



# Catching Gazelles with a Lasso: Big data techniques for the prediction of high-growth firms

Alex Coad · Stjepan Srhoj

Accepted: 13 March 2019 / Published online: 17 June 2019  
© Springer Science+Business Media, LLC, part of Springer Nature 2019

**Abstract** We investigate whether our limited ability to predict high-growth firms (HGF) is because previous research has used a restricted set of explanatory variables, and in particular because there is a need for explanatory variables with high variation within firms over time. To this end, we apply “big data” techniques (i.e., LASSO; Least Absolute Shrinkage and Selection Operator) to predict HGFs in comprehensive datasets on Croatian and Slovenian firms. Firms with low inventories, higher previous employment growth, and higher short-term liabilities are more likely to become HGFs. Pseudo- $R^2$  statistics of around 10% indicate that HGF prediction remains a challenging exercise.

**Keywords** LASSO · High-growth firms · Prediction · Within variation · Firm growth · Post hoc interpretation · Inventories

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s11187-019-00203-3>) contains supplementary material, which is available to authorized users.

---

A. Coad (✉)  
CENTRUM Católica Graduate Business School (CCGBS), Lima, Peru  
e-mail: acoad@pucp.edu.pe

A. Coad  
Pontificia Universidad Católica del Perú (PUCP), Lima, Peru

S. Srhoj  
Department of Economics and Business Economics, Lapadska obala 7, University of Dubrovnik, 20000 Dubrovnik, Croatia  
e-mail: stjepan.srhoj@unidu.hr

**JEL classification** L25 · L26

## 1 Introduction

There is considerable interest among entrepreneurs, investors, and policymakers in predicting tomorrow’s high-growth firms (Henrekson and Johansson 2010; Mason and Brown 2013; McKenzie 2017; Grover Goswami et al. 2019). Since the seminal contribution of David Birch (1979), much excitement has surrounded high-growth firms because of their remarkable ability to create jobs, their potential to create wealth, and their substantial contributions to creative destruction and productivity growth. An improved ability to predict high-growth firms (HGFs) is crucial for investors who seek to allocate funds to the right firms, for policymakers seeking to craft effective framework conditions to support job creation, and for entrepreneurs with ambitions to grow.

Previous research has suggested that HGFs are a heterogeneous group (Delmar et al. 2003; Daunfeldt et al. 2014) and are difficult to predict, although there are a small number of empirical regularities, for example that HGFs are often younger, smaller, and less common in high-tech sectors (Henrekson and Johansson 2010). Are HGFs hard to predict because firm growth is fundamentally random, or because previous investigations had only a small number of (the wrong type of) explanatory variables? This we seek to investigate. We are uniquely positioned to examine the latter explanation, because we have large datasets from two countries with an extensive range of explanatory variables.

Moreover, these explanatory variables have rich information on variation within firms over time.

Amid the proliferation of research into firm growth, new opportunities for HGF prediction have recently been made possible by “big data” approaches to predict firm outcomes (George et al. 2014; van Witteloostuijn and Kolkman 2019), involving sophisticated econometric techniques (Belloni et al. 2014).

We contribute to the sparse literature on HGF prediction in a number of ways. First, our review of the literature emphasizes the richness of our data, in particular with regard to having a large number of time-varying variables, which improves our prediction accuracy and identifies the most relevant variables. Second, we alleviate concerns over the possible over-theorizing of potentially spurious results by analyzing two nationally representative datasets, from Croatia and Slovenia. Third, we apply big data econometric techniques to select which variables from among the hundreds of candidates are the best predictors of HGFs. Previous published work has applied LASSO (Least Absolute Shrinkage and Selection Operator) to bankruptcy prediction (e.g., Tian et al. 2015), and a few working papers have applied LASSO to predicting firm growth and performance (Miyakawa et al. 2017; McKenzie and Sansone 2017).<sup>1</sup> Van Witteloostuijn and Kolkman (2019) apply a big data technique (random forest analysis, not LASSO) to investigate the determinants of a firm’s growth rate of assets (whereas our dependent variable is a dummy for HGF status). We are among the first to apply LASSO to the tasks of predicting firm growth and HGF status.

Our LASSO procedure identifies a number of significant predictors of HGF performance, and the model fit is modest (pseudo- $R^2$  statistics of around 10%). Empirical results suggest that firms with lower inventory, higher previous employment growth, higher short-term liabilities, and higher growth in terms of exports and assets are more likely to become HGFs. Internal finance seems to be more relevant than external finance for predicting rapid growth.

Section 2 discusses the related literature, emphasizes the need for predictor variables with a high within-firm variation, and discusses our post hoc approach to theory

development. Section 3 presents the databases, and Section 4 presents our LASSO estimator and algorithm. Section 5 presents our results on Croatian and Slovenian data. Section 6 discusses our findings. Section 7 concludes.

## 2 Background

### 2.1 Related literature

A “first wave” of early applied economics research into firm growth sought to uncover the factors associated with firm growth, generally using databases on the largest firms that were listed on public stock exchanges. These studies investigated the role of predictor variables such as firm size (according to Gibrat’s law of proportionate growth: Ijiri and Simon 1964), growth rate autocorrelation (Ijiri and Simon 1967; Singh and Whittington 1975), the phenomenon of growth through acquisition (Kumar 1985), and discussed themes such as the contribution of firm growth to industrial concentration (Singh and Whittington 1975; Kumar 1985). Other work investigated the effects on growth of variables such as firm age (Evans 1987) and R&D investments (Hall 1987).

A “second wave” of research into firm growth, in the last few decades, resulted in a large amount of published research on the determinants of firm growth, using richer datasets (often administrative datasets collected by national statistical offices) with a more comprehensive coverage of small and young firms, a wider set of explanatory variables, and more emphasis on longitudinal as opposed to cross-sectional datasets (Davidsson and Wiklund 2000). These studies also benefitted from advanced econometric techniques and more powerful computers. Some exemplary studies include Geroski et al. (1997) on the role of profitability, Harhoff et al. (1998) on the role of legal form, and Audretsch et al. (1999) on the growth of new ventures.

Despite this multiplication of research into firm growth, however, progress was slow, and there was disappointment with our ability to predict which firms will grow (Achtenhagen et al. 2010; Davidsson et al. 2010; McKelvie and Wiklund 2010). Geroski (2000: p. 169)<sup>2</sup> summarized in this way:

<sup>1</sup> The working papers by Miyakawa et al. (2017) and McKenzie and Sansone (2017), who apply LASSO to firm performance data. Miyakawa et al. (2017) seek to predict high growth performance in a sample of Japanese firms, although they use a non-standard definition of high-growth firms. McKenzie and Sansone (2017) investigate top 10% growth among business plan competition winners and non-winners in Nigeria. These working papers came to our attention at an advanced stage of this research.

<sup>2</sup> See also the exchange between Derbyshire and Garnsey (2014) and Coad et al. (2015) on the randomness of growth. We are grateful to a reviewer for this suggestion.

“The most elementary ‘fact’ about corporate growth thrown up by econometric work on both large and small firms is that firm size follows a random walk”

This state of affairs suggests a change of approach. One shift in research focus has been to move away from seeking the determinants of the growth rate of the average firm, towards an emerging strand of literature that uses a binary distinction to reflect whether a firm is included or not in an elite club of “high-growth firms” (Henrekson and Johansson 2010). Another change of research direction has been to expand the list of explanatory variables of firm growth, including finer-grained variables relating to founder characteristics, industry and geographical aspects, productivity, profitability, innovation, and the growth of rivals (Coad 2009), and also including imaginative variables such as whether the firm’s name is concise and whether the firm’s name is eponymous (Guzman and Stern 2016), and the entrepreneur’s score on a Raven test of abstract reasoning (McKenzie and Sansone 2017).

We therefore contribute to the emergence of a “third wave” of empirical research into firm growth, using big data and computationally intensive techniques, and measuring growth using a binary indicator that distinguishes high-growth firms, using the well-known Eurostat-OECD definition (Eurostat-OECD 2007). Previous attempts at HGF prediction (i.e., focusing specifically on cases where the dependent variable is binary and indicates whether a firm is an HGF) are in the literature review table below. Table 1 below shows that many studies seeking to predict HGFs use time-invariant variables, which is hard to justify given that HGF status is transitory and unlikely to be repeated.

## 2.2 The need for explanatory variables with high within-firm variation

A fundamental challenge for research into firm growth concerns the need to include explanatory variables that vary within firms over time:

“If we truly wish to explain corporate growth rates in terms of observables, we need to find variables, which have statistical congruent properties with growth; i.e. that vary much more over time for

particular firms than they vary across firms at any given time.”

(Geroski and Gugler 2004, p. 618)

This is because firm growth is, by its very nature, an erratic and time-varying process (Penrose 1959). Firms can be conceived as configurations of lumpy and inter-dependent resources, such that some resources (e.g., managerial skills and attention, production capacity, distribution channels) are not being fully utilized at any particular moment in time, leading to slack (Nason and Wiklund 2018). This slack spurs firms on to take advantage of growth opportunities, e.g., through diversification, to more efficiently utilize existing resources (Coad and Planck 2012). However, learning effects (whereby the use of existing resources such as human resources becomes more efficient over time) and the further addition of other indivisible resources brought on by growth, means that the degree of slack resources in the firm is constantly shifting and jumping, and that new opportunities for growth are constantly appearing.

Empirical research has shown that the variation in annual growth rates *within* firms over time is greater than the variation in growth rates *between* firms (Geroski and Gugler 2004; see also Coad and Rao 2011). Relatedly, a stylized fact of the HGF literature is that there is little persistence in rapid growth, which has shifted the discussion to refer to “high growth episodes” rather than “high growth firms” (Grover Goswami et al. 2019).

Storey (2011) argues that the erratic and volatile nature of firm growth is hard to reconcile with the focus of entrepreneurship scholars on relatively time-invariant variables such as education of the business owner, opportunity recognition capabilities, networking skills, and human capital.

Indeed, the “usual suspects” in terms of explanatory variables in growth rate regressions are variables that are invariant over time: whether they be founder-level characteristics (gender, education, pre-entry experience) or firm-level characteristics (e.g., legal form, industry sector) or other variables (region dummies). Some variables do vary over time, but in ways that are deterministic (e.g., age of the company), or are the same for large groups of firms (e.g., industry concentration, regional startup density), or have low within-firm variation over time (e.g., R&D expenditure, firm size, capital intensity, number of subsidiary plants) and hence also have a limited capacity to address Geroski’s requirement to include time-varying firm-specific explanatory variables.

