply low flexibility and constant pressure to keep capacity utilization high to cover the recurring assets'

expenses (e. g. for energy and maintenance). We assume that a high far might motivate a firm to grow to spread its high fixed costs over a greater number of products (Schneider and Lindner 2010, 317). Levratto et al. (2010, p. 5) add that too high fixed costs can be a serious threat for future growth.

– Equity fixed asset ratio: $efar$
  This indicator is also known as the 'golden rule of balance sheet' and is designed to enable a rather medium-term or even long-term perspective (Löbbe 2001; Buenstorf et al. 2013, p 250). The intention is to ensure that a firm's assets should be financed by equity to a sufficient degree. This indicator is known to have a high relevance to practitioners (Becker 2010, 16).

It is surprising that the firm's legal form is not among the most important variables. After all, the legal form determines growth-related issues like taxation, whether a firm can issue shares, and who is liable in case of financial problems (Levratto et al. 2010; Harhoff et al. 1998). Therefore, Harhoff et al. (1998, p. 455) expected for their regarded West German firms that limited liability firms will show riskier entrepreneurial activity with more growth potential. They found 4.5% higher growth rates for limited liability firms than for unlimited ones (Harhoff et al. 1998, p. 481).

5.2 Descriptive Results

After identifying the final set of predicting variables, we present some descriptive statistics disaggregated by the HGF status. The results for the median (Table 13), mean (Table 14), standard deviation (Table 15) and inter-quantile range (Table 16) are presented in the Appendix. Please note that $size$ is presented in Millions and spe in Thousands.

It can be observed that HGFs are usually substantially bigger regarding the number of employees and balance sheet total. The differences are particarly big for Portugal and GB. In all cases for $\Delta_1 growth$ and $\Delta_2 growth$ except for Portugal and Poland, the previous employment growth quantified by the Birch-Schreyer growth indicator is much higher for HGFs. Almost the same holds for $\Delta_1 lsize$ and $\Delta_2 lsize$. The only difference is that $\Delta_2 lsize$ for Portugal is bigger for HGFs than for LGFs while $\Delta_1 lsize$ is not.

Moreover, the median HGF is younger than the corresponding LGF in all countries except GB where there is no difference. The age-related observations are in line with the considerations, which are presented in Section 5.1 while the size related findings are not. Still, most HGFs are not start-ups. Only about 18% of all HGFs are less than six years old and approximately every second HGF is 15 years old or even older.

The variable $efar$ also show a clear tendency, being smaller for HGFs than for LGFs in all cases except the Finnish one. This finding is counter-intuitive. Moreover, in most countries, HGFs have a higher roa in the year 2006 than the LGFs.

As far as dr and far are concerned, no clear patterns can be noticed.

To identify sectors with disproportionately high or low occurrences of HGFs, we analyzed the sector distribution in Table 5. The relative frequencies are calculated using the total number of HGFs or LGFs, respectively. By trend, HGFs often seem to originate from the service and social sector and infrequently from the manufacturing and trade sector.

In the HGF literature, studies like Harhoff et al. (1998) assume that being a HGF and going bankrupt are two sides of a coin. The underlying idea is that certain firms make risky decisions. If everything goes according to plan, the firms might eventually become a HGF. In the opposite case, the firm might even fail. According to Harhoff et al. (1998, p. 455), this risky behavior is especially predominant among firms with a limited liability legal form. If

Table 5 Country-specific sector distributions

sec	HGF	DE	ES	FI	FR	GB	IT	PL	PT	SE	All
s_en	FALSE	0.1723	0.1821	0.1529	0.1307	0.0627	0.1097	0.1608	0.1942	0.1207	0.1460
	TRUE	0.1099	0.1590	0.1257	0.1385	0.0635	0.1100	0.1288	0.2857	0.1574	0.1401
s_fi	FALSE	0.1189	0.0849	0.0840	0.1010	0.1078	0.0591	0.0540	0.0540	0.1249	0.0859
	TRUE	0.0604	0.0696	0.0942	0.1269	0.1091	0.0719	0.0539	0.0714	0.1267	0.0851
s_ma	FALSE	0.2952	0.2613	0.3565	0.2193	0.3206	0.4374	0.4003	0.3273	0.2239	0.3030
	TRUE	0.2912	0.2379	0.3613	0.1989	0.2741	0.3656	0.4379	0.2500	0.1713	0.2655
s_se	FALSE	0.1270	0.0662	0.0828	0.0995	0.1595	0.0622	0.0310	0.0827	0.1136	0.0781
	TRUE	0.1813	0.1163	0.1099	0.1695	0.2386	0.1090	0.0328	0.0714	0.1685	0.1292
s_so	FALSE	0.0824	0.0310	0.0356	0.0333	0.0410	0.0266	0.0352	0.0288	0.0323	0.0316
	TRUE	0.2308	0.0779	0.0733	0.0488	0.0609	0.0683	0.0632	0.1071	0.0720	0.0730
s_tr	FALSE	0.2042	0.3746	0.2882	0.4162	0.3083	0.3049	0.3186	0.3129	0.3846	0.3556
	TRUE	0.1264	0.3392	0.2356	0.3173	0.2538	0.2751	0.2834	0.2143	0.3041	0.3072
All	FALSE	1724	50283	1714	12148	3730	28712	3867	278	16149	118605
	TRUE	182	5313	191	1292	394	3017	427	28	1792	12636
	All	1906	55596	1905	13440	4124	31729	4294	306	17941	131241

this is the case, this has serious implications for industrial policy because support for HGFs would also foster risky economic activity, which is presumably not intended. This difficulty is denoted as "survivorship bias" (Becchetti and Trovato 2002, p. 291; Coad et al. 2014a, p. 96).

Table 6 is only supposed to present the train and test data. The findings do not find their way into the RFs so that the performed prediction remains to be a true out-of-sample prediction.

We observed that only in France and in Spain, former HGFs are likely to fail in near future. This is probably also caused by the fact that there is a considerable number of bankrupt firms in the corresponding test data sets. Even for these two countries, it is not appropriate to assume a survivorship bias because LGFs from the same countries are more likely to fail than HGFs. Considering our high numbers of bankrupt firms, this study's results can be seen as in line with the findings of Acs and Mueller (2008) that only three percent of former HGFs fail in the consecutive period

Another question which is, for instance, put forward by Daunfeldt and Halvarsson (2015) is the likelihood that a former HGF remains to be a HGF in the next period. Depending

Table 6 Transitions between HGFs, LGFs, and bankrupt firms between train- and test-data

Transition	DE	ES	FI	FR	GB	IT	PL	PT	SE	All
HGF → HGF	0.2903	0.2442	0.1626	0.2136	0.4054	0.3400	0.1562	0.1538	0.1443	0.2345
HGF → LGF	0.6882	0.6537	0.7724	0.5635	0.5856	0.5990	0.8438	0.7692	0.8391	0.7016
HGF → bankr	0.0215	0.1020	0.0650	0.2229	0.0090	0.0610	0.0000	0.0769	0.0166	0.0639
LGF → HGF	0.2270	0.0689	0.1297	0.0452	0.0775	0.0907	0.0733	0.1667	0.0721	0.1057
LGF → LGF	0.7517	0.8057	0.7978	0.6981	0.8961	0.8403	0.9267	0.6282	0.9063	0.8057
LGF → bankr	0.0214	0.1254	0.0724	0.2567	0.0264	0.0690	0.0000	0.2051	0.0216	0.0887

Table 7 Five country-specific AUC quantiles by employee number

Quantile	DE	ES	FI	FR	GB	IT	PL	PT	SE	AVERAGE
$q_0 - q_{20}$	0.6666	0.6521	0.6975	0.7075	0.6532	0.6806	0.5602	0.6245	0.5116	0.6393
$q_{20} - q_{40}$	0.7534	0.6671	0.7780	0.7496	0.7205	0.7125	0.6009	0.7469	0.6024	0.7035
$q_{40} - q_{60}$	0.7240	0.6384	0.6964	0.7446	0.6624	0.6564	0.8144	0.5843	0.6119	0.6814
$q_{60} - q_{80}$	0.6093	0.6342	0.5452	0.6477	0.6119	0.6754	0.5528	0.5965	0.5828	0.6062
$q_{80} - q_{100}$	0.5749	0.6757	0.6151	0.6576	0.6521	0.6967	0.6799	0.5732	0.5765	0.6335
AVERAGE	0.6657	0.6535	0.6665	0.7014	0.6600	0.6843	0.6416	0.6251	0.5770	0.6528

on the answer, it might not be reasonable in the long-run to invest in HGFs by politicians or private investors. In this study, approximately every fourth HGF manages to repeat its strong growth. However, the results are highly country-specific. Acs et al. (2008) find that approximately every fourth big HGF was a HGF in the previous period, too, and Hölzl (2014) also determines for Austrian firms that Birch-Schreyer based high growth is rather persistent.

However, we can neither confirm the results of Lopez-Garcia and Puente (2012) that the majority of all HGFs also were HGFs in the previous period. Nor can we confirm the conclusion of Coad (2007a, p. 80) for his regarded French manufacturing HGFs that they "[...] may grow a lot in one period, but it is unlikely that the spurt will last long". Daunfeldt and Halvarsson (2015, p. 361) support (Coad 2007a) by stating that (Swedish) HGFs are usually "one hit wonders" and are unlikely to grow intensively in two consecutive periods. Only about one percent of their firms remained to be HGFs during two consecutive periods. However, former HGFs are most likely to experience a modest employment growth.

6 Model Estimation and Prediction Results

After having explained our proceeding and the used training and test data, this section presents the estimation results. It starts with the analysis of the performance of the out-of-sample prediction and an inspection whether HGF prediction of bigger firms is more reliable than of small ones as it is frequently assumed in the literature. Furthermore, it is evaluated if a higher propensity to grow is associated with a higher propensity to default. This is another frequently assumed hypothesis. The last part of this section focusses on the RF's byproducts namely prototypes and the variable importance analysis.

6.1 Prediction Results Evaluation

This part presents the AUC and other measures to illustrate the performance of the RFs. It also discusses whether HGFs differ from firms, which grow but do not belong to the highest growing 10% as assumed by Lopez-Garcia and Puente (2012) and Schreyer (2000) (Table 7).