**Table 1** Previous studies of HGF prediction, where the dependent variable is binary and indicates whether a firm is an HGF

Authors	Data	Model	HGF definition	Explanatory variables	(Pseudo)- $R^2$
Lopez-Garcia and Puente (2012)	1411 Spanish firms, 1996–2003	Probit with correlated random effects	Birch-Schreyer indicator	Young firms, debt ratio, wage premium, share of permanent workers, previous HGF, sector interaction terms	N/A
Arrighetti and Lasagni (2013, Table 4)	777 Italian manufacturing firms, 1998–2003	Probit	Top 10% fastest growing firms by (a) employment and (b) sales	Acquisition dummy, index of industrial production, sum of jobs lost in the industry, credit rationing proxy, log sales, log age, % ownership, human capital index, return on equity, total factor productivity, Pavitt classification dummies	0.03–0.12
Bjuggren et al. (2013, Table 3)	All private firms in Sweden, 1993–2006	Probit	Top 1% fastest growing firms, absolute and relative employment growth	Firm age, number of employees, enterprise group, family ownership, legal form, industry dummies, year dummies	0.004–0.007
Daunfeld, Johansson and Elert (2014)	All limited liability companies in Sweden, 1997–2010	Probit	Top 1% fastest growing firms over 3, 5, or 7 years	Lagged firm size, firm age, period of birth, group membership dummy, industry classification	N/A
Holz (2014, Table 8)	Austrian private sector firms with 1+ employees, 1972–2007	OLS	Eurostat-OECD and Birch-Schreyer indicator	Previous HGF, firm age, firm size, log industry size in employment, industry growth, excess labor turnover as a proxy for mobility barriers	0.007–0.096
Lee (2014, Table 1)	2007–2010 survey data on 4858 UK SMEs	Probit	20% employment growth per year over 2 years	Process and product innovation, change of owner, multiple directors, external advice	0.015
Goedhuys and Sleuwaegen (2016, Table 1)	21,372 Belgian firms with 10+ employees, 2008–2011	Probit	Eurostat-OECD	Firm size, firm age, R&D, higher education, university education	0.079
Guzman and Stern (2016, Table 3)	Over 12 million business registrants in 12 US states, 1988–2014	Logit	Growth dummy variable = 1 if the startup achieves an initial public offering (IPO) or is acquired at a meaningful positive valuation within 6 years of registration	Corporate governance (Corporation, Delaware), firm name (short name, eponymous), patents, trademarks, cluster dummies (local, traded resource intensive, traded), high-tech clusters (biotech, E-commerce, IT, medical devices, semiconductors), state dummies	0.060–0.272
Bianchini et al. (2017)	Panel data of Italian, Spanish, French and UK firms, 2004–2011	Multinomial probit (PHG vs HG vs other)	High-growth (HG) and persistent high-growth (PHG) firms, w.r.t. the top 10% of the growth distribution	Total factor productivity, ROA, intangible assets, interest expenses/sales, leverage, age, sales, number of employees, sector and country dummies	N/A
Weinblat (2017)	179,970 firms from 9 European countries, 2004–2014	Random forest approach	HGF = 1 if a firm's Birch-Schreyer growth index is in the top 10% of all firms from the same country and data set over a period of 3 years	Financial ratios (debt ratio, ROA, ROS, fixed assets ratio, liquidity ratio, equity/fixed assets, sales per employee), Firm descriptive: size (in total assets); legal form; age; Birch-Schreyer growth indicator; number of employees, and sector classifications	N/A
McKenzie and Sansone (2017)	2506 participants in business plan competition in Nigeria	Support vector machines, boosting regression, LASSO	Top 10% in employment, sales, or profits among business plan competition participants	Gender, age, education, location, marital status, family composition, languages spoken, employment history, risk profiles, motivation for running the firms, self-confidence, available assets, loans, challenges, firm name, business plan assessment, digit-span recall, Raven test, personality grit	N/A (accuracy 80.5–84.7%; sensitivity 9–21.7%)
Miyakawa et al. (2017)	800,000–1,700,000 firms for the period 2006–2014 in Japan	Weighted random forest, LASSO	Sales or profit growth exceeds the average plus one standard deviation within the same 2-digit industry	Solvency scores, sales (and sales change), profits (and profit change), number of employees, capital, dividend payment (and its change), firm age, owner age, number of establishments, industry, geographic information, supply-chain network	N/A (sensitivity 22–25%)
		Probit			0.071

Table 1 (continued)

Authors	Data	Model	HGF definition	Explanatory variables	(Pseudo)- $R^2$
Megaravalli and Sampagnaro (2018)	AIDA data on 45,000 family business SMEs in Italy, 2010–2014		> 20% growth for both 2011–2013 and 2012–2014, as well as growth < 20% in 2010	Liquidity ratio, solvency ratio, firm age, cash flow, working capital, industry, and province dummies	
Moschella et al. (2018)	22,988 Chinese manufacturing firms with 8+ employees, 1998–2007	OLS	High-growth dummy = 1 for firms in the top 20% of a period's distribution of average growth rates, in terms of at least 1 of the 2 growth measures (employment or sales)	Lagged HGF status, productivity, ROS, investment intensity, output share of new products, interest expenses, leverage, firm age, firm size, sector and region dummies, exporting dummy, state-controlled dummy	0.03–0.06
Pereira and Temour (2018)	Orbis data on 11 Central & Eastern European countries, 2000–2013	Probit	Eurostat-OECD	Age, number of employees, Herfindahl index, fixed assets, foreign owned, and institutional indicators (relating to corruption, investment profile, and bureaucratic quality)	N/A

Notes: N/A denotes “not applicable.” Holzl (2014, Table 8) reports adjusted  $R^2$  statistics

This paper therefore seeks to investigate the role of time-varying variables in predicting high growth. Indeed, previous work has mentioned that the varying amounts of slack resources over time, and the opportunities offered by idiosyncratic configurations of discrete productive resources, can affect a firm’s growth rates (Coad and Planck 2012). Brown and Mawson (2013) develop the concept of “growth trigger points” to describe how some firms may be well-positioned for a period of rapid growth as a function of time-varying variables such as new capital investments, new bank funding, or boosts to sales coming from obtaining a new contract or customer. However, previous work has not been able to show how periods of slack resources can precede a growth spurt, because previous work (see, e.g., the HGF prediction literature reviewed in Table 1) has not had access to the detailed time-varying firm-specific variables that are found in our data. Our LASSO approach, in combination with our detailed datasets, is well-suited to investigate the role of time-varying firm-specific variables, because a large number of firm-specific variables can be included in the same LASSO model to obtain a parsimonious final model which highlights the most important variables for HGF prediction.

### 2.3 Epistemological approach

Our context of applying big data techniques to large datasets implies that we are engaging in exploratory data-driven empirical research, as a fact-finding exercise that can hopefully contribute to subsequent theory building (Helfat 2007). It would be premature to formulate elaborate hypotheses, given the exploratory nature of our analysis (Hambrick 2007; Helfat 2007; Locke 2007).<sup>3</sup> Instead of formulating a list of hypotheses, we investigate the following broad research question: Which variables are associated with becoming a high-growth firm? In particular, following the recommendations of Geroski and others, we are interested in the explanatory role of time-varying variables with a high within-variance (i.e., vari-

<sup>3</sup> Relatedly, Bernerth et al. (2018) recommend that the control variables be mentioned specifically in the formulation of the hypotheses. Of course, we cannot do this in our context, because we have several hundred explanatory variables, and we apply data-driven procedures to decide which of these explanatory variables to keep. Nevertheless, in step 7 of Algorithm 1, we add in a minimalist set of control variables that are included for theoretical reasons, i.e., sector dummies, year dummies, a dummy for the Zagreb capital region, and firm age.

ables for which there is a large variation over time for any individual firm).

The management field's insistence on formulating and testing hypotheses (Hambrick 2007; Helfat 2007) may lead to a situation whereby the hypotheses are formulated after the results are known (the practice of HARKing—or Hypothesizing After the Results are Known (Kerr 1998)). This kind of post hoc theorizing can be detrimental to scientific progress, if elaborate theoretical explanations are formulated retrospectively to explain results that may be essentially spurious (Denton 1985; Kerr 1998). HARKing and post hoc theorizing can lead to the misinterpreting and over-theorizing of statistically significant results that are simply due to random sampling error.

In this article, we draw on the literature of firm growth, and more specifically high-growth firms, to provide an initial orientation to our big data analysis. In particular, we draw on the interest and curiosity of Geroski and others regarding the potential predictive role of variables with a high within-variance. However, we make no claims to omnisciently predict what our results will look like.

LASSO is a useful statistical tool for variable selection when databases contain large numbers of explanatory variables (Belloni et al. 2012). The selection of relevant variables occurs using statistical rather than theoretical reasons. Our LASSO algorithm will not be left alone to run entirely free though, devoid of theoretical guidance, but it is operated in a “semi-supervised” way, with certain methodological choices being made by the authors during the calculations (e.g., manually fine-tuning the penalization parameter  $\lambda$  in order to obtain a reasonable number of variables in the LASSO output, and dropping LASSO-selected variables that are very highly collinear, and including a minimal set of control variables for theoretical reasons (see footnote 3)). Furthermore, we will obtain independent results using alternative dependent variables (HGFs in terms of either sales growth or employment growth (Delmar 1997, Shepherd and Wiklund 2009)). To avoid overfitting the data, we randomly split our data into train and test samples, for both the Croatian and Slovenian datasets. After finding the important variables in the train sample, we use these variables on the test sample to confirm their importance.

We then engage in post hoc interpretation and discussion of our results. At all costs, we avoid “sharking” (secretly HARKing; Hollenbeck and Wright 2017); rather, we transparently recognize the post hoc nature of our

discussion. Nevertheless, there are advantages to post hoc analysis of scientific data (Hollenbeck and Wright 2017; Vancouver 2018), that are useful in our context of exploratory analysis of big data (Hollenbeck and Wright 2017; Vancouver 2018), because our discussions can benefit from being guided by newly discovered phenomena.

We then return to some exploratory data analysis after discussing the initial results, as recommended by Hollenbeck and Wright (2017). In particular, we apply the LASSO-selected variables from the Croatian data to the Slovenian dataset, as a further check against any overfitting and sampling bias that could be specific to any one country's dataset. Hence, although we engage in post hoc interpretation of exploratory data-driven empirical analysis, nevertheless our methodology is robust against the perils of post hoc interpretation and possible “data mining.”

### 3 Data

#### 3.1 Croatian data

The main database in this paper stems from the census data of the Financial Agency (FINA) of the Republic of Croatia. All limited liability firms are obliged by law to report their balance sheets as well as their profit and loss statements to FINA. The advantage of having a census dataset is coverage of firms from all industries and of all sizes, while at the same time missing values do not pose a serious issue. Previous work on this same dataset includes Peric and Vitezic (2016) as well as Vitezic et al. (2018). The dataset spans 2003–2016, the year 2003 corresponds to the year of financial reporting changes at FINA (hence reducing the comparability of data from previous years), while 2016 is the last reported year. We deflate all the monetary variables by the Eurostat's NACE 2-digit output deflators.<sup>4</sup>

For the dependent variables, we apply the Eurostat-OECD definition of HGF (Eurostat-OECD 2007) to create a dummy variable that takes 1 for firms that have

<sup>4</sup> At the Eurostat web site, under the National accounts aggregates by industry (up to NACE A\*64), we obtain current prices, million units of national currency and previous year prices, million units of national currency. To obtain the share of current in previous year prices, the two are divided and were set at constant prices in 2010. When two-digit NACE deflators were not possible to obtain, the one-digit deflators were used (e.g., mining and quarrying, the NACE one-digit deflators were used for the four separate NACE 2-digit sectors).

at least 10 employees in the initial period, and a geometric average of at least 20% growth per year, over a 3-year period, i.e.:

$$E_{t=0} \geq 10 \quad (1)$$

$$\left( \frac{S_{t+3}}{S_t} \right)^{\frac{1}{3}} - 1 \geq 20\% \quad (2)$$

Where  $E$  is the number of employees, and  $S$  is firm size (measured in terms of either sales or employees). The dependent variable is the Eurostat-OECD HGF dummy, calculated for growth of either sales or employment.<sup>5</sup>

The starting number of observations is 1,189,275 (138,766 unique firms) in the period 2003–2016 from which 1.34% are HGFs (by either Eurostat-OECD employment or turnover growth indicator). We construct lagged variables (similar to van Witteloostuijn and Kolkman 2019) for the period  $t-2$  and then drop observations in years 2003 and 2004, as these have missing values in their lagged variables. We also exclude observations in years 2014, 2015, and 2016 as there is no 3-year period and insufficient information to construct an HGF indicator. We needed to clean the variable age which occasionally had incorrectly specified values, thus we excluded observations with negative age and age over 100 years. This leads to 734,773 observations (120,389 unique firms), with 1.53% HGFs. This is lower than the well-known “vital 6%” figure obtained for the UK (NESTA 2009), although it is about twice as large as the proportions found in neighboring Slovenia (Vitezić et al. 2018).

We further exclude firms with fewer than 3 employees,<sup>6</sup> annual turnover lower than one average monthly wage in the Republic of Croatia, and firms from the public sector, agriculture and construction. This leads to 212,769 observations (45,465 unique firms) and 4% HGFs. We here split the dataset into two, as micro firms cannot become HGFs by the Eurostat-OECD (2007)

definition (because technically they need to have 10 or more employees in the initial period). For model 1, we drop firms with fewer than 10 employees which leads to 79,109 observation and 10.75% HGFs.<sup>7</sup> For the model 2, we keep firms with 3 or more employees, but modify the HGF definition. Within this dataset, for firms with 10 and more employees in period  $t$ , the Eurostat-OECD (2007) is applied (turnover or employment criteria), while for firm with 3–9 employees in period  $t$ , we apply a definition of employment growth of 7.8 employees over the next 3-year period ( $t$  to  $t+3$ ) in order to be classified as HGFs.<sup>8</sup> The dataset for model 2 consists of 212,769 observations (45,465 unique firms) and 5.22% HGFs.

In regard to the independent variables, the dataset is composed of 172 variables, of which 109 come from the balance sheets and 45 from profit and loss statements, these are enriched with variables on demographic information, including firm age, year of financial report, capital region dummy, economic activity by technological intensity and knowledge intensity, as well as number of employees, exporting value, and importing value. We construct dummies for micro, small, medium, and large firms (following the classification of the European Commission). Our independent variables are set in period  $t$ . In addition, we add log changes in independent variables between the period  $t-2$  and  $t$ . This way, we insert log changes of all continuous financial independent variables, to end up with the final number of independent variables being 325. While it is good to have a large number of candidate variables for HGF prediction, nevertheless we have too many variables to include them all in the same regression. LASSO therefore is an ideal tool to select the most relevant variables from among the 325. Many of our independent variables are right-skewed, which motivates the log-transformation of variables to reduce the influence of outliers.<sup>9</sup> Online Appendix 1 gives information on the

<sup>5</sup> Previous research has shown that employment growth and sales growth are the two most common indicators of firm growth, and we include them both because they are alternative and complementary indicators that capture different aspects of the firm growth process (Delmar 1997; Shepherd and Wiklund 2009).

<sup>6</sup> Including firms with 1 or 2 employees was not possible, because the LASSO computations could not converge to a solution. However, this does not seem to be a problem because, despite the large number of firms with one or two employees, nevertheless these firms make a small aggregate contribution to the national economy, and moreover these micro firms are relatively unlikely to become HGFs (Neumark et al. 2011). Note also that the Eurostat-OECD HGF definition excludes all firms with fewer than 10 employees.

<sup>7</sup> Note that the share of HGFs jumps up from 1.53 to 10.75% when we exclude firms with fewer than 10 employees. This could explain why some countries have higher HGF shares than others—it could be because the databases being used have different coverage of micro firms (e.g., Coad and Scott 2018).

<sup>8</sup> The number 7.8 comes from the minimum possible growth increment to become an HGF according to the OECD definition. A firm with 10 employees in the first year, with average annual growth of 20% over 3 years, will need to grow by  $10 \times [1.20^3 - 1] = 7.28$  employees.

<sup>9</sup> By applying the natural log transformation on all variables, we are in line with the recommendations of Makridakis et al. (2018, p. 21) to automate the preprocessing of data before the application of data-intensive forecasting methods, to avoid the role of potentially ad hoc decisions being made by the researcher.

Table 2 LASSO-selected variables—summarized results

	Croatian sample		Slovenian sample		Croatian sample		Slovenian sample	
	Model 1 (Employment growth, 10+ employees)	Model 2 (Employment growth, 3+ employees)	Model 2 (Employment growth, 3+ employees)	Model 2 (Employment growth, 3+ employees)	Model 1 (Turnover growth, 10+ employees)	Model 2 (Turnover growth, 3+ employees)	Model 1 (Turnover growth, 10+ employees)	Model 2 (Turnover growth, 3+ employees)
Exports	.	.	.	.	YES (+)	YES (+)	YES (+)	YES (+)
Sales	YES (+)	YES (+)	YES (+)	YES (+)	YES (-)	.	.	.
Profits	YES (+)	YES (+)	YES (+)	YES (+)	.	.	.	.
Reserves	YES (-)	YES (-)	YES (-)	YES (-)	.	.	.	YES (-)
Cash in bank	YES (+)	.	.	.	.	.	.	.
Inventories	YES (-)	YES (-)	YES (-)	YES (-)	YES (-)	YES (-)	YES (-)	YES (-)
Intangible assets	YES (+)	YES (+)	YES (+)	.	.	YES (+)	YES (+)	.
Fixed assets	.	.	.	.	YES (+)	YES (+)	YES (+)	.
Subsidies, and grants	.	.	YES (+)	YES (+)	YES (+)	YES (+)	YES (+)	.
Short-term liabilities	.	YES (+)	YES (+)	YES (+)	YES (+)	YES (+)	YES (+)	YES (+)
Long-term liabilities	.	.	.	.	.	YES (+)	YES (+)	YES (+)
Long-term liabilities towards group firms	.	.	.	.	.	YES (+)	YES (+)	.
Other expenses	.	.	.	.	YES (+)	YES (+)	YES (+)	.
Costs of services	.	.	.	.	.	.	.	YES (-)
Amortization	.	.	YES (-)	YES (-)	.	.	.	.
Capital region	YES (+)	YES (+)	YES (+)	YES (-)	YES (+)	YES (+)	YES (+)	YES (-)
Log age	YES (-)	YES (-)	YES (-)	Not available	YES (-)	YES (-)	YES (-)	Not available
Growth in exports	YES (+)	YES (+)	YES (+)	YES (+)	.	.	.	.
Growth in assets	YES (+)	YES (+)	YES (+)	YES (+)	.	.	.	.
Growth in employees	YES (+)	YES (+)	YES (+)	YES (+)	YES (+)	YES (+)	YES (+)	YES (+)
Growth in profits	.	YES (+)	.	.	.	.	.	.
Growth in inventories	YES (+)	YES (+)	YES (+)	YES (+)	.	.	.	.
Growth in other long-term operating liabilities	.	.	YES (-)	YES (-)	.	.	.	.
Growth in intangible assets	.	.	.	.	YES (+)	YES (+)	YES (+)	.

Note: Table reports whether independent variable was selected and confirmed on the test sample. In case it was not selected or confirmed on the sample, it is denoted by “.”. If the variable was selected, the sign in the brackets denotes if the association between the variable and high-growth status is positive or negative



level of detail from balance sheets and profit and loss statements. Online Appendix 2 gives a description of variables in the Croatian dataset, while summary statistics are given in Online Appendix 3.

### 3.2 Slovenian data

In addition to the dataset of firms in the Republic of Croatia, we also use a very similar database of firms in the Republic of Slovenia. This dataset stems from the Agency of the Republic of Slovenia for Public Legal Records and Related Services (AJPES). Firms of all sizes and types registered in Slovenia are obliged to deliver their annual financial statements to AJPES. This dataset was used in several research papers (e.g., De Loecker 2007; Srhoj et al. 2018). The database provides text files with balance sheets, profit and loss statements and additional financial information, encompassing 193 different financial variables. The initial dataset consists of 455,925 observations (85,172 unique firms) in the period 2007–2014, out of which 0.47% are HGFs. The small initial percentage of HGFs shows Eurostat-OECD (2007) definition is overly restrictive for smaller countries<sup>10</sup> (for case of Slovenia: Srhoj et al. 2018) which is why the modified HGF definition (model 2) is used for the Slovenian dataset. We repeat the variable creation and data cleaning procedure as for the dataset of firms in the Republic of Croatia. The final sample consists of 35,758 observations (14,096 unique firms), 2.83% HGFs, and 403 independent variables. The description of variables and their summary statistics is available in Online Appendices 5 and 6.

## 4 Methods

In a time of big data and increased computational power, an important question is which variables should be selected in statistical models. Least Absolute Shrinkage and Selection Operator (LASSO), first introduced by Tibshirani (1996), is a powerful method that performs regularization and variable selection (see Tibshirani 2011). The assumption behind LASSO is the approximate sparsity condition, that is, the relatively small subset among predictors used is different from zero (Belloni et al. 2014). It applies a penalization process

to the independent variables, decreasing some variables to zero, thus leaving only those most important variables for explaining the dependent variable. It can be said that LASSO is the state-of-art method for variable selection, as it outperforms the standard stepwise logistic regressions (e.g., Tong et al. 2016) and also outperforms adaptive LASSO and elastic net (e.g., Fan et al. 2015). There are also different views, some suggest using elastic net instead of LASSO when the number of independent variables is larger than the sample size, and when variables are correlated (Zou and Hastie 2005). In our setting, the sample size is many times larger than the number of independent variables, and although some variables are correlated, the firm-level literature finds elastic net not to outperform LASSO in variable selection (Sermpinis et al. 2018), which is why LASSO is often used for variable selection in bankruptcy prediction studies (e.g., Tian et al. 2015) and lately is used in prediction of firm growth (e.g., McKenzie and Sansone 2017; Miyakawa et al. 2017).

The LASSO estimator can be written as a solution to the following optimization problem:

$$\beta^{lasso} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \hat{Q}_l(\beta) + \frac{\lambda}{n} \|\hat{\Upsilon}_l \beta\|_1, \quad (3)$$

Where  $\hat{\Upsilon}_l = \operatorname{diag}(\hat{\gamma}_{l1}, \dots, \hat{\gamma}_{lp})$  is a diagonal matrix specifying penalty loadings.<sup>11</sup> The key idea behind the penalty loading is to introduce self-normalization of the FOC of the LASSO problem using data-dependent penalty loadings, therefore applying self-normalized moderate deviation theory (see Belloni et al. 2012). Loadings enable obtaining sharp convergence results for the LASSO estimator. In addition to the diagonal matrix of penalty loadings, a penalty level  $\frac{\lambda}{n}$  has to be selected in order to dominate the noise to all  $k_e$  regression problems simultaneously.

$$P\left(\frac{\lambda}{n} \geq c \max_{1 \leq l \leq k_e} \|S_l\|_\infty\right) \rightarrow 1 \quad (4)$$

where  $\lambda = c2\sqrt{n}\Phi^{-1}(1-\gamma/(2k_e p))$ , with  $\gamma \rightarrow 0$ ,  $\log(\frac{1}{\lambda}) \leq \log(pvn)$ , that implements (4). The parameter  $p$  denotes covariates and  $n$  is number of observations. We use the recommended (Belloni et al. 2012)

<sup>10</sup> Croatia has twice larger population (4.15 million) in comparison to Slovenia (2.07 million).

<sup>11</sup> These data-driven penalty loadings for LASSO are different from the canonical penalty loadings proposed in Tibshirani (1996).

**Table 3** Categorization of financial variables by importance

Category	Variables
Generally influential (selected and validated in at least five out of six models)	Growth in employees (positive) Inventories (negative) Short-term liabilities (positive)
Sensitive to growth indicator (selected and validated in all three models of particular growth indicator)	Exports (positive for turnover indicator) Sales (positive for employment indicator) Profits (positive for employment indicator) Reserves (negative for employment indicator) Growth in exports (positive for employment indicator) Growth in assets (positive for employment indicator)
Sensitive to growth indicator and country context (selected and validated in both models in Croatia, but not confirmed in Slovenia or selected with a combination of three models)	Fixed assets (positive for turnover in Croatia) Subsidies and grants (positive for turnover in Croatia and positive for employment in Slovenia) Intangible assets (positive for employment in Croatia) Other expenses (positive for turnover in Croatia) Growth in inventories (positive for employment in Croatia)
Sensitive to model selection (selected and validated with model 2 of both countries, but not with model 1)	Log long-term liabilities (positive for turnover indicator)
Generally non-influential (selected and validated in only one out of six models)	Cash in bank (positive for employment indicator in Croatia model 1) Amortization (negative for employment indicator in Slovenia model 2) Growth in profits (positive for employment indicator in Croatia model 2) Growth in other long-term operating liabilities (negative for employment indicator in Slovenia model 2) Growth in intangible assets (positive for turnover indicator in Croatia model 2) Cost of services (positive for turnover indicator in Slovenia model 2) Long-term liabilities towards group firms

confidence level of  $\gamma = 0.1 / \log(p \vee n)$ , and constant  $c = 1.1$ . These are used in the R package *hdm* (*High-Dimensional Metrics*) for penalty parameter calculation in the function *lambdaCalculation* (Chernozhukov et al. 2016). The penalty level obtained this way is used as a starting penalty level. The higher the penalty, the lower is the number of variables selected. Given the exploratory nature of our investigation, we decrease the level of penalty gradually until the number of selected financial variables is 6–8.<sup>12</sup>

We focus on the logistic LASSO regression<sup>13</sup> (Belloni et al. 2016; p. 8) where  $y$  can take either values

<sup>12</sup> The decision on the number of financial variables per LASSO procedure is left to the researchers. When *lambdaCalculation* gave penalty that selected only few variables, we gradually decrease the penalty. Details on the penalty level are given in the Online Appendix 8.