Table 8 presents the obtained results of the country-specific HGF prediction. The true positive rate (TPR), false positive rate (FPR), the Accuracy and Precision are all calculated by predicting a firm as highly growing if the majority of all trees of the forest believe the firm to be highly growing.[19] Using a cutoff value of 50% is reasonable because SMOTE was

[19]The correlation column will be discussed at the end of this section.

Table 8 Out-of-sample prediction results using 0.5 as cutoff for all measures except AUC

Country	AUC	TPR	FPR	Accuracy	Precision	Correlation
DE	0.6737	0.3518	0.1738	0.7814	0.1742	0.0404
ES	0.7157	0.2986	0.0813	0.8643	0.2612	−0.0046
FI	0.6439	0.2645	0.1442	0.8026	0.1536	0.0373
FR	0.7720	0.6528	0.2554	0.7380	0.1653	0.0092
GB	0.8110	0.5348	0.1328	0.8354	0.2991	0.0902
IT	0.7624	0.3839	0.0952	0.8566	0.2913	0.0131
PL	0.7067	0.5952	0.2513	0.7332	0.2101	0.0828
PT	0.6685	0.1582	0.0383	0.9016	0.2500	0.0450
SE	0.5232	0.3942	0.3446	0.6302	0.1090	0.0120

applied in a way that established an even class distribution. Still, higher values can be used to only predict firms to be a HGF when the RF is "very sure" or lower values in order not to oversee as many true HGFs as possible. In both cases, the four mentioned measures will change while the AUC is independent of such cutoff points. Moreover, the later discussed Table 7 illustrates the models' performance for different thresholds. Please be reminded that the accuracy and the precision measures are also influenced by the share of HGFs in the test data as it is mentioned in Section 4.3. A better recognition of LGFs has a higher effect on these two measures than of HGFs, which is against the intention of the study.

Based on the AUC, the following ranking can be obtained for true out-of-sample predictions: GB (highest prediction performance), FR, IT, ES, PL, DE, PT, FI, SE (lowest prediction performance). To evaluate if country-specific differences are caused by different predictabilities of HGFs in this countries and not by chance, we compared the ROCs using Venkatraman's unpaired test.

The test results are presented in Table 9. According to these results, there is strong evidence for the superior performance of the British RF and the inferiority of the Swedish RF applying the usual 5% level of significance. The Polish model only significantly differs from the British and Swedish one. Moreover, it cannot be said with certainty whether the results of two consecutive country-specific models always deviate significantly. Examples for such countries are France and Italy but also Finland and Poland. However, the test seems to mainly confirm the mentioned ranking.

Table 9 p-values of the ROC based unpaired Venkatraman test

Country	DE	ES	FI	FR	GB	IT	PL	PT	SE
DE	–	0.0040	0.4800	0.0000	0.0000	0.0000	0.5680	0.1240	0.0000
ES	0.0040	–	0.0040	0.0000	0.0000	0.0000	0.5760	0.0000	0.0000
FI	0.4800	0.0040	–	0.0000	0.0000	0.0000	0.4680	0.0760	0.0000
FR	0.0000	0.0000	0.0000	–	0.0120	0.8240	0.1200	0.0000	0.0000
GB	0.0000	0.0000	0.0000	0.0120	–	0.0000	0.0120	0.0000	0.0000
IT	0.0000	0.0000	0.0000	0.8240	0.0000	–	0.4280	0.0000	0.0000
PL	0.5680	0.5760	0.4680	0.1200	0.0120	0.4280	–	0.2120	0.0000
PT	0.1240	0.0000	0.0760	0.0000	0.0000	0.0000	0.2120	–	0.0000
SE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	–

According to the AUC's interpretation, the British RF has a probability of 81.1% to assign a randomly chosen HGF a higher HGF-propensity than a randomly chosen LGF (cf. Table 8).

$$P(P(\widehat{HGF}|HGF) > P(\widehat{HGF}|LGF)) = 0.811$$

For a Swedish HGF, this probability is only 52.32%. In France, the highest share of HGFs (65.28%) is correctly anticipated (TPR) whereas the same anticipation works worst in Portugal (15.82%). Portuguese decision maker might want to use a lower threshold at the expense of a higher number of false positives. For the current threshold of 50%, the Portuguese, Spanish and Finnish RFs act conservatively referred to Fawcett (2006, p. 863) since they are rather hesitant to predict high growth (low TPR) but make only a few incorrect high growth predictions (low FPR). The Polish and French models can be designated as liberal since they seem to predict high growth more willingly so that more true HGFs are anticipated (high TPR) but also more wrong high growth predictions occur (high FPR).

For the 50% threshold, almost every third British and Italian firm, which is assumed to be a HGF, is truly a HGF which can be seen from the precision measure. This is a considerable result because by simple random sampling, only about every tenth firm would have been a HGF. However, especially in Sweden but also in Finland and France, the RF's results are only slightly better than random choice. The Portuguese RF has the highest share of correct predictions (accuracy). Because of the low TPR and FPR value, this is achieved by seldom predicting high growth so that most of the LGFs (90% of all firms) are correctly predicted.

Table 10 shows the RF's prediction performance from another perspective and illustrates it for six different cutoff points. In this section, only the HGF-rows are analyzed. The

Table 10 Relative frequencies in the test data of all predicted HGFs and bankrupt firms per country among x% of all firms with the highest propensity to be a HGF

Country	Variable	1%	5%	10%	15%	25%	50%
DE	HGF	0.0278	0.1238	0.1889	0.2899	0.4365	0.7296
	BANKR	0.0000	0.0051	0.0202	0.0455	0.1061	0.2879
ES	HGF	0.0456	0.1809	0.2986	0.3919	0.5405	0.7674
	BANKR	0.0071	0.0374	0.0809	0.1225	0.2260	0.5391
FI	HGF	0.0065	0.0968	0.1684	0.2645	0.3806	0.6935
	BANKR	0.0167	0.0556	0.0833	0.1111	0.2000	0.4278
FR	HGF	0.0708	0.2323	0.3769	0.4629	0.6113	0.8309
	BANKR	0.0030	0.0234	0.0528	0.0912	0.1758	0.4060
GB	HGF	0.0542	0.2294	0.3900	0.4875	0.6657	0.8830
	BANKR	0.0062	0.0248	0.0559	0.0621	0.1553	0.3913
IT	HGF	0.0561	0.2143	0.3409	0.4393	0.5866	0.8266
	BANKR	0.0014	0.0236	0.0636	0.1164	0.2154	0.4875
PL	HGF	0.0476	0.0655	0.2381	0.3333	0.5000	0.7619
	BANKR	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
PT	HGF	0.0311	0.1638	0.2825	0.3814	0.4774	0.7034
	BANKR	0.0050	0.0350	0.0922	0.1451	0.2602	0.5455
SE	HGF	0.0308	0.0917	0.1534	0.2014	0.3025	0.5193
	BANKR	0.0000	0.0049	0.0584	0.0779	0.1509	0.3431

remaining rows are covered in Section 6.3. The HGF-rows show for each country, what share of all HGFs in the corresponding test data set is correctly predicted when regarding the top x% of all firms with the highest propensity to grow substantially. When regarding Italian, French, and British models, the top one percent group contains around five times more HGFs than random sampling would identify. When regarding the five percent group, similar statements can be made. The density of correctly anticipated HGFs is much bigger between 0 and 25% than between 25% and 50% for all countries except for Finland and Sweden so that a big number of true HGFs receive rather high propensities. This is a desirable outcome.

The boxplots in Fig. 2 confirm this finding. They show the distribution of the estimated propensities to be a HGF subject to the growth category. Once again, the results of the bankrupt firms are discussed in Section 6.3. On average, the Italian, French, and British HGFs have a much higher propensity to be a HGF than the LGF. Their medians are always over the 75% quantile of the LGFs. In contrast to that, no difference can be seen for the

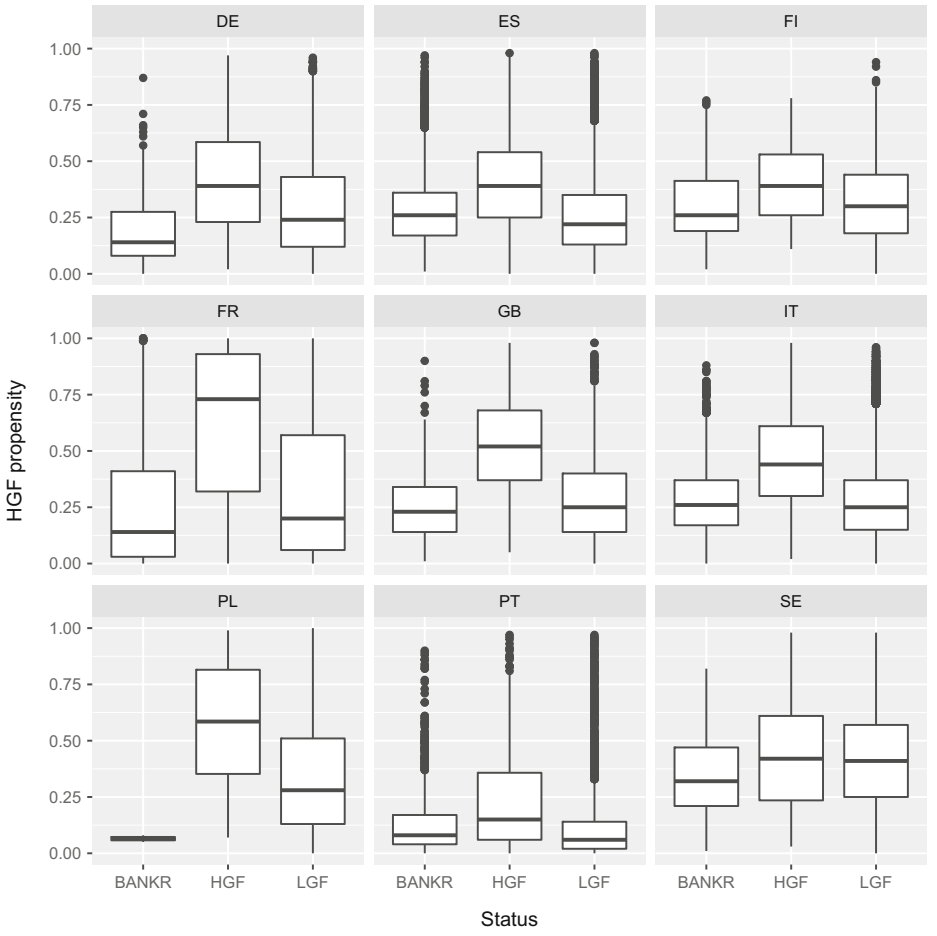

Fig. 2 Country-specific boxplots of the estimated out-of-sample propensity to be a HGF subject to the firm's true status

mentioned two groups in Sweden. Countries like Germany and Portugal lie between these two extremes. By trend, HGFs have a higher propensity to be HGFs than the LGFs counterparts do. However, every possible cutoff point (parallel line to the x-axis) will always lead to a remarkable number of false positives which receive high propensity scores but are LGFs.