<sup>13</sup> Logit LASSO has the same intuition as in the linear LASSO case because logit regression can be reduced to the linear case by employing reweighted regression.

of 1 or 0. The regularization works by adding the penalty to the log-likelihood function:

$$\sum_{i=1}^n (-Y_{i,t}(\beta_0 + \beta' X_{i,t})) + \log(1 + \exp(\beta_0 + \beta' X_{i,t})) - \lambda \sum_{k=1}^p |\beta_k| \quad (5)$$

The logit LASSO selects only those variables with highest predictive power of HGF status. The logistic LASSO is implemented using the function *rlassologit* (in R package *hdm*).

We consider two models. Model 1 includes observations with at least 10 employees, where two dichotomous dependent variables are constructed for turnover and employment-based indicators, in line with the Eurostat and OECD (2007) definition. The second model includes observations with at least 3 employees in period  $t$  and a modified HGF definition is used where a firm needs to have an increase of at least 7.8 employees in the forthcoming 3-year period to be classified as an HGF.

## ALGORITHM 1:

1. Randomly split sample into two datasets (train and test) with stratification based on HGF status. 70% of the data is training (for in-sample prediction) and the other 30% is test sample (for out-of-sample prediction).
2. Calculate the data-driven penalty on the training sample.
3. Logistic LASSO regression on HGF indicators (see regression equation (5) above) on the 70% of the full sample (the so-called train sample).
4. LASSO will deliver a set of selected variables
5. Penalization is incrementally modified (if needed) until 10-12 financial variables are selected.
6. Some LASSO-selected variables will be highly correlated between them (e.g. 'profits before taxation' & 'net profits') leading to multicollinearity. If this is the case, identify which variables have pairwise correlations over 70%, to check for severe multicollinearity, and decide which one to remove on a case-by-case basis. Hence, our algorithm is semi-supervised rather than fully automated.
7. Take these selected variables as the basis of a second set of regressions, which are estimated using logistic regression and include a set of controls that we add in for theoretical reasons (i.e. sector controls, year dummies, a dummy for the capital region (the Zagreb region), and firm age). This second set of regressions is used on the 30% of the sample (so-called test sample) which helps establish the statistical significance of the LASSO-selected variables. In addition, LASSO selected variables are applied to the full sample.
8. Based on the predicted values from the test sample and full sample calculate the accuracy, sensitivity and specificity rates.

		Predicted values	
		Not predicted as high-growth episode	Predicted as high-growth episode
Actual values	Not high-growth episode	True Negative (TN)	False Positive (FP)
	High-growth episode	False Negative (FN)	True Positive (TP)

These formulas are used to calculate accuracy, sensitivity and specificity rates:

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ number\ of\ observations}$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

Accuracy, sensitivity, and specificity statistics are reported at the bottom of the regression results tables.

As a sensitivity check, we repeat the *Algorithm* ten times to check whether the variable selection is sensitive to the random split into train and test samples. This sensitivity check shows stability in variable selection.<sup>14</sup> Finally, we use variables selected on the training sample of the Croatian dataset and apply them to the Slovenian dataset.

## 5 Analysis

We begin by presenting the results for Croatia, before investigating their external validity with our Slovenian data. Table 2 summarizes the LASSO-selected variables across models, and Table 3 summarizes the variables according to their importance.

### 5.1 Main results for Croatia

Our main results for Croatia are in Tables 4, 5, 6, and 7 in Appendix 1. These results are estimated for two subsamples: model 1 is estimated for firms with 10+ employees, while model 2 is estimated for firms with 3+ employees (as explained in Section 3). Model 2 therefore has far more observations (e.g., 212,769 in Table 6 compared to 79,109 in Table 4, for employment HGFs).

Overall, the predictive power of our LASSO methodology is relatively high with respect to the previous literature surveyed in Table 1.<sup>15</sup> The McFadden pseudo- $R^2$  varies from 0.085 to 0.136 in the 6 results tables. Irrespective of whether our estimated coefficients correspond to associations or causal effects, the predictive power of our model is mildly encouraging.

Our most stable results for Croatia, that are observed irrespective of sample (model 1 or model 2), and irrespective of growth indicator (employment HGFs or sales HGFs) are that previous growth of employees, and short-term liabilities, are positively associated with subsequent HGF status, while raw materials, supplies, and inventories are negatively associated with HGF status.<sup>16</sup>

The LASSO selection of some variables is sensitive to growth indicator (employment HGFs or sales HGFs).

For the employment HGF indicator, a number of variables corresponding to firm size are significant predictors of HGF status; these variables are sales, profits, and assets. Sales and profits are positively associated with HGF status, and “cash in bank & cash in hand” is significant in model 1. Intangible assets are also positively associated with HGF status. Therefore, holding all other influences constant (including some crude size dummies for micro, small, medium-sized and large firms), firms with higher sales, profits, and fixed assets are more likely to be employment HGFs.

For employment HGFs, the coefficient for reserves is negative, perhaps because HGFs have many productive opportunities, and they face the urgent challenges of preparing for rapid growth, and they reinvest their profits in capital assets and corporate infrastructure. Some variables that correspond to growth (prior to the HGF episode) are selected by the LASSO model: such as growth in exports, growth of assets, and growth of profits. Each of these three growth variables is positively associated with HGF status.<sup>17</sup>

Regarding the sales HGF indicator, exports and fixed assets are positive predictors of HGF status. Intangible assets are also positively related to HGF status.<sup>18</sup>

The role of some of the variables varies from model 1 to model 2, therefore being sensitive to the inclusion (or not) of micro firms with three or more employees. The logarithm of long-term liabilities is positive for the turnover indicator in model 2, i.e., when micro firms are included. This could be because the availability of long-term liabilities is especially valuable for micro firms as a source of financial resources. Relatedly, the variable “liabilities towards group firms” is also positive for model 2—this is an interesting (and surely endogenous) finding, whereby micro firms that perceive attractive growth opportunities may benefit from the financial support of their enterprise group. Finally, cash in bank is positive for employment HGFs in model 1, which provides further support of the role of financial performance for subsequent HGF status.

### 5.2 Analysis of Slovenian data

One of the dangers of post hoc theorizing after exploratory data analysis is that sampling error could be

<sup>14</sup> These results are available from the authors upon request.

<sup>15</sup> Nevertheless, note that the studies in Table 1 display heterogeneity regarding their HGF indicators (Birch index, top 10% of firms, etc.) as well as size of firms in the samples, which limits the comparability of the pseudo- $R^2$  statistics across studies.

<sup>16</sup> It is possible that the usage and significance of inventories differs between manufacturing and services sectors. We therefore repeated the analysis on subsamples of manufacturing and services sectors, and the results for inventories remained.

<sup>17</sup> Note that growth of profits is only selected by LASSO in model 2, for Croatian employment HGFs.

<sup>18</sup> The level of intangible assets is positive in the subsample of all firms with 3 or more employees (i.e., model 2), while growth of intangible assets is positive in the subsample of all firms with 10 or more employees (i.e., model 1).

mistakenly construed as evidence of economically meaningful effects (Denton 1985; Kerr 1998; Hollenbeck and Wright 2017). A high model fit on one country's dataset is not necessarily a good predictor of forecasting accuracy with a different country's sample (Makridakis et al. 2018). Therefore, we continue our analysis of the determinants of HGFs using a new dataset (census data on Slovenian firms, described in Section 3.2).

Our main results for Slovenia are in Tables 8 and 9 in Appendix 2. Overall, there is substantial overlap with the Croatian results, in terms of the variables selected by LASSO. In particular, the variables that overlap most prominently are growth of employees, inventories, and short-term liabilities. Log of long-term liabilities is also positive for the turnover HGF indicator.

In the Slovenian data, some of the LASSO-selected variables overlap with the Croatian results for one growth indicator, but not for the other. In such cases, therefore, the differences between employment HGFs and sales HGFs are larger than the differences between Slovenian firms and Croatian firms. Regarding employment HGFs, it is sales, profits, reserves, growth in exports, and growth in assets that are selected by LASSO for Slovenia as well as for Croatia. With regard to sales HGFs, it is growth of exports that matters for HGF status in both Slovenia and Croatia.

In some cases, there are variables associated with HGF status in Slovenia that were not relevant for Croatia. For example, subsidies and grants are positively associated with employment HGFs in Slovenia, but not for Croatia. (In fact, in Croatia, "subsidies, donations and compensations" are positively associated with sales HGFs). Cost of services is also positively associated with sales HGF status in Slovenia.

In a few cases, the results for Slovenia contrast with those for Croatia. For instance, being located in the capital region is positively related to HGF status in Croatia, but the relation is negative in Slovenia. Regarding sector of activity, high-tech KIS firms are *ceteris paribus* more likely to be HGFs in Croatia (for both sales and employment HGF indicators), but high-tech KIS firms are less likely to be HGFs in Slovenia (for the employment HGF indicator).<sup>19</sup>

<sup>19</sup> One possible explanation for the varying results could be that the regression specifications for Slovenia do not include an age variable, because this variable is not present in the Slovenian data.

Finally, some of the variables selected by LASSO for Croatia were not selected for Slovenia. These variables include fixed assets and "other expenses" (for sales HGFs) and intangible assets (for employment HGFs).

### 5.3 Applying the Croatian LASSO-selected variables to Slovenia

As a further robustness check, the LASSO-selected variables from the Croatian data were taken and applied to the Slovenian data (see Online Appendix 9). Many of the Croatian LASSO-selected variables are significant in the Slovenian data, and the McFadden  $R^2$  statistics are reasonably high, which suggests that there is substantial overlap in the predictors of HGFs in Croatia and in Slovenia.

## 6 Discussion

### 6.1 General comments

Overall, there is considerable overlap in terms of the LASSO-selected variables in the two countries. This suggests that the LASSO-selected variables are not simply being chosen due to random sampling error, but that there is a more systematic relationship between these variables and HGF status. In some cases, the predictor variables are more sensitive to the choice of growth indicator (employment growth or sales growth) than they are to the choice of country, indicating that the differences between growth indicators overshadow the differences between country contexts (at least for the cases of Croatia and Slovenia). Another interesting observation is that there are more variables selected as being associated with subsequent employment growth than there are for being associated with subsequent turnover growth. Nevertheless, this should be interpreted together with the observation that the McFadden  $R^2$  statistics are roughly similar for model 1, for the two growth indicators, and in the case of model 2, the McFadden  $R^2$  for the sales growth logit regressions is actually slightly *higher* than the McFadden  $R^2$  for the employment growth regressions (0.136 vs 0.103 for the full samples). This latter observation on the basis of  $R^2$  statistics suggests that it is very slightly easier to predict the HGF status of micro firms in terms of sales than in terms of their employment growth.

Our theoretical discussion in Section 2.2 proposed that there is an important role of firm-specific time-

varying variables as predictors of HGF status. Online Appendices 4 and 7 show the proportions of within and between variables for the LASSO-selected variables in Croatian and Slovenian data, respectively. As expected, the LASSO-selected variables that predict HGF status are not time-invariant, but have a relatively high share of within-firm variation over time. This suggests that HGF prediction with the usual set of time-invariant explanatory variables (mentioned in Section 2.2) will not go far in understanding the determinants of rapid growth.

Our analysis has put forward a large number of predictor variables, as could be expected from our big data approach. However, our post hoc inductive theorizing on the basis of our exploratory data analyses will take the form of focusing on several variables that are strongly and robustly associated with HGF status. Three prominent variables are inventories, growth in employees, and short-term liabilities (discussed in the subsections below). We also discuss the role of internal finance and external finance, even though these variables are not always selected by LASSO as important predictors of HGF status, but because of previous theoretical interest in this matter.

It is also worth mentioning some variables that were *not* selected by our LASSO procedure. Several variables often put forward as key drivers of rapid growth are found *not* to be important in our analysis, such as “R&D expenditures,” “concession rights, patents, commodity and service brands, software and other rights,” and “goodwill.” These seem not to be important in the Croatian nor Slovenian datasets. Interesting also is that variables relating to the use of external finance (such as bank loans) do not appear prominently as predictors of HGF status (this will be discussed further below).

## 6.2 Specific variables

### 6.2.1 Inventories

An original and yet intriguing finding concerns the relationship between inventories (also referred to as “raw materials and supplies”) and HGFs. The relationship between inventories and rapid growth episodes has received little attention in the previous literature (e.g., Table 1), although the importance of lean inventories—from the perspective of management practices relating to “Lean Management”—has often been lauded by management consultants, international organizations such as the OECD, and government support schemes for SMEs.