Another claim in the literature, for instance by Schreyer (2000), is that HGFs differ from firms which grow but do not belong to the highest growing 10%. To contribute to this discussion, the correlation measure by Bravais and Pearson is presented in Table 8. It is calculated by confronting the firms' Birch-Schreyer growth indicator value with its estimated propensity to be a HGF: $cor(growth, \hat{p})$. The basic idea of this analysis is that if there is no structural difference between HGFs and the remaining firms, the indicator value should correlate with the estimated propensity: the higher the propensity, the higher the growth. Although all except the Spanish correlation are indeed positive, they are quite small which hints to the frequently mentioned structural difference of HGFs. These findings support the results from Lopez-Garcia and Puente (2012) who have also stressed the peculiarities of HGFs.

Another assumption, which is vividly discussed in the literature, is whether high growth of big firms is more predictable than of small firms. For instance, Coad (2007a) and Acs et al. (2008) concluded that big firms have a rather steady growth path due to longer planning horizons and investment plans. Coad (2007b, p. 51) and Coad and Hölzl (2009) described the growth of small firms as erratic and, therefore, harder to predict.

To contribute to this discussion and to evaluate whether big firms can be more reliably predicted by the RF algorithm, the firms from the test data were evenly divided in five groups based on their number of employees. Table 7 shows the corresponding AUCs for each of these groups. If small firms' growth is truly more irregular than big firms' growth, small firm's growth should be harder to predict leading to smaller AUC values. It can be confirmed that the country-specific AUCs of the smallest firms ($q_0 - q_{20}$) are most often slightly smaller or substantially smaller (Poland) than the average value. However, for countries like Spain, France, Great Britain, Italy, and Portugal, no remarkable differences can be observed. Concerning the biggest firms ($q_{80} - q_{100}$), no cross-national statements can be made either. While prediction seems to work better-than-average in Spain, Italy, Poland, and Sweden, the same prediction delivers below average results for the remaining countries. Based on these results, above average results can be found for the second smallest ($q_{20} - q_{40}$) size category of firms for all countries except Poland, which does not go in line with the usual findings.

To find out if these differences could have also been caused by chance, we performed the unpaired Venkatraman test, which is mentioned in Section 4.3. We tested the difference between the second smallest size category and the biggest size category which is assumed to be most predictable in literature. We found the deviation to be "significant" for Germany, Finland, and Portugal using the usual 5% level of significance and also for Sweden using 10%. This is why we cannot find any clear tendency as far as predictability of different size groups is concerned.

6.2 Variable Importance and Prototypes

As mentioned before, the RF does not only enable predictions but also several interpretation mechanisms. This section starts with a visual analysis of a "typical" tree and continues with a more illustrative variable importance ranking. It concludes with an observation of typical HGFs, LGFs, and bankrupt firms.

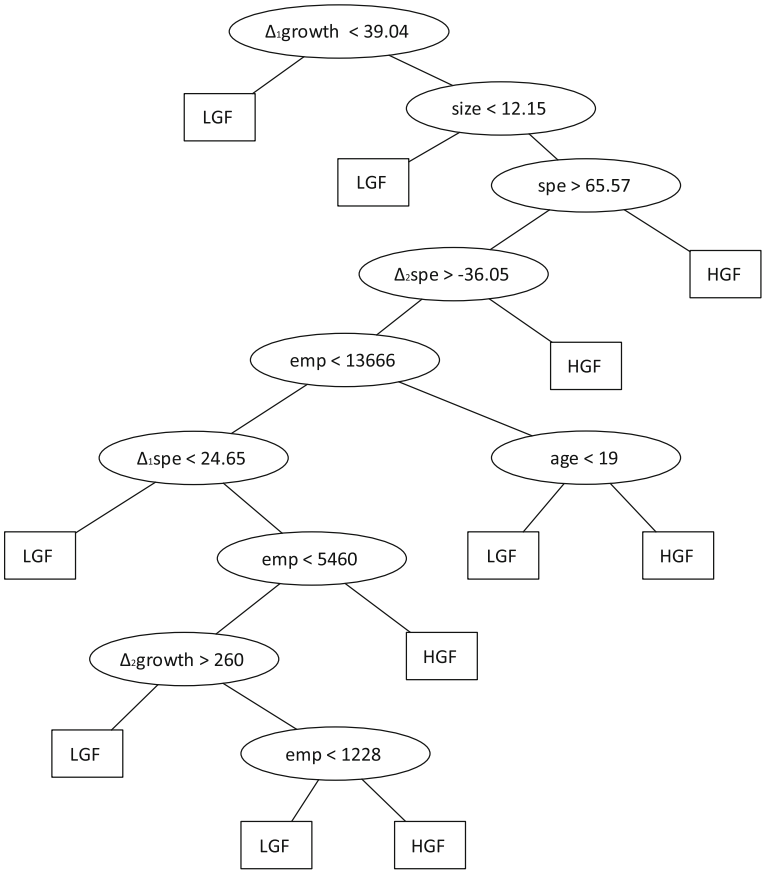

Fig. 3 British example decision tree created by CART

Figure 3 exemplifies one "typical" decision tree of the British RF.[20] The presented tree is the outcome of the CART algorithm and is similar to the ones, which are used inside the RF. This tree is more suited for demonstration purposes because CART chooses the best split out of all possible ones based on its Gini measure and not out of m randomly chosen variables, all of which might be inappropriate. The tree illustrates the 1 to 9 univariate evaluations which need to be conducted to generate a prediction for a given firm. CART also carries out an implicit dimension reduction since only nine out of the fifteen available variables are considered for a prediction. For firms with $\Delta_1 growth < 39.04$, even only one evaluation is sufficient. As a provisional result, it can be stated that the prediction relies on the absolute and delta values of the growth indicator, the company size, and age, the number of employees, and the sales per employee. This tree already confirms the findings

[20] We have chosen Great Britain because of the high prediction performance of the corresponding RF and because of its relatively small size so that it is more clearly arranged than the other countries' trees.

of Becchetti and Trovato (2002, p. 291) that there are several determinants of firm growth in addition to the firm's size and age. Financial variables like roa and dr and the sector designation are also not considered which is not in line with Becchetti and Trovato (2002) who stress the importance of creditworthiness for future growth.

Another approach to understanding the functionality of the RF is the analysis of the RF's variable importance, which also enables a cross-national comparison of HGF patterns. The x-axis of Fig. 4 contains the analyzed countries and the y-axis the regarded predictors. These predictors are ordered based on the average RF's variable importance over all nine countries so that the variable at the top has the highest importance and the one at the bottom the lowest. The numbers in the diagram indicate the importance rank for the corresponding country and variable. In this context, 1 represents the highest rank and 15 the lowest. Moreover, the higher a variable's importance, the bigger the radius of the circle.

The findings confirm and complement the existing literature, which is mainly driven by regression analysis. The size, both measured using the number of employees and the balance sheet total, is highly important to predict future high growth. This is in line with Levratto et al. (2010, p. 6) who state that size is a determinant of growth when firms of different sizes are analyzed simultaneously. According to this authors, small firms grow at a higher rate than bigger firms. The number of employees is the almost certain most important variable. Other very useful variables are the two Birch-Schreyer indicator values $\Delta_1 growth$

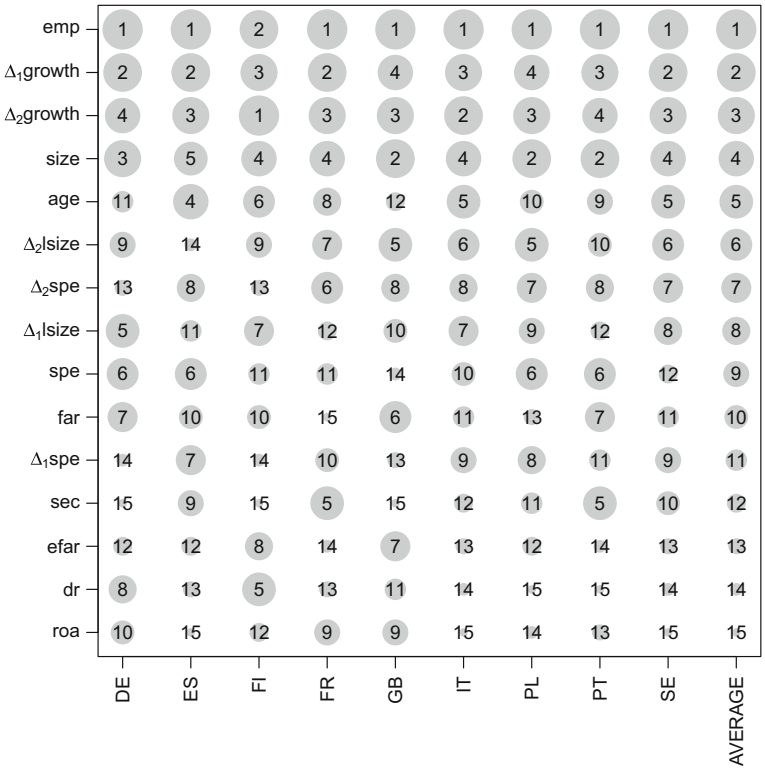

Fig. 4 Variable importance rankings of the nine regarded country-specific RFs

and $\Delta_2 growth$. Moreover, the absolute change of the balance sheet total over one or two years is also valuable to anticipate future HGFs in most countries. This means that both the firm's size and its variation over time play a substantial role. These patterns seem to be valid for all regarded countries. Acs et al. (2008, p. 46) determined that past HGF status is an important predictor of future growth. Coad (2007a, p. 80) could also show that even less recent growth changes have an influence on current growth.