Lean management suggests that firms should try to keep inventories low to boost efficiency and minimize waste.

In addition to the standard advantages of lean management, there are some advantages that are particularly relevant for HGFs. It is well known that HGFs are under great pressure to balance costs and revenues (Churchill and Mullins 2001). Costs of production and costs of growth often are paid long before the corresponding revenues can be recovered. Indeed, it can take a long time for firms to send invoices and receive payments from clients, even before taking late payments into account. As a result, many HGFs have difficulties balancing costs and revenues, and these difficulties may increase their chances of exit (Churchill and Mullins 2001; Davidsson et al. 2009). HGFs that can keep inventories low will enjoy lower costs of production, hence improving their cash cycle.

One reason why firms may seek to have large inventories is because they want to have a certain amount of slack to face up to future demand growth. However, our results suggest that this kind of slack is best kept in the form of cash. Cash is a fungible resource, it is versatile (Nason and Wiklund 2018), and it can be redeployed across different uses. In contrast, inventory is not a fungible resource, and is difficult to redeploy into different uses. Our results suggest that HGFs are ideally lean (in terms of inventory), although they may be “plump” in terms of cash holdings.<sup>20</sup> Despite having low levels of inventory, HGFs can boost their readiness for growth by investing in capital assets and employees.

Furthermore, firms often overestimate the efficiency gains of large batch production, and underestimate the gains from flexibility from small batch production under “single-piece flow” (Ries 2011).<sup>21</sup> Having a small batch production process gives flexibility in production, lowers the costs of producing and storing inventory, and increases the ability to detect production errors and to redesign products to better address consumer

<sup>20</sup> Table 2 shows that “cash in bank” is positive and significant in model 1 (i.e., for firms with 10+ employees) for Croatia.

<sup>21</sup> Ries (2011, p. 184) gives the example of folding newsletters, sealing them into envelopes, and attaching a stamp. The standard approach might be to begin by folding all newsletters, then afterward putting them all into envelopes. However, this approach has drawbacks relating to time taken to sort, stack, and move around large piles of half-complete envelopes. Also it is possible that the letters do not fit in the envelopes, a problem which would only be discovered late into the production process. Instead, “single-piece flow” (see also “continuous flow manufacturing”), which corresponds to completing each envelope one at a time, is a surprisingly efficient production method, and the superiority of “single-piece flow” has been confirmed by studies (Ries, 2011, p. 184).

needs. Scale economies may come to HGFs from investing in capital infrastructure and employees, rather than from producing a large inventory.

Although our evidence on the importance of low inventories for HGFs is not causal, but based on associations, nevertheless it signals a relatively neglected area that would benefit from further research.

### 6.2.2 Growth in employees

Our results have shown, in a clear and robust way, that the employment growth rate from  $t-2$  to  $t$  is positively associated with HGF status (a dummy variable) for the period  $t:t+3$ . We interpret this in terms of firms preparing for periods of high growth (via investing in employees) to have the necessary human resources to carry out their growth projects. Relatedly, growth of assets is associated with subsequent HGF status (for the employment HGF indicator in both Croatia and Slovenia), which we also interpret as evidence of the need for firms to prepare for rapid growth by investing proactively. Employment and physical assets are converted into sales growth, with a lag. Penrose (1959) explains how a critical part of the growth process involves taking the time to train up new employees before executing a firm's growth plans. Firms should proactively invest in employment growth before embarking on an ambitious growth trajectory, because these employees will need to build up their firm-specific skills and knowledge before they can start to effectively implement the growth plans (Coad and Guenther 2014).

Our results therefore suggest that employment growth is positively associated with subsequent HGF status. Previous research, however, has generally observed that high-growth status in one period does not improve the probability of high-growth status in the following period, but rather that HGF status in subsequent periods is roughly statistically independent (Holzl, 2014; Daunfeldt et al. 2014; Daunfeldt and Halvarsson 2015). Nevertheless, caution is required because our results are not closely comparable. It is plausible that the different results are due to differences in the measurement of growth.<sup>22</sup> Here, we find that growth rate ( $t-2:t$ ) is positively associated to HGF status (dummy variable for growth over  $t:t+3$ ). This is a different

specification from that used in other studies, because our focus is on HGF prediction more generally, and not just on the autocorrelation of growth.

### 6.2.3 Short-term liabilities

A robust finding from our analysis is that “short-term liabilities” is an important predictor of HGF status. Our interpretation is that firms with access to short-term liabilities have more resources available than those without. This availability of access to short-term liabilities could furnish firms with a little more financial security in order to carry out their ambitious growth projects. This could help HGFs to grow without disrupting their cash cycle—the balance of costs and revenues.

An alternative interpretation could be that future HGFs are more desperate to seek financing, and that they make more efforts to seek finance, even if they can only obtain short-term finance rather than longer-term financing. We remain unsure about the causal direction, and recommend that future research could better identify the role of short-term liabilities as a contributing factor for rapid growth.

If indeed the availability of short-term liabilities does have a causal effect on the likelihood of rapid growth, the implications could be that government could support HGFs by facilitating access to short-term loan facilities. This could be especially useful for HGFs, while take-up among non-HGFs could be lower. A size disaggregation analysis (not shown here, available from the authors) shows that the coefficient on short-term liabilities is positive for micro firms (3–9 employees), and zero or negative for larger firms (10–19 employees, and 20+ employees, respectively). To the extent that micro firms are buffeted about by volatile cash flow streams, that may even threaten their survival, then short-term liabilities can provide micro firms a lifeline during short-lived cash flow crises.

### 6.2.4 Internal finance and external finance

Previous research has shown interest in the relationship between financial performance and firm growth (Cowling 2004; Cho and Pucic 2005; Davidsson et al. 2009; Delmar et al. 2013; Coad et al. 2017). Cowling (2004) observes that growth and profits move in parallel. Davidsson et al. (2009) investigate whether SMEs that grow become more profitable, or whether SMEs that are profitable are more likely to grow. They observe that profits precede growth, rather than vice

<sup>22</sup> One possibility could be that the effects of previous growth rate on subsequent HGF status are nonlinear across the distribution of previous growth rates.

versa. Coad et al. (2017) obtain causal estimates that while sales growth leads to profits growth in the overall sample, nevertheless in the subsample of high-growth firms, it is the growth of profits that drives sales growth. Possible explanations for this are that, on the one hand, profits are reinvested into the growth projects of cash-starved firms, and on the other hand that profits act as a signal of credibility to stakeholders and providers of external finance.

Our LASSO algorithm finds profits to be important for predicting high-growth episodes, in the case of the employment growth HGF indicator, although not for the sales HGF indicator. This offers partial support for the role of profits on subsequent chances of rapid growth. Relatedly, a firm's reserves are negatively related to HGF status (again, for the employment HGF indicator only), which suggests that while profits have a beneficial role on growth probabilities, nevertheless it is important to reinvest these profits into growth projects rather than storing the profits as reserves.

An interesting finding is that variables relating to the use of external finance (such as longer-term bank loans) do not appear prominently as predictors of HGF status, while internal financial resources (i.e., variables relating to cash and profits) are positively related to HGF status. Whatever the reasons may be (imperfect capital markets or firms' low demand for borrowing), it seems that Croatian HGFs tend to finance their growth using their internal financial resources.

## 7 Conclusion

Previous research has had only a modest success in predicting high-growth firms. Reasons for this could be that previous research has applied a restrictive set of explanatory variables, and in particular has not included variables with the statistical properties that are congruent with those of firm growth: i.e., there remains a pressing need to include explanatory variables with a high amount of variation within firms over time. To address this, we explore whether big data techniques (i.e., LASSO) applied to comprehensive datasets with hundreds of explanatory variables (many of which have high within-variance) can be useful for HGF prediction. Pseudo- $R^2$  statistics of around 10% suggest that the prediction of HGFs remains a challenge. Machine learning is therefore no panacea for predicting HGFs, even with variables that vary over time.

Similar results are found for Croatia and Slovenia, suggesting that our post hoc discussion of the observed results is not simply an exercise in over-theorizing about spurious random sampling error, but rather that our results are robust.

Our LASSO analysis suggests that HGFs are already performing well, in terms of (growth of) exports, sales, assets, and employment, in the period before the high-growth episode (both in period  $t$ , but also growing from  $t-2$  to  $t$ ), HGFs tend to rely on profits to finance their growth, rather than external finance. HGFs are on average younger firms, and are less likely to be from high-tech manufacturing sectors (in line with Henrekson and Johansson 2010). An increase in inventories is associated with a lower probability of becoming a HGF. Finally, HGFs have more intangible assets.

Firms that are well prepared for growth are firms with high profits, high investment, and low reserves (presumably due to high rates of reinvesting their profits), and also low inventories (hence, operating according to "lean" principles to boost efficiency and to reduce waste). Internal finance variables seem to be a stronger predictor of HGF status than external finance variables. Exports especially beneficial for micro firms seeking to grow. Investment in fixed assets also helps improve chances of rapid growth.

Analysis of the accuracy and sensitivity statistics confirms an intuition made by Shane (2009, p141)—that although it is rather easy to predict which firms will certainly not become HGFs, nevertheless the error rates are higher when it comes to predicting which firms are HGFs.

Our LASSO procedure was operated in a semi-supervised way, and was not fully automated. Indeed, the raw output of our calculations is not knowledge, nor information, but rather data. LASSO output is a raw material that still requires much effort for interpretation, and to distinguish the theoretically interesting significant results from the relatively unimportant significant results. AI and machine learning are tools to augment human decision-making, rather than autonomous robots that can replace human decision-making (Brynjolfsson and McAfee 2014).

Broadly speaking, we expect that big data techniques (such as LASSO) will become more widely used in entrepreneurship research in future. But will machines ever be able to accurately predict HGFs? We expect that improved methods will enhance our predictive power, but that there will always remain a large amount of chaos, surprises, and unpredictability.



**Acknowledgments** We are grateful to Martin Spindler (maintainer of the HDM package in R) for advice on the software and to Iris Loncar, accounting professor, for discussions on accounting practice and the composition of particular variables. Thanks also go to Barbara Zitek for translating the accounting variables from

Slovenian to English, and to Margherita Bacigalupo for introducing the authors of this manuscript to each other, and to Ivan Zilic for helpful comments on machine learning. Three anonymous reviewers provided many helpful comments. Any remaining errors are ours alone.

## Appendix 1. LASSO results for the Croatian sample

**Table 4** Logit model 1, employment indicator

	Test sample					Full sample				
	Coeff.	Effect (in %)	Std. Error	<i>t</i> stat	<i>p</i> value	Coeff.	Effect (in %)	Std. Error	<i>t</i> stat	<i>p</i> value
Intangible assets	0.00149	0.14888	0.00024	6.09176	0.00000	0.00104	0.10361	0.00013	7.71436	0.00000
Buildings	-0.00061	-0.06072	0.00020	-3.01543	0.00257	-0.00039	-0.03855	0.00011	-3.50989	0.00045
Raw materials and supplies	-0.00070	-0.07005	0.00024	-2.92562	0.00344	-0.00061	-0.06095	0.00013	-4.66669	0.00000
Cash in bank and cash on hand	0.00075	0.07478	0.00068	1.10124	0.27080	0.00047	0.04736	0.00037	1.29280	0.19609
Reserves from retained earnings	-0.00088	-0.08801	0.00030	-2.92904	0.00340	-0.00111	-0.11106	0.00017	-6.63006	0.00000
Sales	0.00607	0.60736	0.00138	4.41662	0.00001	0.00493	0.49297	0.00076	6.46006	0.00000
Profit of the year	0.00169	0.16897	0.00032	5.21193	0.00000	0.00182	0.18165	0.00018	9.93443	0.00000
Growth in cash in bank and cash on hand	0.00180	0.17998	0.00059	3.03873	0.00238	0.00104	0.10399	0.00034	3.09575	0.00196
Growth in assets	0.00378	0.37768	0.00131	2.88258	0.00395	0.00469	0.46919	0.00072	6.52159	0.00000
Growth in exports	0.00090	0.09028	0.00027	3.38644	0.00071	0.00072	0.07219	0.00015	4.87819	0.00000
Growth in number of employees	0.00744	0.74388	0.00177	4.21216	0.00003	0.00837	0.83708	0.00093	8.98041	0.00000
Log age	-0.02196	-2.19647	0.00213	-10.28948	0.00000	-0.01995	-1.99513	0.00118	-16.85758	0.00000
Small firm	0.06310	6.30969	0.01180	5.34751	0.00000	0.04740	4.74049	0.00587	8.07673	0.00000
Medium firm	0.04767	4.76653	0.01154	4.12871	0.00004	0.03487	3.48749	0.00572	6.09273	0.00000
Capital region	0.00380	0.38045	0.00276	1.37647	0.16869	0.00712	0.71211	0.00150	4.74476	0.00000
Mid-high-tech manufacturing	0.01999	1.99926	0.01275	1.56835	0.11681	0.02022	2.02199	0.00708	2.85460	0.00431
Mid-low-tech manufacturing	0.00781	0.78123	0.01349	0.57915	0.56250	0.01337	1.33747	0.00746	1.79224	0.07310
Low-tech manufacturing	0.01595	1.59482	0.01287	1.23881	0.21543	0.01451	1.45138	0.00718	2.02081	0.04330
High-tech KIS	0.02279	2.27883	0.01280	1.78014	0.07507	0.02750	2.74969	0.00708	3.88555	0.00010
Other KIS	0.02312	2.31190	0.01212	1.90685	0.05655	0.02330	2.32964	0.00675	3.45240	0.00056
Less KIS	0.01830	1.83030	0.01215	1.50648	0.13196	0.02050	2.05006	0.00676	3.03487	0.00241
Number of observations	23,733					79,109				
McFadden $R^2$	0.096					0.092				
Dependent variable mean	0.043					0.043				
Accuracy (in %)	76.48					75.88				
Sensitivity (in %)	56.90					57.10				
Specificity (in %)	77.34					76.73				

Note: Year dummies included. LASSO was trained on the train sample which included 70% of the full sample. Variables selected in the train sample were used on the test sample which includes the other 30% of the full sample. Threshold of fitted probabilities to calculate accuracy, sensitivity, and specificity rates is 0.05; the formulas are given in the *Algorithm* under the section *Method*

**Table 5** Logit model 1, turnover indicator

	Test sample					Full sample				
	Coeff.	Effect (in %)	Std. Error	<i>t</i> stat	<i>p</i> value	Coeff.	Effect (in %)	Std. Error	<i>t</i> stat	<i>p</i> value
Tools, transportation equipment, and vehicle	0.00172	0.17224	0.00044	3.87705	0.00011	0.00176	0.17606	0.00025	7.10417	0.00000
Merchandise goods (at inventory)	-0.00124	-0.12352	0.00037	-3.29537	0.00098	-0.00126	-0.12573	0.00020	-6.17034	0.00000
Subsidies, donations, and compensations	0.00110	0.11005	0.00049	2.26838	0.02332	0.00138	0.13817	0.00027	5.14246	0.00000
Short-term liabilities	0.01181	1.18107	0.00174	6.79200	0.00000	0.01267	1.26742	0.00095	13.34335	0.00000
Sales	-0.03180	-3.17991	0.00211	-15.04354	0.00000	-0.03117	-3.11719	0.00116	-26.78622	0.00000
Other expenses	0.00157	0.15711	0.00034	4.59314	0.00000	0.00170	0.17009	0.00019	8.99866	0.00000
Cost of goods sold	-0.00108	-0.10823	0.00040	-2.68654	0.00722	-0.00069	-0.06879	0.00022	-3.11050	0.00187
Exports	0.00155	0.15538	0.00031	5.02606	0.00000	0.00142	0.14214	0.00017	8.38779	0.00000
Growth in intangible assets	0.00057	0.05714	0.00043	1.33785	0.18096	0.00109	0.10928	0.00023	4.69149	0.00000
Growth in inventories	0.00218	0.21752	0.00049	4.40939	0.00001	0.00144	0.14427	0.00027	5.31194	0.00000
Growth in assets	0.00721	0.72086	0.00202	3.56852	0.00036	0.00813	0.81302	0.00112	7.26589	0.00000
Growth in number of employees	0.01657	1.65658	0.00238	6.96896	0.00000	0.01571	1.57134	0.00131	12.02439	0.00000
Log age	-0.02494	-2.49403	0.00310	-8.04718	0.00000	-0.02205	-2.20469	0.00172	-12.84907	0.00000
Small firm	0.00251	0.25067	0.01348	0.18591	0.85252	0.01729	1.72909	0.00757	2.28407	0.02237
Medium firm	0.00501	0.50115	0.01316	0.38089	0.70329	0.01773	1.77322	0.00742	2.39097	0.01681
Capital region	0.01003	1.00334	0.00374	2.68126	0.00734	0.01143	1.14266	0.00206	5.55156	0.00000
Mid-high-tech manufacturing	0.04368	4.36788	0.01645	2.65499	0.00794	0.04623	4.62310	0.00917	5.04213	0.00000
Mid-low-tech manufacturing	-0.00294	-0.29399	0.01805	-0.16292	0.87059	0.02079	2.07859	0.00984	2.11236	0.03466
Low-tech manufacturing	0.03685	3.68520	0.01663	2.21570	0.02672	0.02430	2.43043	0.00938	2.59172	0.00955
High-tech KIS	0.06223	6.22323	0.01650	3.77150	0.00016	0.05780	5.77978	0.00922	6.26701	0.00000
Other KIS	0.04053	4.05253	0.01571	2.57930	0.00991	0.04253	4.25317	0.00878	4.84579	0.00000
Less KIS	0.04997	4.99730	0.01575	3.17279	0.00151	0.05275	5.27508	0.00880	5.99566	0.00000
Number of observations	23,733					79,109				
McFadden $R^2$	0.090					0.093				
Dependent variable mean	0.090					0.090				
Accuracy (in %)	72.41					72.86				
Sensitivity (in %)	57.75					58.22				
Specificity (in %)	73.86					74.30				

Note: Year dummies included. LASSO was trained on the train sample which included 70% of the full sample. Variables selected in the train sample were used on the test sample which includes the other 30% of the full sample. Threshold of fitted probabilities to calculate accuracy, sensitivity, and specificity rates is 0.10; the formulas are given in the *Algorithm* under the section *Method*

**Table 6** Model 2, employment indicator

	Test sample					Full sample				
	Coeff.	Effect (in %)	Std. Error	<i>t</i> stat	<i>p</i> value	Coeff.	Effect (in %)	Std. Error	<i>t</i> stat	<i>p</i> value
Intangible assets	0.00045	0.04540	0.00013	3.57504	0.00035	0.00066	0.06589	0.00007	9.31097	0.00000
Inventories	-0.00092	-0.09224	0.00012	-7.89514	0.00000	-0.00078	-0.07826	0.00007	-11.76446	0.00000
Reserves from retained earnings	-0.00023	-0.02271	0.00016	-1.39225	0.16385	-0.00050	-0.04975	0.00009	-5.30721	0.00000
Short-term liabilities	0.00224	0.22403	0.00066	3.38576	0.00071	0.00290	0.28982	0.00037	7.78904	0.00000
Financial expenses	0.00044	0.04405	0.00017	2.54079	0.01106	0.00037	0.03737	0.00010	3.86446	0.00011
Sales	0.00071	0.07124	0.00019	3.67204	0.00024	0.00408	0.40826	0.00048	8.45447	0.00000
Profit of the year	0.00528	0.52768	0.00087	6.09370	0.00000	0.00097	0.09744	0.00011	8.94241	0.00000
Exports	0.00023	0.02307	0.00013	1.84491	0.06506	0.00017	0.01745	0.00007	2.47887	0.01318
Growth in assets	0.00379	0.37935	0.00061	6.24430	0.00000	0.00414	0.41426	0.00033	12.64907	0.00000
Growth in profit of the year	0.00034	0.03380	0.00016	2.14366	0.03206	0.00031	0.03118	0.00009	3.52239	0.00043
Growth in exports	0.00021	0.02125	0.00015	1.40021	0.16145	0.00036	0.03581	0.00009	4.18249	0.00003
Growth in number of employees	0.00805	0.80543	0.00089	9.03273	0.00000	0.00659	0.65884	0.00051	12.93915	0.00000
Log age	-0.01235	-1.23481	0.00105	-11.77337	0.00000	-0.01418	-1.41796	0.00059	-24.18856	0.00000
Micro firm	0.04775	4.77462	0.00819	5.82789	0.00000	0.03672	3.67207	0.00400	9.18372	0.00000
Small firm	0.05797	5.79654	0.00790	7.34136	0.00000	0.04833	4.83292	0.00381	12.69091	0.00000
Medium firm	0.03776	3.77574	0.00793	4.75895	0.00000	0.03176	3.17584	0.00381	8.33435	0.00000
Capital region	0.00357	0.35712	0.00131	2.72120	0.00651	0.00315	0.31478	0.00074	4.28020	0.00002
Mid-high-tech manufacturing	0.01846	1.84611	0.00754	2.44730	0.01440	0.01839	1.83864	0.00410	4.48609	0.00001
Mid-low-tech manufacturing	0.00790	0.79049	0.00796	0.99368	0.32038	0.01282	1.28232	0.00427	3.00209	0.00268
Low-tech manufacturing	0.01333	1.33322	0.00771	1.72899	0.08382	0.01576	1.57569	0.00416	3.78533	0.00015
High-tech KIS	0.01681	1.68147	0.00749	2.24627	0.02469	0.02076	2.07644	0.00405	5.13176	0.00000
Other KIS	0.01840	1.84019	0.00722	2.54988	0.01078	0.01732	1.73155	0.00391	4.43092	0.00001
Less KIS	0.01883	1.88262	0.00726	2.59454	0.00947	0.01951	1.95070	0.00392	4.97030	0.00000
Number of observations	63,831					212,769				
McFadden $R^2$	0.108					0.101				
Dependent variable mean	0.028					0.028				
Accuracy (in %)	86.14					86.55				
Sensitivity (in %)	45.44					45.32				
Specificity (in %)	87.35					87.75				

Note: Year dummies included. LASSO was trained on the train sample which included 70% of the full sample. Variables selected in the train sample were used on the test sample which includes the other 30% of the full sample. Threshold of fitted probabilities to calculate accuracy, sensitivity, and specificity rates is 0.05; the formulas are given in the *Algorithm* under the section *Method*