However, different studies came up with contradicting results as Coad (2007b, p. 15–16) summarized in his review. This disagreement can also be observed for the number of relevant lags.

Besides that, the firm's age also has a noteworthy prediction contribution in most but not in all countries. As already determined by the CART tree, age is quite unimportant in Great Britain but also in Germany, Poland, and Portugal. In contrast to that, age is the fourth most important predictor in Spain. In the literature, it is assumed that age is very important for HGF prediction and that its contribution ranges slightly behind the firm's size (Harhoff et al. 1998, p. 479; Acs et al. 2008, p. 10). The explained country-specific differences have not been mentioned in the literature yet.

The ranking also confirms the regression-based finding from previous studies that size- and age-based variables are more important than financial data based predictors (Coad 2007b, p. 23). However, there are exceptions to this heuristic. For instance, far is rather important in Portugal and Great Britain. spe and its variation are worth considering in Germany, Spain, Poland, and Portugal. On average, the first difference of the sales per employee ratio is the most important financial variable. Variables like dr and roa but also the firm's sector are only of minor (average) importance. For some countries, these variables are quite important like roa for France and Great Britain and sec for France and Portugal. The low importance of dr is in line with the findings of Lopez-Garcia and Puente (2012) for Spanish firms, with Levratto et al. (2010) findings for French firms and with Fagiolo and Luzzi (2006) results for Italian firms. Lopez-Garcia and Puente (2012, p. 1038) explain the low importance of dr for young HGFs by their reliance on risk capital instead of bank credits. However, contradictory findings like Becchetti (1995) for Great Britain exist as well.

Similar to this study's results, Boeri and Cramer (1992) did not find an important sector effect for the regarded German firms and Coad (2007b, p. 29) reported that most other regression studies made similar findings. A possible reason might be that HGFs can be found in all sectors although they are known to occur more frequently in certain ones (Schreyer 2000, p. 5; Acs and Mueller 2008, p. 86; Acs et al. 2008, p. 2). Meanwhile, Becchetti and Trovato (2002) found a sector effect in both regarded models. Note that the mentioned variables have all been assessed to be relatively important by an initial variable importance check (c.f. Section 4.4) so that even being on the last position does not mean that the variable is absolutely unimportant.

Besides being an illustration of the different RF, the ranking might provide first hints for the bad performance of the Swedish, Finnish, Portuguese, and German models. This does not seem to be the case in this study because these countries' variable importance rankings do not seem to substantially differ from the other countries. This is why other explanations must be taken into consideration like the presence of other variables, which are omitted in this analysis. An alternative explanation might be that HGF prediction is simply more challenging for these countries.

RF's variable importance ranking illustrates the intensity of the single variables' contribution to the prediction of a HGF. Which values a "typical" HGF or LGF has in a certain country remains unclear. Table 11 presents the values of the six most important variables of

Table 11 Prototypes - six most important variables

Country	Variable	emp	$\Delta_1 growth$	$\Delta_2 growth$	$size$	age	$\Delta_2 lsize$
DE	LGF	212	1.0909	1.0026	10.8883	22	0.1068
	HGF	924	23.8464	34.2333	11.8686	23	0.1312
	BANKR	227	0.0000	1.0175	10.1091	16	0.0886
ES	LGF	17	0.0000	1.0556	7.8099	14	0.1967
	HGF	210	26.2948	56.0000	9.2902	16	0.2907
	BANKR	24	1.0244	2.1333	8.0193	11	0.3780
FI	LGF	20	0.0000	0.0000	7.8698	17	0.1135
	HGF	172	9.6532	15.8824	9.9565	17	0.1637
FR	LGF	18	0.0000	0.0000	7.5939	15	0.1227
	HGF	104	3.1765	7.7778	9.3796	16	0.1862
	BANKR	17	0.0000	0.0000	7.3330	15	0.2217
GB	LGF	126	1.0417	2.0519	10.1365	18	0.1311
	HGF	1528	121.6351	181.3740	13.0414	26	0.3166
	BANKR	80	0.0000	1.0213	9.0107	14	0.1624
IT	LGF	25	0.0000	0.0000	8.7942	20	0.1260
	HGF	251	14.8370	30.7909	10.2083	17	0.2233
	BANKR	23	0.0000	0.0000	8.5702	13	0.1980
PL	LGF	100	0.0000	0.0000	8.1887	14	0.1275
	HGF	250	0.0000	8.1905	9.8240	13	0.4242
	BANKR	60	0.0000	0.0000	7.5027	14	0.0776
PT	LGF	34	−8.5714	−9.4819	9.4661	15	0.1890
	HGF	707	−6.9663	−7.9599	12.1114	23	0.0576
	BANKR	79	−8.5235	−4.6711	8.7823	20	0.0175
SE	LGF	12	0.0000	1.0385	7.2689	17	0.1713
	HGF	90	13.5745	28.8000	8.8899	14	0.4784

a country-specific prototype of a HGF and LGF.[21] The values of all 15 available variables can be found in Table 17 in the Appendix.

In all countries, such a typical HGF has a higher number of employees and a larger size than a corresponding LGF like it was already discovered in Section 5.2. Schreyer (2000, p. 13) explains this by high research costs in certain areas like pharmaceuticals which only big firms can afford to pay. Whether a HGF is younger or older than a LGF seems to depend on the regarded country. However, a typical HGF is older than 13 years. For the USA, Acs and Mueller (2008) concluded that HGFs are at least five years in business. Using a more restricting definition of HGFs[22], Acs et al. (2008, p. 1) find that their average

[21] For the most countries, an additional prototype describes a "typical" bankrupt firm. This will be further explained in Section 6.3.

[22] Acs et al. (2008) denote their regarded firms as "high-impact-firms" and regards both employment and revenue growth.

HGFs are between 17 and 34 years old and, therefore, explicitly mention that HGFs are not start-ups. Schreyer (2000, p. 19) also determined for Spain that the propensity to be a HGF does not decline with age and that there is no unidirectional effect of age. Henrekson and Johansson (2010, p. 227) concluded that HGFs are "not necessarily small and young". Coad (2007b, p. 51–52) added that since most start-ups do not come up with innovating ideas but rather replicate already existing business models so that there is no reason to expect remarkable growth from the majority of start-ups. Since a high rate of start-ups fails soon after market entry, start-ups often lead to a waste of economic resources. Shane (2009) argued similarly. Several past studies often came to the opposite conclusion that HGFs are usually younger than LGFs (Birch 1981, p. 8; Lopez-Garcia and Puente 2012, p. 1037; Becchetti and Trovato 2002, p. 17; Harhoff et al. 1998, p. 455; Boeri and Cramer 1992). This study's findings contradict the notion that a support of HGFs can be achieved by subsidizing (small and young) start-ups. Many countries have support programs, which mainly focus on start-ups. Still, the high contribution of small HGFs to job creation should not be overseen (Schreyer 2000, p. 7). Our findings might be caused by choice of the growth indicator. Hölzl (2014, p. 226) argues that the Birch-Schreyer growth indicator favors large firms. Since it is easier for a large firm to achieve a high indicator value, our RFs might be biased towards bigger firms. Additional to the previously stated arguments, this might be another reason why large firms become the center of attention as far as HGFs are concerned.

Moreover, typical HGFs most often have substantially higher Birch-Schreyer growth indicator values in the past than LGFs similar to the findings from the variable description section. This finding is in line with Lopez-Garcia and Puente (2012, p. 1037) who have found that "past extreme growth episodes increase the probability of current fast growth" and are "a significant predictor of current fast growth". Boeri and Cramer (1992) comes to a similar conclusion. In Portugal, both values are negative, but the HGFs' values are still higher.

As far as financial variables are concerned, in all countries except Poland, Sweden, and France, typical HGFs have a lower dr than their LGFs counterparts do (cf. Table 17). This confirms the findings of Becchetti and Trovato (2002) who conclude that financial restrictions limit the growth of Italian SMEs. Constrained HGFs are not able to finance all their potentially profitable projects (Becchetti and Trovato 2002, p. 297). Meanwhile, for French manufacturing HGFs this limitation could not be confirmed (Levratto et al. 2010). Our descriptive findings also do not hint to such a tendency.

Most but not all countries' HGF prototypes have a higher roa than the corresponding LGFs. It is surprising that only in Poland, HGFs have a higher spe than LGFs because one could expect that a firm will hire new employees if they "contribute" to a relatively large amount of sales. However, Coad and Broekel (2012) and Becchetti and Trovato (2002) came to a similar conclusion. For North American HGFs, Acs et al. (2008) found a higher revenue per employee values than for LGFs.

As far as the industry is concerned, previous findings that a large number of HGFs belongs to the service sector are confirmed. This is also the case for Spain, which was analyzed by Lopez-Garcia and Puente (2012) and also for Sweden, Germany, and the Netherlands (Schreyer 2000, p. 22). Because of similar results, Coad et al. (2014a, p. 98) argue that it is a common misconception among politicians that most HGFs are high-technology firms, which leads to misguided economic development schemes. To the contrary, Schreyer (2000, p. 22) find that a lot of HGFs came from manufacturing sector which we can confirm

for Poland and Finland. Levratto et al. (2010, p. 5) argue that only a few big manufacturing firms grow substantially due to a survival bias caused by high sunk costs and high capital investment. Our prototype-based findings are in line with the purely descriptive observations from Section 5.2.

6.3 Connection between high growth and bankruptcy

Section 5.2 explains why it is of major interest if a close relationship between HGFs and bankrupt firms (survivorship bias) exists as it is frequently mentioned in the literature. This section takes up on this question.[23]

A graphical out-of-sample approach can be seen in Fig. 2. Besides the two already discussed boxes, every country's boxplot contains a third box which depicts the HGF propensity distribution for companies which are known to have failed between 2011 and 2014. This period is also regarded to determine the HGF/LGF status of the firms. Therefore, a HGF propensity score was also estimated for every failing firm. These propensities are also presented in Fig. 2. The results clarify that defaulting firms usually do not have a higher propensity to grow and will usually not be wrongly predicted to be a HGF. The HGF-box always lies (more or less distinctively) above the box of the failed firms. In GB, where most accurate predictions have been obtained according to AUC, the two boxes do not overlap at all. For the second-best and third-best prediction (FR and IT), there is a small overlap so that high growth and bankruptcy rather seem to be antipodes. Still, the median HGF propensity of all HGFs lies above the 75% quantile of the HGF propensity of the bankrupt firms. The less accurate the prediction, the higher the intersection between these two boxes seems to be. This can be seen for countries like Finland, Portugal, and Sweden. The reason for these differences might be that for the last three countries, the corresponding RF was almost not able to recognize any useful HGF patterns so that virtually arbitrary propensities are assigned to both failing firms and HGFs. This leads to similar distributions.

Besides the already discussed HGF prediction results, Table 10 also presents what share of all failing firms can be found among the x% of all firms with the highest HGF propensity. Even for the three countries Finland, Portugal, and Sweden, the presented quantiles almost always contain a higher share of all HGFs than of all bankrupt firms. The latter share is also almost always smaller than the regarded quantile so that the hypothesis that bankruptcy and high growth are "two sides of one coin" is rejected.

Table 12 presents the results of a correlation analysis. For all countries except Finland and Sweden,[24] a second RF was estimated which predicts bankruptcy instead of high growth using train data. These RFs were used to generate an out-of-sample bankruptcy propensity score for all test data firms of their countries so that both a bankruptcy propensity and a HGF propensity are estimated for every firm. Table 12 shows the correlations between these

[23] Please note that no meaningful out-of-sample analysis is possible for Poland due to the low number of bankruptcies in the test data.

[24] Finland and Sweden are excluded due to the low number of bankrupt firms in the train data. In contrast to them, a RF can be created for Poland since its test data contains enough bankrupt firms. However, low numbers of bankrupt firms will also lead to less accurate bankruptcy propensity scores.

Table 12 Correlation between the propensity to be a HGF and the propensity to fail

Country	DE	ES	FR	GB	IT	PL	PT
Correlation	−0.3073	0.1559	−0.2219	−0.4481	0.0522	−0.5032	−0.0710

two propensities per country. A correlation of 1 would mean that high HGF propensities are always accompanied with high bankruptcy propensities whereas −1 would mean that high HGF propensities are always accompanied with low bankruptcy propensities. However, both results are highly unrealistic because neither high growth nor bankruptcy can be faultlessly predicted as shown in Kumar and Ravi (2007) for bankruptcy prediction.

In countries like GB and PL, there is a rather strong negative relationship between the two propensities. In Germany, France, and Portugal this relationship is also negative, but the corresponding correlation is quite low. A weak positive correlation is observed for Spain and Italy.

To confront HGFs and bankrupt firms, prototypes of bankrupt firms are presented as well for all countries where enough bankrupt firms are observed in the test data. The results can be found in Tables 11 and 17. It can be seen that the likelihood of confusion does not seem to be very high. The values of most of the variables differ strongly between bankrupt and highly growing firms. Based on the prototypes, it is more difficult to distinguish between LGFs and bankrupt firms. For example, typical HGFs always have a much higher number of employees than LGFs and bankrupt firms. Bankrupt firms are often younger than HGFs, always have a lower roa, and are stronger indebted than HGFs.

As a result, it can be said that no evidence could be found for the hypothesis that firms that are likely to grow have a high risk to fail. Based on the mentioned analysis, it is unlikely that subsidies for HGFs would lead to riskier economic activities. This does not mean that no firm that is going to be supported is going to fail, but the apprehensions from the literature cannot be confirmed for the regarded countries. Additionally, in countries where HGF prediction is rather reliable, the corresponding RF can usually distinguish between bankrupt and highly growing firms although they are not explicitly trained to do so.

7 Conclusion

Highly growing firms (HGF) or "gazelles" have been regarded in many studies because they are assumed to have an above-average contribution to job creation and revenue growth (Coad 2007a, p. 81; Acs and Mueller 2008, p. 86; Coad et al. 2014b, p. 92; Acs et al. 2008, p. 8; Lopez-Garcia and Puente 2012, p. 1030). Nowadays, important political institutions like the European Commission and the OECD foster HGFs to spur growth (Coad et al. 2014b, p. 93). To achieve a precise promotion of HGFs, it is crucial to have a mechanism to predict future HGFs using readily accessible data. The literature has been rather skeptical of the existence of such a mechanism due to the high heterogeneity of HGFs (Acs et al. 2008, p. 45).

This study concludes that although not every HGF could have been predicted in the past, a prediction is often still possible in most of the regarded countries. Therefore, we agree with Coad (2007b, p. 58) that regression-based analysis might be inappropriate to analyze HGFs.

In our study, we analyzed 179970 unique firms from nine European countries between 2004 and 2014 using the random forest (RF) algorithm. The RF is a modern data mining algorithm which evaluates a user-defined number of decision trees to assign a binary prediction to every regarded firm. This study describes a true out-of-sample prediction using 15 structural and financial variables, which have been determined using preliminary country-specific RFs. After having found a promising set of variables, the country-specific RFs were obtained by extensive cross-validations. It was shown for the first time that HGF prediction is most reliable in Great Britain, France, Italy and Spain and also worthwhile in Poland, Germany, Portugal, and Finland. In Sweden, a reliable out-of-sample prediction could not be realized. In the first four countries, the probability to assume a randomly chosen true HGF being more likely to be a HGF than a true LGF is above 76% for a 50% threshold. When regarding the top 15% of all firms according to their propensity to be a HGF, up to 49% of all contained HGFs can be correctly predicted. In contrast to studies like Coad (2007a) and Acs et al. (2008), it was found that especially precise predictions could be obtained for the second fifth of all firms with respect to their number of employees.

The most important variables for this prediction turned out to be the previous absolute value of the firm's size, the number of employees, and their first differences as well as the firm's age. However, certain financial variable based predictors like the fixed assets ratio, equity fixed asset ratio, and the sales per employee also contribute, but to a lower extent. This variables' ranking is to some extent country-specific. This is also the case for "typical" HGFs (prototypes) which is another contribution of this study. Prototype HGFs have more employees than a prototype LGF and have usually grown stronger in the past. Most HGFs are not start-ups which questions political initiatives to foster start-ups to enable future growth.

No evidence was found for a common concern in the literature that HGFs have often made risky but eventually successful decisions in the past so that fostering them also fosters risky behavior. The correlation between the propensity to be a HGF and the propensity to be a bankrupt firm is either close to zero or even negative which is the case in most countries. High propensities to grow intensively are only rarely assigned to failing firms because they turned out to be more similar to LGFs than to HGFs.

For future research, our findings imply that single countries' results should not be generalized due to the often substantial country-specific differences. Out-of-sample HGF prediction should no longer be considered to be impossible but analyzed thoroughly using further countries, variables, growth indicators and prediction models. Especially other prediction models and further variables might further improve the prediction quality. For instance, Lopez-Garcia and Puente (2012) and Barringer et al. (2005) point out the importance of extensive staff training and financial incentives for the employees, which could be considered as additional variables. The engagement in export of a firm and its research and development efforts also seem to be important predictors (Wagner 2007; Becchetti and Trovato 2002; Schreyer 2000). Moreover, Barringer et al. 2005, p. 664) highlighted the characteristics of the founder as an important factor of success. We did not analyze these variables due to very high amounts of missing values in our data set.

Before fostering HGFs, their effect on the economy should be analyzed carefully. The effects of existing HGFs have already been studied but the consequences of an increased share of HGFs is still unclear (Coad et al. 2014a, p. 101). It might even happen that a higher rate of HGFs comes at the expense of a lower industry growth because of high numbers of bankruptcies, which may not be desired by policy-makers (Bravo Biosca 2010, p. 2).

Appendix

Table 13 Median values of the regarded predictors

VAR	Y	DE	ES	FI	FR	GB	IT	PL	PT	SE
emp	1	741.0000	24.0000	68.0000	53.0000	510.5000	61.0000	200.0000	451.0000	30.0000
	0	228.0000	17.0000	22.5000	20.0000	116.0000	23.0000	100.0000	67.0000	12.0000
$\Delta_1 growth$	1	18.6794	2.2667	6.0000	3.1639	38.6296	3.2647	2.0571	−2.9789	3.8182
	0	−0.9310	1.0476	1.0455	0.0000	1.1250	0.0000	0.0000	0.0000	1.0303
$\Delta_2 growth$	1	12.7405	1.0286	2.4000	1.0556	19.2973	1.0039	0.0000	−4.4181	2.0000
	0	0.0000	0.0000	0.0000	0.0000	1.0051	0.0000	0.0000	0.0000	0.0000
$size$	1	141.1180	2.9740	8.1160	6.5435	132.2940	12.7570	13.7080	90.2925	3.4170
	0	56.2200	2.4500	2.6090	2.4480	23.6585	6.4180	5.1240	12.9535	1.5890
$\Delta_1 lsize$	1	0.0673	0.1260	0.0869	0.0783	0.1060	0.1040	0.1775	0.0206	0.1740
	0	0.0278	0.0903	0.0636	0.0524	0.0797	0.0626	0.0826	0.0477	0.1191
$\Delta_2 lsize$	1	0.1140	0.2583	0.1774	0.1694	0.2649	0.2068	0.3788	0.0978	0.2633
	0	0.0498	0.1982	0.1289	0.1187	0.1564	0.1250	0.2172	0.0847	0.1656
age	1	18.5000	12.0000	14.0000	15.0000	20.0000	18.0000	13.0000	16.5000	15.0000
	0	22.0000	14.0000	17.0000	16.0000	20.0000	20.0000	14.0000	22.0000	17.0000
spe	1	210.7599	149.4286	163.1852	170.4419	244.7566	254.5294	108.9800	177.9567	191.8990
	0	295.4376	157.7692	158.8596	185.4618	273.2857	263.0000	72.9714	139.7856	210.3636
$\Delta_1 spe$	1	6.6410	7.7500	3.8953	5.7953	11.9978	15.4286	9.2453	3.5886	16.8000
	0	12.3653	6.4284	8.4161	5.6667	13.8305	13.4173	3.0238	3.7458	15.4643
$\Delta_2 spe$	1	12.3388	13.4000	13.4286	11.2720	24.3979	21.4338	15.2050	10.2124	19.2692
	0	23.5572	10.2609	13.9341	11.1890	25.5095	17.0946	5.6219	5.1478	17.2549
far	1	0.4978	0.3604	0.4242	0.2000	0.3973	0.2204	0.4575	0.3595	0.2390
	0	0.4663	0.3365	0.4611	0.1898	0.2716	0.2019	0.4296	0.3338	0.2597