**Table 7** Model 2, turnover indicator

	Test sample					Full sample				
	Coeff.	Effect (in %)	Std. Error	<i>t</i> stat	<i>p</i> value	Coeff.	Effect (in %)	Std. Error	<i>t</i> stat	<i>p</i> value
Fixed assets	0.00063	0.06339	0.00041	1.54943	0.12128	0.00074	0.07432	0.00023	3.26366	0.00110
Intangible assets	0.00010	0.00980	0.00016	0.60859	0.54280	0.00036	0.03642	0.00009	4.11156	0.00004
Merchandise goods (at inventory)	-0.00138	-0.13801	0.00013	-10.53613	0.00000	-0.00134	-0.13359	0.00007	-18.53208	0.00000
Subsidies, donations, and compensations	0.00032	0.03201	0.00021	1.54550	0.12223	0.00050	0.04983	0.00012	4.30303	0.00002
Long-term liabilities	0.00050	0.04965	0.00014	3.60895	0.00031	0.00036	0.03584	0.00008	4.73369	0.00000
Liabilities towards group companies	0.00096	0.09554	0.00032	2.99490	0.00275	0.00079	0.07897	0.00018	4.48096	0.00001
Short-term liabilities	0.00292	0.29200	0.00071	4.11675	0.00004	0.00242	0.24239	0.00039	6.19291	0.00000
Other expenses	0.00059	0.05945	0.00015	3.87847	0.00011	0.00064	0.06434	0.00008	7.58041	0.00000
Exports	0.00033	0.03345	0.00014	2.41108	0.01591	0.00041	0.04053	0.00008	5.30572	0.00000
Growth in inventories	0.00077	0.07672	0.00021	3.61104	0.00031	0.00062	0.06195	0.00012	5.21261	0.00000
Growth in assets	0.00504	0.50391	0.00082	6.16657	0.00000	0.00581	0.58104	0.00046	12.54537	0.00000
Growth in profit of the year	0.00075	0.07497	0.00016	4.61618	0.00000	0.00056	0.05558	0.00009	6.26277	0.00000
Growth in number of employees	0.00863	0.86322	0.00112	7.71617	0.00000	0.00908	0.90785	0.00063	14.49462	0.00000
Log age	-0.01532	-1.53235	0.00136	-11.26685	0.00000	-0.01612	-1.61203	0.00075	-21.47673	0.00000
Micro firm	0.00049	0.04872	0.00819	0.05946	0.95258	-0.01085	-1.08452	0.00408	-2.65504	0.00793
Small firm	0.06016	6.01642	0.00785	7.66872	0.00000	0.04985	4.98540	0.00387	12.88661	0.00000
Medium firm	0.03741	3.74068	0.00789	4.73963	0.00000	0.02938	2.93817	0.00389	7.54554	0.00000
Capital region	0.00454	0.45417	0.00166	2.73099	0.00632	0.00425	0.42470	0.00092	4.61970	0.00000
Mid-high-tech manufacturing	0.02335	2.33512	0.00814	2.86954	0.00411	0.03101	3.10120	0.00453	6.84484	0.00000
Mid-low-tech manufacturing	0.02105	2.10469	0.00849	2.47906	0.01318	0.01679	1.67897	0.00480	3.49474	0.00047
Low-tech manufacturing	0.01847	1.84722	0.00827	2.23270	0.02557	0.02090	2.09038	0.00463	4.51930	0.00001
High-tech KIS	0.03238	3.23843	0.00803	4.03521	0.00005	0.03354	3.35362	0.00451	7.44213	0.00000
Other KIS	0.02244	2.24384	0.00773	2.90338	0.00369	0.02525	2.52544	0.00434	5.82239	0.00000
Less KIS	0.03019	3.01868	0.00776	3.89146	0.00010	0.03192	3.19197	0.00435	7.33155	0.00000
Number of observations	63,831					212,769				
McFadden $R^2$	0.133					0.136				
Dependent variable mean	0.046					0.046				
Accuracy (in %)	88.78					88.36				
Sensitivity (in %)	38.14					41.03				
Specificity (in %)	91.13					90.63				

Note: Year dummies included. LASSO was trained on the train sample which included 70% of the full sample. Variables selected in the train sample were used on the test sample which includes the other 30% of the full sample. Threshold of fitted probabilities to calculate accuracy, sensitivity, and specificity rates is 0.10; the formulas are given in the *Algorithm* under the section *Method*

## Appendix 2. LASSO results for the Slovenian sample

**Table 8** Model 2, employment indicator, Slovenia

	Test sample					Full sample				
	Coeff.	Effect (in %)	Std. Error	<i>t</i> stat	<i>p</i> value	Coeff.	Effect (in %)	Std. Error	<i>t</i> stat	<i>p</i> value
Inventories	-0.00057	-0.05693	0.00015	-3.89664	0.00010	-0.00076	-0.07637	0.00008	-9.25379	0.00000
Capital reserves	-0.00029	-0.02861	0.00015	-1.97345	0.04847	-0.00030	-0.02955	0.00009	-3.40629	0.00066
Short-term liabilities	0.00408	0.40815	0.00062	6.54174	0.00000	0.00395	0.39483	0.00035	11.41020	0.00000
Sales	0.00038	0.03838	0.00013	2.96443	0.00304	0.00043	0.04299	0.00007	5.86775	0.00000
Profit	0.00050	0.05015	0.00022	2.24697	0.02466	0.00023	0.02253	0.00010	2.16088	0.03071
Subsidies, donations, and compensations	0.00046	0.04613	0.00016	2.86041	0.00424	0.00050	0.05040	0.00009	5.31583	0.00000
Amortization	-0.00108	-0.10834	0.00031	-3.52184	0.00043	-0.00139	-0.13909	0.00017	-8.15877	0.00000
Growth in assets	-0.00019	-0.01905	0.00037	-0.51241	0.60837	0.00009	0.00861	0.00019	0.45922	0.64608
Growth in production machinery and equipment	0.00045	0.04453	0.00023	1.96875	0.04901	0.00037	0.03725	0.00012	3.21555	0.00130
Growth in other long-term operating liabilities	-0.00034	-0.03428	0.00016	-2.11494	0.03446	0.00025	0.02496	0.00010	2.49052	0.01276
Growth in sales outside EU	0.00043	0.04331	0.00019	2.30041	0.02145	0.00029	0.02905	0.00009	3.08364	0.00205
Growth in other material costs	0.00026	0.02592	0.00032	0.82016	0.41215	0.00038	0.03817	0.00017	2.25856	0.02392
Growth in other expenses	0.00031	0.03093	0.00019	1.63512	0.10205	0.00041	0.04138	0.00011	3.74508	0.00018
Growth in number of employees	0.00379	0.37900	0.00118	3.21855	0.00129	0.00292	0.29248	0.00065	4.52516	0.00001
Micro firm	0.00088	0.08842	0.00524	0.16868	0.86605	0.00303	0.30303	0.00387	0.78356	0.43331
Small firm	0.00384	0.38354	0.00657	0.58394	0.55927	0.00919	0.91883	0.00690	1.33072	0.18329
Medium firm	-0.00370	-0.37025	0.00321	-1.15218	0.24927	-0.00226	-0.22555	0.00342	-0.65898	0.50991
Mid-high-tech manufacturing	-0.00508	-0.50826	0.00214	-2.37229	0.01770	-0.00341	-0.34123	0.00195	-1.74937	0.08024
High-tech KIS	-0.00697	-0.69721	0.00161	-4.33873	0.00001	-0.00609	-0.60903	0.00128	-4.75014	0.00000
Other KIS	-0.00987	-0.98671	0.00280	-3.52625	0.00042	-0.00897	-0.89729	0.00176	-5.09268	0.00000
Less KIS	-0.01231	-1.23102	0.00521	-2.36300	0.01815	-0.00980	-0.98040	0.00297	-3.30481	0.00095
Mid-low-tech manufacturing	-0.00575	-0.57495	0.00215	-2.67452	0.00750	-0.00326	-0.32604	0.00198	-1.64869	0.09922
Low-tech manufacturing	-0.00571	-0.57123	0.00210	-2.71792	0.00658	-0.00583	-0.58304	0.00142	-4.11137	0.00004
Capital region	-0.00759	-0.75930	0.00147	-5.18101	0.00000	-0.00361	-0.36061	0.00139	-2.60368	0.00923
Number of observations	9930					33,101				
McFadden $R^2$	0.130					0.140				
Dependent variable mean	0.015					0.015				
Accuracy (in %)	94.79					94.47				
Sensitivity (in %)	24.82					32.94				
Specificity (in %)	95.80					95.43				

Note: Year dummies included. LASSO was trained on the train sample which included 70% of the full sample. Variables selected in the train sample were used on the test sample which includes the other 30% of the full sample. Threshold of fitted probabilities to calculate accuracy, sensitivity, and specificity rates is 0.05; the formulas are given in the *Algorithm* under the section *Method*

**Table 9** Model 2, turnover indicator, Slovenia

	Test sample					Full sample				
	Coeff.	Effect (in %)	Std. Error	<i>t</i> stat	<i>p</i> value	Coeff.	Effect (in %)	Std. Error	<i>t</i> stat	<i>p</i> value
Merchandise goods (at inventories)	-0.00062	-0.06231	0.00017	-3.61574	0.00030	-0.00074	-0.07404	0.00010	-7.35889	0.00000
Capital reserves	-0.00050	-0.04959	0.00018	-2.77171	0.00559	-0.00053	-0.05326	0.00010	-5.18578	0.00000
Long-term accrued costs and deferred revenues	0.00067	0.06722	0.00023	2.93852	0.00331	0.00047	0.04740	0.00013	3.64140	0.00027
Exports	0.00096	0.09604	0.00019	5.15550	0.00000	0.00086	0.08641	0.00010	8.28455	0.00000
Subsidies, donations, and compensations	0.00019	0.01901	0.00021	0.90055	0.36785	0.00034	0.03432	0.00012	2.87218	0.00408
Short-term liabilities	0.00327	0.32745	0.00095	3.44958	0.00056	0.00255	0.25462	0.00054	4.68473	0.00000
Costs of services	-0.00026	-0.02640	0.00087	-0.30308	0.76183	0.00157	0.15739	0.00049	3.24152	0.00119
Growth in fixed assets	0.00125	0.12467	0.00049	2.54162	0.01105	0.00034	0.03408	0.00028	1.20931	0.22655
Growth in number of employees	0.00467	0.46658	0.00158	2.96029	0.00308	0.00553	0.55266	0.00087	6.36558	0.00000
Micro firm	-0.03131	-3.13073	0.01671	-1.87338	0.06105	-0.01010	-1.00996	0.00693	-1.45820	0.14479
Small firm	-0.00040	-0.03980	0.00651	-0.06111	0.95127	0.01309	1.30921	0.00753	1.73956	0.08195
Medium firm	-0.00626	-0.62642	0.00383	-1.63555	0.10197	0.00090	0.09044	0.00527	0.17150	0.86383
Mid-high-tech manufacturing	-0.00356	-0.35631	0.00499	-0.71475	0.47478	-0.00141	-0.14073	0.00326	-0.43235	0.66549
High-tech KIS	-0.00631	-0.63142	0.00403	-1.56873	0.11674	-0.00612	-0.61204	0.00234	-2.62086	0.00877
Other KIS	-0.00742	-0.74240	0.00484	-1.53544	0.12471	-0.00735	-0.73528	0.00265	-2.77333	0.00555
Less KIS	-0.00695	-0.69482	0.00628	-1.10577	0.26885	-0.00930	-0.93000	0.00342	-2.71672	0.00660
Mid-low-tech manufacturing	-0.00378	-0.37755	0.00488	-0.77421	0.43882	-0.00268	-0.26781	0.00288	-0.92882	0.35299
Low-tech manufacturing	-0.00731	-0.73112	0.00365	-2.00325	0.04518	-0.00767	-0.76736	0.00205	-3.74146	0.00018
Capital region	-0.00510	-0.50984	0.00286	-1.78264	0.07468	-0.00536	-0.53565	0.00173	-3.09942	0.00194
Number of observations	9930					33,101				
Dependent variable mean	0.026					0.026				
McFadden $R^2$	0.155					0.157				
Accuracy (in %)	87.67					86.95				
Sensitivity (in %)	51.27					53.29				
Specificity (in %)	88.56					87.84				

Note: Year dummies included. LASSO was trained on the train sample which included 70% of the full sample. Variables selected in the train sample were used on the test sample which includes the other 30% of the full sample. Threshold of fitted probabilities to calculate accuracy, sensitivity, and specificity rates is 0.05; the formulas are given in the *Algorithm* under the section *Method*