Table 13 (continued)

VAR	Y	DE	ES	FI	FR	GB	IT	PL	PT	SE
$efar$	1	0.8078	0.8426	0.9268	1.4248	0.8833	0.8928	1.0188	0.9620	1.2184
	0	0.8221	0.9380	0.8074	1.5803	1.1823	0.9723	1.0495	0.9959	1.3182
dr	1	0.6042	0.7223	0.6578	0.7107	0.6662	0.8093	0.5564	0.6858	0.6878
	0	0.6387	0.6957	0.6517	0.6733	0.6733	0.8069	0.5466	0.6888	0.6708
roa	1	0.0519	0.0706	0.0958	0.0766	0.0870	0.0685	0.0946	0.0665	0.1089
	0	0.0691	0.0644	0.0995	0.0722	0.0696	0.0646	0.0831	0.0543	0.0984

Table 14 Mean values of the regarded predictors

VAR	Y	DE	ES	FI	FR	GB	IT	PL	PT	SE
emp	1	5677.8846	152.9983	267.1623	178.0519	3637.3655	191.1883	532.2178	1274.6071	133.8549
	0	988.3213	46.0511	115.2054	78.4108	907.1166	66.3766	244.0096	180.1978	50.9804
$\Delta_1 growth$	1	677.3162	793.4141	68.1034	36.2335	406.7525	625.7118	101.9336	168.4152	178.4938
	0	621.5584	292.0567	536.3263	170.8490	166.0791	416.9338	31.2504	12.5770	21454.2848
$\Delta_2 growth$	1	138.6440	373.3770	19.9932	19.3966	330.3120	475.4572	55.6323	156.7642	147.0959
	0	49.1195	13.7324	95.2130	175.3367	30.0141	14.7091	9.7189	4.9278	28.3932
$size$	1	1920.5585	49.0501	109.7627	60.1924	1156.3686	61.8310	54.0854	306.4221	71.0709
	0	450.2017	17.7833	53.3718	18.9517	343.0546	27.3088	19.1491	150.9828	26.2142
$\Delta_1 lsize$	1	0.1063	0.1766	0.1741	0.1151	0.1500	0.1295	0.2212	0.1214	0.2167
	0	0.0486	0.1373	0.0917	0.0778	0.1042	0.0826	0.1236	0.0601	0.1579
$\Delta_2 lsize$	1	0.1804	0.3499	0.3008	0.2310	0.3462	0.2536	0.4433	0.2304	0.3403
	0	0.0940	0.2776	0.1888	0.1613	0.2010	0.1621	0.2701	0.1306	0.2288
age	1	39.5055	15.3567	22.9058	19.2841	29.7766	20.3586	18.0234	23.6429	21.1864
	0	40.9559	16.2553	22.4463	20.2071	28.3660	22.2729	21.7484	25.6043	22.5340
spe	1	635.3374	477.2196	397.7776	409.3885	1122.9355	814.3285	225.8191	409.5362	1616.5531
	0	825.4330	354.8207	395.0327	440.5974	1460.4445	614.6585	200.2155	1690.5017	472.6724
$\Delta_1 spe$	1	27.5143	73.1757	64.4675	21.1805	60.1420	148.3788	27.8188	−3.2468	141.0294
	0	−191.5782	20.9587	39.1417	−12.7828	7.1480	46.2830	9.5283	802.7140	72.4615
$\Delta_2 spe$	1	86.2378	101.8402	32.6458	45.9987	318.0350	167.3698	48.6757	13.3005	695.0565
	0	153.7807	31.3809	53.3803	63.7736	354.5191	59.1364	11.8587	1163.9828	91.7343
far	1	0.4901	0.3977	0.4393	0.2786	0.4131	0.2807	0.4545	0.3883	0.3213
	0	0.4757	0.3835	0.4611	0.2729	0.3322	0.2627	0.4501	0.3827	0.3442
$efar$	1	2.5147	2.5222	2.6550	5.9878	2.4655	2.8989	1.5172	2.0151	6.5225
	0	2.9678	4.2772	2.3694	6.3280	13.0545	3.7741	2.6434	4.0881	8.3166
dr	1	0.5863	0.6779	0.6131	0.6786	0.6503	0.7580	0.5465	0.6676	0.6618
	0	0.6171	0.6526	0.6374	0.6476	0.6446	0.7547	0.5315	0.6574	0.6405
roa	1	0.0714	0.0876	0.1129	0.0977	0.0889	0.0813	0.1048	0.0731	0.1154
	0	0.0739	0.0762	0.1120	0.0873	0.0718	0.0724	0.1019	0.0600	0.1089

Table 15 Standard deviation values of the regarded predictors

VAR	Y	DE	ES	FI	FR	GB	IT	PL	PT	SE
emp	1	28830.7662	1361.6994	714.6374	577.9930	10424.1198	625.8577	1217.0847	2118.0101	578.0722
	0	6497.8354	448.9777	1006.8135	734.0461	7221.5914	928.1347	1449.8252	401.5992	393.4336
$\Delta_1 growth$	1	4411.3361	37064.7653	380.3559	230.8458	3275.5254	24025.9313	347.6507	475.1003	5298.2440
	0	22316.8106	42415.9375	13304.3893	16291.2957	6497.3030	58696.0712	573.5593	129.8472	2724607.3287
$\Delta_2 growth$	1	1233.8456	25827.4607	116.6187	158.8220	1496.6733	21020.8576	213.6589	488.6376	5285.0937
	0	1439.0608	597.0636	2638.0963	18437.9133	1230.4187	600.5433	112.6738	64.3513	2694.7544
$size$	1	7819.4832	577.7689	319.6263	633.8934	3430.0210	244.2366	177.4183	647.4667	482.5197
	0	3830.5877	436.6211	597.2834	156.0408	4096.5009	466.9743	103.8775	889.0011	344.0090
$\Delta_1 lsize$	1	0.2063	0.3131	0.3048	0.2475	0.3549	0.2546	0.2693	0.3113	0.3513
	0	0.2647	0.2800	0.2455	0.2341	0.3095	0.2216	0.2416	0.1852	0.2833
$\Delta_2 lsize$	1	0.3003	0.4717	0.4347	0.3535	0.4758	0.3662	0.3915	0.4041	0.4934
	0	0.3905	0.4162	0.3561	0.3385	0.4036	0.3224	0.3397	0.3150	0.4092
age	1	44.7791	11.9639	23.6321	16.7824	25.9930	15.1737	18.3708	19.5206	20.1721
	0	46.8478	10.7741	19.2879	15.6908	24.5517	15.0214	24.3378	19.6956	18.4134
spe	1	2197.1564	2977.4410	794.5392	1647.5468	8875.2952	6682.7322	566.5598	970.1824	50770.0761
	0	6087.9559	1505.3229	1898.3675	3579.5516	21299.2975	3830.5393	690.5690	19041.9126	3014.8204
$\Delta_1 spe$	1	501.5383	1181.8304	261.6618	256.3288	496.5089	2107.6768	203.3222	156.0060	4559.5764
	0	11566.4343	925.8546	558.1152	3853.2029	4790.6484	1900.3039	218.9534	11619.5247	2148.9316
$\Delta_2 spe$	1	472.6791	1624.6064	432.7594	375.4259	3262.7889	2332.0586	347.4825	363.6978	27263.7027
	0	3720.1863	1518.4593	883.0259	2866.7033	5976.6092	2997.2225	504.7661	16447.8974	2392.5463
far	1	0.2525	0.2601	0.2523	0.2412	0.2648	0.2225	0.2321	0.2387	0.2779
	0	0.2789	0.2676	0.2677	0.2445	0.2736	0.2231	0.2547	0.2689	0.2975
$efar$	1	11.1517	14.3911	8.3543	49.6171	6.1820	14.6744	2.7107	4.4262	30.8425
	0	32.3533	59.6532	14.8754	63.8361	114.6051	32.9446	25.3305	15.5598	70.2268
dr	1	0.2318	0.2157	0.2171	0.1977	0.2058	0.1841	0.2169	0.1905	0.2047
	0	0.2256	0.2299	0.2089	0.2084	0.2163	0.1930	0.2213	0.1892	0.2178
roa	1	0.1214	0.1090	0.1207	0.1175	0.1692	0.0851	0.1760	0.0852	0.1743
	0	0.2082	0.0957	0.1358	0.1430	0.1407	0.0830	0.1206	0.0763	0.1581