## References

- Achtenhagen, L., Naldi, L., & Melin, L. (2010). "Business growth"—Do practitioners and scholars really talk about the same thing? *Entrepreneurship Theory and Practice*, 34(2), 289–316.
- Arrighetti, A., & Lasagni, A. (2013). Assessing the determinants of high-growth manufacturing firms in Italy. *International Journal of the Economics of Business*, 20(2), 245–267. <https://doi.org/10.1080/13571516.2013.783456>.
- Audretsch, D. B., Santarelli, E., & Vivarelli, M. (1999). Start-up size and industrial dynamics: Some evidence from Italian manufacturing. *International Journal of Industrial Organization*, 17, 965–983.
- Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6), 2369–2429.
- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2), 29–50.
- Belloni, A., Chernozhukov, V., & Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34(4), 606–619.
- Bernerth, J. B., Cole, M. S., Taylor, E. C., & Walker, H. J. (2018). Control variables in leadership research: A qualitative and quantitative review. *Journal of Management*, 44(1), 131–160.
- Bianchini, S., Bottazzi, G., & Tamagni, F. (2017). What does (not) characterize persistent corporate high-growth? *Small Business Economics*, 48(3), 633–656.
- Birch, D. L. (1979). *The job generation process*. Cambridge, MA: MIT program on neighborhood and regional change, Massachusetts Institute of Technology.
- Bjuggren, C.-M., Daunfeldt, S.-O., & Johansson, D. (2013). High-growth firms and family ownership. *Journal of Small Business & Entrepreneurship*, 26(4), 365–385. <https://doi.org/10.1080/08276331.2013.821765>.
- Brown, R., & Mawson, S. (2013). Trigger points and high-growth firms. *Journal of Small Business and Enterprise Development*, 20(2), 279–295.
- Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.
- Chernozhukov, V., Hansen, C., & Spindler, M. (2016). High-dimensional metrics in R. arXiv preprint arXiv:1603.01700.
- Cho, H. J., & Pucik, V. (2005). Relationship between innovativeness, quality, growth, profitability, and market value. *Strategic Management Journal*, 26(6), 555–575.
- Churchill, N. C., & Mullins, J. W. (2001). How fast can your company afford to grow? *Harvard Business Review*, 79(5), 135–143.
- Coad A., (2009). *The growth of firms: A survey of theories and empirical evidence*. Edward Elgar, Cheltenham, UK and Northampton, MA, USA.
- Coad, A., Cowling, M., & Siepel, J. (2017). Growth processes of high-growth firms as a four-dimensional chicken and egg. *Industrial and Corporate Change*, 26(4), 537–554.
- Coad A., Frankish J.S., Roberts R.G., Storey D.J., (2015). Are firm growth paths random? A reply to "firm growth and the illusion of randomness." *Journal of Business Venturing Insights* 3, 5–8.
- Coad, A., & Guenther, C. (2014). Processes of firm growth and diversification: Theory and evidence. *Small Business Economics*, 43, 857–871.
- Coad, A., & Planck, M. (2012). Firms as bundles of discrete resources—Towards an explanation of the exponential distribution of firm growth rates. *Eastern Economic Journal*, 38, 189–209.
- Coad, A., & Rao, R. (2011). The firm-level employment effects of innovations in high-tech US manufacturing industries. *Journal of Evolutionary Economics*, 21(2), 255–283.
- Coad A., Scott G., (2018). High-growth firms in Peru. *Cuadernos de Economia*, 37(75), 671–696.
- Cowling, M. (2004). The growth-profit nexus. *Small Business Economics*, 22(1), 1–9.
- Davidsson, P., Steffens, P., & Fitzsimmons, J. (2009). Growing profitable or growing from profits: Putting the horse in front of the cart? *Journal of Business Venturing*, 24(4), 388–406.
- Davidsson, P., & Wiklund, J. (2000). Conceptual and empirical challenges in the study of firm growth. In D. Sexton, & H. Landström (Eds.), *The Blackwell Handbook of Entrepreneurship* (reprinted 2006 in *Entrepreneurship and the Growth of Firms*, Elgar): 26–44. Oxford, MA: Blackwell Business.
- Daunfeldt, S.-O., Elert, N., & Johansson, D. (2014). The economic contribution of high-growth firms: Do policy implications depend on the choice of growth indicator? *Journal of Industry, Competition and Trade*, 14(3), 337–365.
- Daunfeldt S.-O., Halvarsson D., (2015). Are high-growth firms one-hit wonders? Evidence from Sweden. *Small Business Economics* 44, 361–383.
- Daunfeldt, S.-O., Elert, N., & Johansson, D. (2014). The economic contribution of high-growth firms: Do policy implications depend on the choice of growth indicator? *Journal of Industry, Competition and Trade* 14(3), 337–365.
- De Loecker, J. (2007). Do exports generate higher productivity? Evidence from Slovenia. *Journal of International Economics*, 73(1), 69–98.
- Delmar, F. (1997). Measuring growth: Methodological considerations and empirical results. In R. Donckels, & A. Miettinen (Eds.), *Entrepreneurship and SME research: On its way to the next millennium* (also reprinted 2006 in *Entrepreneurship and the Growth of Firms*, Elgar): 190–216. Aldershot, UK and Brookfield, VA: Ashgate.
- Delmar, F., Davidsson, P., & Gartner, W. B. (2003). Arriving at the high-growth firm. *Journal of Business Venturing*, 18, 189–216.
- Delmar, F., McKelvie, A., & Wennberg, K. (2013). Untangling the relationships among growth, profitability and survival in new firms. *Technovation*, 33(8–9), 276–291.
- Derbyshire, J., & Gamsey, E. (2014). Firm growth and the illusion of randomness. *Journal of Business Venturing Insights*, 1–2, 8–11.
- Denton, F. T. (1985). Data mining as an industry. *Review of Economics and Statistics*, 124–127.
- Evans, D. S. (1987). Tests of alternative theories of firm growth. *Journal of Political Economy*, 95(4), 657–674.

- Eurostat-OECD (2007). Eurostat-OECD Manual on Business Demography Statistics, Office for Official Publications of the European Communities, Luxembourg.
- Fan, L., Chen, S., Li, Q., & Zhu, Z. (2015). Variable selection and model prediction based on lasso, adaptive lasso and elastic net, 2015 4th International Conference on Computer Science and Network Technology (ICCSNT), Harbin, 2015, 579–583.
- George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy of Management Journal*, 57(2), 321–326.
- Geroski, P. A., Machin, S. J., & Walters, C. F. (1997). Corporate growth and profitability. *Journal of Industrial Economics*, 45(2), 171–189.
- Geroski, P. A. (2000). The growth of firms in theory and in practice. Pages 168–186 in Nicolai Foss and Volker Mahnke (eds): Competence, governance and entrepreneurship. Oxford University Press: Oxford, UK.
- Geroski, P., & Gugler, K. (2004). Corporate growth convergence in Europe. *Oxford Economic Papers*, 56, 597–620.
- Goedhuys, M., & Sleuwaegen, L. (2016). High-growth versus declining firms: The differential impact of human capital and R&D. *Applied Economics Letters*, 23(5), 369–372.
- Grover Goswami, A., Medvedev, D., & Olafsen, E. (2019). *High-growth firms: Facts, fiction, and policy options for emerging economies*. Washington, DC: World Bank.
- Guzman, J., Stern S. (2016). The state of American entrepreneurship: New estimates of the quantity and quality of entrepreneurship for 15 US states, 1988–2014. No. w22095. National Bureau of Economic Research.
- Hall, B. H. (1987). The relationship between firm size and firm growth in the US manufacturing sector. *Journal of Industrial Economics*, 35(4), 583–606.
- Hambrick, D. C. (2007). The field of management's devotion to theory: Too much of a good thing? *Academy of Management Journal*, 50(6), 1346–1352.
- 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. *Journal of Industrial Economics*, 46(4), 453–488.
- Helfat, C. E. (2007). Stylized facts, empirical research and theory development in management. *Strategic Organization*, 5(2), 185–192.
- Henrekson, M., & Johansson, D. (2010). Gazelles as job creators: A survey and interpretation of the evidence. *Small Business Economics*, 35, 227–244.
- Hollenbeck, J. R., & Wright, P. M. (2017). Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data. *Journal of Management*, 43(1), 5–18 <https://doi.org/10.1177/0149206316679487>.
- Hölzl, W. (2014). Persistence, survival, and growth: A closer look at 20 years of fast-growing firms in Austria. *Industrial and Corporate Change*, 23(1), 199–231.
- Ijiri, Y., & Simon, H. A. (1964). Business firm growth and size. *American Economic Review*, 54(2), 77–89.
- Ijiri, Y., & Simon, H. A. (1967). A model of business firm growth. *Econometrica*, 35(2), 348–355.
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196–217.
- Kumar, M. S. (1985). Growth, acquisition activity and firm size: Evidence from the United Kingdom. *Journal of Industrial Economics*, 33(3), 327–338.
- Lee, N. (2014). What holds back high-growth firms? Evidence from UK SMEs. *Small Business Economics*, 43(1), 183–195.
- Locke, E. A. (2007). The case for inductive theory building. *Journal of Management*, 33(6), 867–890.
- Lopez-Garcia, P., & Puente, S. (2012). What makes a high-growth firm? A dynamic probit analysis using Spanish firm-level data. *Small Business Economics*, 39, 1029–1041.
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS One*, 13(3), e0194889 <https://doi.org/10.1371/journal.pone.0194889>.
- Mason, C., & Brown, R. (2013). Creating good public policy to support high-growth firms. *Small Business Economics*, 40(2), 211–225.
- McKelvie, A., & Wiklund, J. (2010). Advancing firm growth research: A focus on growth mode instead of growth rate. *Entrepreneurship Theory and Practice*, 34(2), 261–288.
- McKenzie, D. (2017). Identifying and spurring high-growth entrepreneurship: Experimental evidence from a business plan competition. *American Economic Review*, 107(8), 2278–2307.
- McKenzie, D., & Sansone, D. (2017). Man vs. machine in predicting successful entrepreneurs: Evidence from a business plan competition in Nigeria. World Bank Policy Research Working Paper 8271.
- Megaravalli, A. V., & Sampagnaro, G. (2018). Predicting the growth of high-growth SMEs: Evidence from family business firms. *Journal of Family Business Management*. <https://doi.org/10.1108/JFBM-09-2017-0029>.
- Miyakawa, D., Miyauchi, Y., & Perez, C. (2017). Forecasting firm performance with machine learning: Evidence from Japanese firm-level data. Research Institute of Economy, Trade and Industry (RIETI).
- Moschella, D., Tamagni, F., & Yu, X. (2018). Persistent high-growth firms in China's manufacturing. *Small Business Economics*, in press. <https://doi.org/10.1007/s11187-017-9973-4>
- Nason, R. S., & Wiklund, J. (2018). An assessment of resource-based theorizing on firm growth and suggestions for the future. *Journal of Management*, 44(1), 32–60.
- NESTA. (2009). *The vital 6 per cent: How high growth innovative businesses generate prosperity and jobs*. London: NESTA.
- Neumark, D., Wall, B., & Zhang, J. (2011). Do small businesses create more jobs? New evidence for the United States from the National Establishment Time Series. *Review of Economics and Statistics*, 93(1), 16–29.
- Penrose, E.T., (1959). *The Theory of the Growth of the Firm*. Basil Blackwell: Oxford, UK.
- Pereira, V., & Temouri, Y. (2018). Impact of institutions on emerging European high-growth firms. *Management Decision*, 56(1), 175–187 <https://doi.org/10.1108/MD-03-2017-0279>.
- Peric, M., & Vitezic, V. (2016). Impact of global economic crisis on firm growth. *Small Business Economics*, 46(1), 1–12.
- Ries, E. (2011). *The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses*. Crown Books: New York.



- Sermpinis, G., Tsoukas, S., & Zhang, P. (2018). Modelling market implied ratings using LASSO variable selection techniques. *Journal of Empirical Finance*. Forthcoming.
- Shane, S. (2009). Why encouraging more people to become entrepreneurs is bad public policy. *Small Business Economics*, 33, 141–149.
- Shepherd, D., & Wiklund, J. (2009). Are we comparing apples with apples or apples with oranges? Appropriateness of knowledge accumulation across growth studies. *Entrepreneurship Theory and Practice*, 33(1), 105–123.
- Singh, A., & Whittington, G. (1975). The size and growth of firms. *Review of Economic Studies*, 42(1), 15–26.
- Srroj, S., Zupic, I., & Jaklič, M. (2018). Stylized facts about Slovenian high-growth firms. *Economic Research-Ekonomska Istraživanja*, 31(1), 1851–1879. <https://doi.org/10.1080/1331677X.2018.1516153>.
- Storey, D. J. (2011). Optimism and chance: The elephants in the entrepreneurship room. *International Small Business Journal*, 29(4), 303–321.
- Tian, S., Yu, Y., & Guo, H. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, 52, 89–100.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.
- Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3), 273–282.
- Tong, et al. (2016). Comparison of predictive modeling approaches for 30-day all-cause non-elective readmission risk. *BMC Medical Research Methodology*, 16, 26.
- Vancouver, J. B. (2018). In defense of HARKing. *Industrial and Organizational Psychology*, 11(1), 73–80.
- van Witteloostuijn, A., & Kolkman, D. (2019). Is firm growth random? A machine learning perspective. *Journal of Business Venturing Insights*, forthcoming. <https://doi.org/10.1016/j.jbvi.2018.e00107>
- Vitezić, V., Srroj, S., & Perić, M. (2018). Investigating industry dynamics in a recessionary transition economy. *South East European Journal of Economics and Business*, 13(1), 43–67.
- Weinblat, J. (2017). Forecasting European high-growth firms—A random forest approach. *Journal of Industry, Competition and Trade*, 1–42.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.