Table 16 Inter-quantile range values of the regarded predictors

VAR	Y	DE	ES	FI	FR	GB	IT	PL	PT	SE
emp	1	1409.5000	57.0000	187.0000	129.2500	1750.0000	133.0000	390.0000	1128.7500	73.0000
	0	462.2500	25.0000	43.7500	36.0000	259.7500	43.0000	150.0000	140.0000	20.0000
$\Delta_1 growth$	1	130.9497	15.3425	34.3354	16.6039	235.0628	19.4161	73.4069	99.5071	15.6592
	0	31.5758	5.9429	6.3089	4.3788	26.2063	4.3333	10.7692	10.3314	3.5000
$\Delta_2 growth$	1	61.8226	6.7660	15.5376	8.4323	135.9505	8.4916	32.5611	48.4631	7.4818
	0	15.9745	3.1500	3.0536	3.0518	13.9673	3.0151	3.2812	7.1279	2.0909
$size$	1	348.7130	10.1900	43.9755	20.6760	557.3617	31.4080	29.4925	234.8020	12.1622
	0	121.0163	4.8380	8.6508	5.9845	67.6825	12.0715	9.5140	39.8882	4.1150
$\Delta_1 lsize$	1	0.1546	0.2777	0.2604	0.2245	0.1990	0.2219	0.2582	0.1698	0.2905
	0	0.1418	0.2461	0.2121	0.2007	0.2065	0.1895	0.2285	0.1648	0.2359
$\Delta_2 lsize$	1	0.2646	0.4779	0.4037	0.3492	0.3600	0.3567	0.4157	0.3992	0.4833
	0	0.2205	0.4067	0.3526	0.3140	0.3311	0.2997	0.3678	0.3076	0.3818
age	1	45.5000	12.0000	18.0000	15.0000	29.0000	17.0000	7.0000	26.2500	19.0000
	0	44.0000	12.0000	16.0000	15.0000	26.0000	18.0000	8.0000	16.7500	19.0000
spe	1	292.1207	243.2346	248.6313	210.6365	376.2819	373.6874	180.8992	247.5000	253.1873
	0	378.4740	221.5217	194.3732	231.2708	365.7970	370.7930	123.7588	236.1324	250.5208
$\Delta_1 spe$	1	30.8477	44.5920	35.6581	28.1453	40.6132	66.0369	28.5548	34.1024	48.9081
	0	50.1585	42.2998	34.4149	35.1058	48.1015	65.2497	16.8140	28.9500	46.6190
$\Delta_2 spe$	1	48.2824	62.1036	60.3904	39.9918	67.3602	89.1186	47.2561	63.6533	67.5282
	0	85.5997	57.3127	48.6966	48.5457	76.9600	85.0203	24.0680	43.0643	61.3112
far	1	0.3587	0.4113	0.3912	0.3432	0.4201	0.3112	0.3603	0.2575	0.4508
	0	0.4840	0.4098	0.4420	0.3174	0.4183	0.2937	0.3991	0.4201	0.4866
$efar$	1	0.8134	1.2764	1.2787	2.2869	1.2145	1.4176	0.9629	0.9554	3.1517
	0	0.8839	1.4266	1.1483	2.8920	2.5975	1.7646	0.9943	1.3131	3.4148
dr	1	0.3631	0.3066	0.3008	0.2833	0.3207	0.2425	0.3033	0.1760	0.2937
	0	0.3225	0.3347	0.3149	0.2947	0.3212	0.2559	0.3388	0.2385	0.3190
roa	1	0.0891	0.0869	0.1482	0.1129	0.0942	0.0710	0.1246	0.0701	0.1450
	0	0.1053	0.0761	0.1324	0.1073	0.0998	0.0651	0.1156	0.0708	0.1243

Table 17 Prototypes of all analyzed variables

Country	Variable	*emp*	*size*	$\Delta_2 growth$	$\Delta_1 growth$	*spe*	*far*	$\Delta_2 spe$	*age*	$\Delta_1 spe$	*efar*	$\Delta_2 lsize$	*sec*	$\Delta_1 lsize$	*dr*	*roa*
DE	LGF	212	10.8883	1.0026	1.0909	254.8451	0.4517	22.6727	22	11.9823	0.8804	0.1068	s_ma	0.0536	0.5968	0.0674
	HGF	924	11.8686	34.2333	23.8464	164.1937	0.5394	10.5135	23	4.8141	0.8847	0.1312	s_se	0.0749	0.5739	0.0572
	BANKR	227	10.1091	1.0175	0.0000	224.8625	0.2469	15.3910	16	12.4552	0.9570	0.0886	s_ma	0.0406	0.8168	0.0484
ES	LGF	17	7.8099	1.0556	0.0000	160.2222	0.3373	10.9982	14	6.7778	0.9362	0.1967	s_ma	0.0869	0.6948	0.0657
	HGF	210	9.2902	56.0000	26.2948	67.8500	0.4209	2.6162	16	1.1175	0.8065	0.2907	s_se	0.1421	0.6943	0.0787
	BANKR	24	8.0193	2.1333	1.0244	133.6250	0.2058	11.6733	11	6.9545	0.5955	0.3780	s_ma	0.1754	0.8848	0.0658
FI	LGF	20	7.8698	0.0000	0.0000	180.9000	0.4557	22.4138	17	11.3083	0.8965	0.1135	s_ma	0.0598	0.6356	0.1081
	HGF	172	9.9565	15.8824	9.6532	176.3741	0.3897	5.9240	17	2.0084	0.9341	0.1637	s_ma	0.0934	0.6251	0.0863
FR	LGF	18	7.5939	0.0000	0.0000	183.0000	0.1830	8.7559	15	4.1118	1.6333	0.1227	s_se	0.0550	0.6627	0.0757
	HGF	104	9.3796	7.7778	3.1765	173.3333	0.2174	10.1831	16	5.7687	1.3589	0.1862	s_ma	0.0855	0.7197	0.0828
	BANKR	17	7.3330	0.0000	0.0000	175.2000	0.0894	14.8205	15	7.2941	1.8893	0.2217	s_se	0.0795	0.8040	0.0567
GB	LGF	126	10.1365	2.0519	1.0417	281.7551	0.2911	25.2520	18	13.7278	1.1663	0.1311	s_ma	0.0745	0.6915	0.0737
	HGF	1528	13.0414	181.3740	121.6351	240.9103	0.4906	26.9205	26	11.1804	0.7672	0.3166	s_se	0.1158	0.6735	0.0932
	BANKR	80	9.0107	1.0213	0.0000	204.6889	0.1641	15.8000	14	11.8331	1.8592	0.1624	s_se	0.0750	0.7239	0.0428
IT	LGF	25	8.7942	0.0000	0.0000	260.5500	0.2058	16.3799	20	13.8476	0.9384	0.1260	s_ma	0.0659	0.8080	0.0642
	HGF	251	10.2083	30.7909	14.8370	142.0491	0.2639	3.5093	17	2.5962	0.7678	0.2233	s_so	0.0946	0.8011	0.0656
	BANKR	23	8.5702	0.0000	0.0000	192.4545	0.1762	11.2367	13	8.0770	0.3250	0.1980	s_ma	0.0948	0.9458	0.0598
PL	LGF	100	8.1887	0.0000	0.0000	54.6250	0.4596	2.8846	14	1.3844	1.0096	0.1275	s_ma	0.0451	0.5271	0.0792
	HGF	250	9.8240	8.1905	0.0000	142.2100	0.4425	24.3900	13	13.1162	0.9903	0.4242	s_ma	0.1943	0.5718	0.0999
	BANKR	60	7.5027	0.0000	0.0000	49.7800	0.3699	−4.8619	14	−4.5976	0.7669	0.0776	s_ma	0.0267	0.7008	0.0595
PT	LGF	34	9.4661	−9.4819	−8.5714	209.8571	0.3287	27.1071	15	12.7738	0.8936	0.1890	s_ma	0.0827	0.7196	0.0659
	HGF	707	12.1114	−7.9599	−6.9663	186.2193	0.3619	13.1029	23	5.6433	0.9790	0.0576	s_fi	0.0103	0.6916	0.0652
	BANKR	79	8.7823	−4.6711	−8.5235	99.3548	0.3248	−8.8693	20	3.2867	0.6572	0.0175	s_ma	0.0714	0.7989	0.0554
SE	LGF	12	7.2689	1.0385	0.0000	211.5000	0.2595	15.5840	17	14.3810	1.2640	0.1713	s_se	0.1204	0.6778	0.0971
	HGF	90	8.8899	28.8000	13.5745	167.1801	0.1967	15.4583	14	17.9141	1.5841	0.4784	s_se	0.2384	0.6869	0.1459

References

Ablameyko S (2003) Neural networks for instrumentation, measurement and related industrial applications, 1st edn. IOS Press, Crema

Acs Z, Parsons W, Tracy S (2008) High-impact firms: gazelles revisited. Small Business Research Summary (328):1–92. http://econpapers.repec.org/bookchap/elgeebook/16552.htm

Acs ZJ, Mueller P (2008) Employment effects of business dynamics: Mice, gazelles and elephants. Small Bus Econ 30(1):85–100

Aiginger K (2006) Competitiveness: from a dangerous obsession to a welfare creating ability with positive externalities. J Indust Compet Trade 6(2):161–177

Albrecht WS, Stice EK, Stice JD (2007) Financial Accounting, 1st edn. Cengage Learning

Alpaydin E (2004) Introduction to machine learning, vol 1. MIT Press, Massachusetts

Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J Finance 23(4):589–609

Audretsch DB, Mahmood T (1994) Firm selection and industry evolution: the post-entry performance of new firms. J Evol Econ 4(3):243–260

Baily MN, Bartelsman EJ, Haltiwanger J (1996) Downsizing and productivity growth: Myth or reality? Small Bus Econ 8(4):259–278

Barringer BR, Jones FF, Neubaum DO (2005) A quantitative content analysis of the characteristics of rapid-growth firms and their founders. J Bus Ventur 20(5):663–687

Batista G, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. ACM Sigkdd Explor Newslett 6(1):20–29

Becchetti L (1995) Finance, investment and innovation: a theoretical and empirical comparative analysis. Empirica 22(3):167–184

Becchetti L, Trovato G (2002) The determinants of growth for small and medium sized firms. the role of the availability of external finance. Small Bus Econ 19(4):291–306

Becker HP (2010) Investition und finanzierung: grundlagen der betrieblichen finanzwirtschaft, 4th edn. Gabler Verlag, Wiesbaden

Behr A, Weinblat J (2017) Default patterns in seven eu countries: A random forest approach. Int J Econ Bus 24(2):181–222

Birch D, Medoff J (1994) Gazelles. In: Solmon L, Levenson A (eds) Labor Markets, Employment Policy and Job Creation. Westview Press, Boulder, pp 159-168

Birch DL (1981) Who creates jobs?. The public interest 65:3–14

Boeri T, Cramer U (1992) Employment growth, incumbents and entrants: evidence from Germany. Int J Indust Organ 10(4):545–565

Bravo Biosca A (2010) Growth dynamics: Exploring business growth and contraction in europe and the us. Research report, NESTA

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (2001) Random forests. Mach Learn 45(1):5—-32

Breiman L, Cutler A (2004) Random forests. http://www.math.usu.edu/adele/forests/cc_home.htm

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Wadsworth International Group, Belmont

Brown I, Mues C (2012) An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst Appl 39(3):3446–3453

Buenstorf G, Cantner U, Hanusch H, Hutter M, Lorenz HW, Rahmeyer F (2013) The Two Sides of Innovation: Creation and Destruction in the Evolution of Capitalist Economies. Springer Science & Business Media, Dordrecht, London

Chandra DK, Ravi V, Bose I (2009) Failure prediction of dotcom companies using hybrid intelligent techniques. Expert Syst Appl 36(3):4830–4837

Chawla NV (2005) Data mining for imbalanced datasets: An overview. In: Data Mining and Knowledge Discovery Handbook. Springer, pp 853–867

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

Chen KS, Babb EM, Schrader LF (1985) Growth of large cooperative and proprietary firms in the us food sector. Agribusiness 1(2):201–210

Coad A (2007a) A closer look at serial growth rate correlation. Rev Indust Organ 31(1):69–82

Coad A (2007b) Firm growth: A survey. Doc Trav Centre d'Econ Sorbonne 24:1–72

Coad A, Broekel T (2012) Firm growth and productivity growth: evidence from a panel var. Appl Econ 44(10):1251–1269

Coad A, Hölzl W (2009) On the autocorrelation of growth rates. J Indust Compet Trade 9(2):139–166
Coad A, Daunfeldt SO, Hölzl W, Johansson D, Nightingale P (2014a) High-growth firms: introduction to the special section. Indust Corp Chang 23(1):91–112
Coad A, Daunfeldt SO, Johansson D, Wennberg K (2014b) Whom do high-growth firms hire? Indust Corp Chang 23(1):293–327
Cross EP, Rarnchandani H (1995) Comparing classification accuracy of neural networks, binary logit regression and discriminant analysis for insolvency prediction of life insurers. J Econ Finan 19(13):1–18
Daunfeldt SO, Halvarsson D (2015) Are high-growth firms one-hit wonders? evidence from Sweden. Small Bus Econ 44(2):361–383
Daunfeldt SO, Elert N, Johansson D (2014) The economic contribution of high-growth firms: do policy implications depend on the choice of growth indicator? J Indust Compet Trade 14(3):337–365
Dunne T, Roberts MJ, Samuelson L (1989) The growth and failure of us manufacturing plants. Q J Econ 104(4):671–698
European Commission (2010) Communication from the commission europe 2020: A strategy for smart, sustainable and inclusive growth. Technical report
Fagiolo G, Luzzi A (2006) Do liquidity constraints matter in explaining firm size and growth? some evidence from the italian manufacturing industry. Indust Corp Chang 15(1):1–39
Fawcett T (2006) An introduction to roc analysis. Pattern Recogn lett 27(8):861–874
Fotopoulos G, Louri H (2004) Firm growth and fdi: Are multinationals stimulating local industrial development? J Indust Compet Trade 4(3):163–189
Frydman H, Altman EI, Kao DL (1985) Introducing recursive partitioning for financial classification: The case of financial distress. J Finan 40(1):269–291
Gibrat R (1931) Les inégalités économiques. Recueil Sirey
Gorunescu F (2011) Data Mining: Concepts, Models and Techniques, vol 1. Springer Science & Business Media
Han J, Kamber M, Pei J (2011) Data mining: concepts and techniques, 3rd edn. Morgan Kaufmann, Amsterdam, Boston, Heidelberg, London
Härdle W, Moro R, Schäfer D (2005) Predicting bankruptcy with support vector machines. In: Statistical Tools for Finance and Insurance. Springer, pp 225–248
Harhoff D, Stahl K, Woywode M (1998) Legal form, growth and exit of west german firms—empirical results for manufacturing, construction, trade and service industries. J Indust Econ 46(4):453–488
Hart WE, Krasnogor N, Smith JE (2005) Recent advances in memetic algorithms, 1st edn. Springer Science and Business Media, Berlin, Heidelberg
Hassan MR, Ramamohanarao K, Karmakar C, Hossain MM, Bailey J (2010) A novel scalable multi-class roc for effective visualization and computation. In: Zaki MJ, Yu JX, Ravidran B, Pudi V (eds) Advances in Knowledge Discovery and Data Mining, Part I: 14th Pacific-Asia Conference. Springer-Verlag, Berlin, Heidelberg, pp 107-120
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, 2nd edn. Science + Business, Springer, New York
Henrekson M, Johansson D (2010) Gazelles as job creators: a survey and interpretation of the evidence. Small Bus Econ 35(2):227–244
Hölzl W (2014) Persistence, survival, and growth: a closer look at 20 years of fast-growing firms in Austria. Indust Corp Chang 23(1):199–231
Jovanovic B (1982) Selection and the evolution of industry. Econ: J Econ Soc 50(3):649–670
Kartasheva AV, Traskin M (2011) Insurers' insolvency prediction using random forest classification. http://anastasiakartashevaphd.com/3.pdf
Krzanowski WJ, Hand DJ (2009) ROC curves for continuous data. CRC Press
Kumar PR, Ravi V (2007) Bankruptcy prediction in banks and firms via statistical and intelligent techniques – a review. Eur J Oper Res 180(1):1–28
Lam M (2004) Neural network techniques for financial performance prediction: integrating fundamental and technical analysis. Decis Support Syst 37(4):567–581
Levratto N, Zouikri M, Tessier L (2010) The determinants of growth for smes-a longitudinal study from french manufacturing firms. Technical report, CNRS-EconomiX, https://www.researchgate.net/profile/Nadine_Levratto/publication/228271512_The_Determinants_of_Growth_for_SMEs_-_A_Longitudinal_Study_from_French_Manufacturing_Firms/links/02e7e514a1ea6bf9f0000000.pdf
Löbbe H (2001) Klassifizierung landwirtschaftlicher jahresabschlüsse mittels neuronaler netze und fuzzy systeme PhD thesis. Rheinischen Friedrich-Wilhelms-Universität zu Bonn, Hamm
Lopez-Garcia P, Puente S (2012) What makes a high-growth firm? a dynamic probit analysis using spanish firm-level data. Small Bus Econ 39(4):1029–1041

Maimon O, Rokach L (2006) Data mining and knowledge discovery handbook. Springer Science & Business Media, Tel-Aviv
National Commission on Entrepreneurship (2011) High-growth companies: Mapping america's entrepreneurial landscape. Technical report, National Commission on Entrepreneurship
Ohlson JA (1980) Financial ratios and the probabilistic prediction of bankruptcy. J Account Res 18(1):109–131
Olson DL, Delen D, Meng Y (2012) Comparative analysis of data mining methods for bankruptcy prediction. Decis Support Syst 52(2):464–473
Organisation for Economic Co-operation and Development (2010) High-growth enterprises: What governments can do to make a difference. OECD Publish 1(1):1–238
Pagans FG (2015) Predictive Analytics Using Rattle and Qlik Sense. Packt Publishing Ltd
Penner SJ (2004) Introduction to health care economics & financial management: fundamental concepts with practical applications, 1st edn. Lippincott Williams & Wilkins, New York, London
Puri S (2012) Introduction to retail math, vol 1. Introduction to Retail Math, India
Pytlik M (1995) Diskriminanzanalyse und künstliche Neuronale Netze zur Klassifizierung von Jahresabschlüssen: Ein empirischer Vergleich. Europäischer Verlag der Wissenschaft, Frankfurt am Main
Rokach L (2007) Data mining with decision trees: theory and applications. series in machine perception and artificial intelligence world scientific. Hackensack, London
Schneider O, Lindner A (2010) The value of lead logistics services. In: Vallespir B, Alix T (eds) Advances in Production Management Systems. New Challenges, New Approaches, pp 315–322
Schreyer P (2000) High-growth firms and employment, oECD Science, Technology and Industry Working Papers
Shane S (2009) Why encouraging more people to become entrepreneurs is bad public policy. Small Bus Econ 33(2):141–149
Shin KS, Lee TS, jung Kim H (2005) An application of support vector machines in bankruptcy prediction model. Expert Syst Appl 28(1):127–135
Shirata CY (1998) Financial ratios as predictors of bankruptcy in Japan: an empirical research. Tsukuba Coll Technol Jpn 1(1):1–17
Stickney C, Weil R, Schipper K, Francis J (2009) Financial accounting: an introduction to concepts, methods and uses, 1st edn. Cengage Learning, Mason
Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. Bioinformatics 8(25):1–21. http://www.statistik.lmu.de/carolin/research/varimppaper_techreport.pdf
Vause B (2009) Guide to analysing companies the economist. Wiley, New York
Van Dijk Electronic Publishing GmbH B (2015) amadeus. http://www.bvdinfo.com/de-de/our-products/company-information/international-products/amadeus
Venkatraman E (2000) A permutation test to compare receiver operating characteristic curves. Biometrics 56(4):1134–1138
Venkatraman E, Begg CB (1996) A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. Biometrika 83(4):835–848
Verikas A, Gelzinis A, Bacauskiene M (2010) Mining data with random forests: A survey and results of new tests. Pattern Recogn 44(2):330–349
Wagner J (2007) Exports and productivity: A survey of the evidence from firm-level data. World Econ 30(1):60–82
Williams G (2011) Data mining with rattle and R: The art of excavating data for knowledge discovery. Springer Science & Business Media, New York
Witten IH, Frank E, Hall MA (2011) Data mining: Practical machine learning tools and techniques: practical machine learning tools and technique. Elsevier, Amsterdam, Boston
Yeh CC, Chi DJ, Lin YR (2014) Going-concern prediction using hybrid random forests and rough set approach. Inf Sci 254:98–110
Zhou XH, Obuchowski NA, McClish DK (2014) Statistical Methods in Diagnostic Medicine. Wiley
Zighed DA, Komorowski J, Zytkow JM, Zytkow J (2000) Principles of data mining and knowledge discovery: 4th european conference, PKDD, 2000, Lyon, France, Proceedings, vol 1. Springer Science & Business Media, Berlin, Heidelberg, New York