Predicting the past: A machine learning approach to detect innovative firms in times of crisis

Marco Guerzoni,[1,4] Massimiliano Nuccio[2,1], Consuelo R. Nava[3,1]

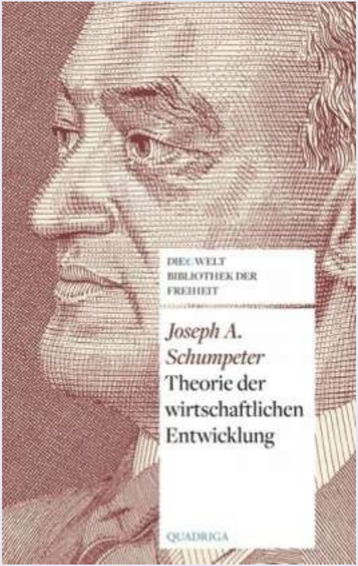
[1]Despina, Department of Economics and Statistics, University of Turin
[2]City REDI, University of Birmingham
[3]University of Aosta Valley
[4]ICRIOS, Bocconi

City REDI - University of Birmingham
October 23rd, 2019

A roadmap

A roadmap

Large and small Firms



Innovation in Large and Small Firms: An Empirical Analysis

By Zoltan J. Acs and David B. Audretsch*

We present a model suggesting that innovative output is influenced by R&D and market structure characteristics. Based on a new and direct measure of innovation, we find that (1) the total number of innovations is negatively related to concentration and unionization, and positively related to R&D, skilled labor, and the degree to which large firms comprise the industry; and (2) these determinants have disparate effects on large and small firms.

As Simon Kuznets (1962) observed, perhaps the greatest obstacle to understanding the role of innovation in economic processes has been the lack of meaningful measures of innovative inputs and outputs. More recently, there has been the development of new data sources measuring different aspects of technical change. These new sources of data have included measures of patented inventions from the computerization by the U.S. Patent Office (Bronwyn Hall et al., 1986; Adam B. Jaffe, 1986; Ariel Pakes and Zvi Griliches, 1980), better measures of research and development (John Bound et al., 1984, and F. M. Scherer, 1982), and stock market values of inventive output (Pakes, 1985). While several of these new and improved data sources have been used to examine the relationship between innovative activity and firm size, there have been virtually no studies able to apply a more direct measure of the innovative output. For example, the limitations of using patent data were significant enough to supplement them with renewal data (Pakes and Mark Schankerman, 1984). Further, while most of the empirical research has examined only the innovative activity contributed by relatively large firms, the innovative output of the smallest firms has received only scant attention and quantification.[1] Thus, most of the inferences which have been made about the causes of innovative activity have been based on observing only the behavior of larger firms.[2] Such inferences may be misleading since, as we show, almost half of the number of innovations are contributed by firms which employ fewer than 500 workers.

The purpose of this paper is to add to the literature on new measures examining technical change by introducing a more direct measure of innovative activity, to determine some of its basic properties, and to illustrate its use with a reduced form empirical model. We present a model which investigates the degree to which innovative output is affected by different industry characteristics, and the extent to which small and large firms respond differently to various stimuli. The econometric analysis enables the testing of

*Research Fellows, Wissenschaftszentrum Berlin, Reichpietschufer 50, D-1000 Berlin 30, Federal Republic of Germany. We wish to thank Georg Bathgenner, Bo Carlsson, Paul Geroski, Albert N. Link, Richard R. Nelson, William K. Schelter, Joachim Schwalbach, J.-Matthias Graf von der Schulenburg, Hideki Yamawaki, Klaus Zimmermann, two anonymous referees, and seminar participants at the U.S. Small Business Administration, Case Western Reserve University, and the University of Bradford for helpful comments. We are especially grateful for the suggestions by F. M. Scherer and the computational assistance of Michael Karge and Jianping Yang. All errors and omissions remain our responsibility.

[1]For a thorough review of the literature relating technical change to innovation activity, see Morton I. Kamien and Nancy L. Schwartz (1975), F. M. Scherer (1980), and Richard C. Levin et al. (1985).

[2]For example, Scherer (1965) related market structure to the number of patents for fewer than 500 of the largest U.S. corporations.

678

Do innovative start-ups perform better?

Pros

- Better products and services (Guerzoni, 2010)
- Less myopic (Christensen, 1995)
- No sunk cost bias (Aestebro et al., 2007)
- More dynamic (Teece, 2012)

Cons

- Uncertainty in demand (Guerzoni, 2010)
- Uncertainty in technological evolution (Dosi, 1982)
- Uncertainty in competition (Fudenberg et al., 1983)
- Financial constraints (Stucki, 2013)

Audretsch, 1995

'The evidence therefore suggests that a highly innovative environment exerts a disparate effect on the post-entry performance of new entrants.'

The sectoral dimension

The Schumpterian patterns of innovation

Malerba and Orsenigo (1997) surmized that sectors can explain innovative behaviour much better rather than the micro characteristics of the firm. Namely the technological base of a sector can explain a firm's innovativeness, performance, size and turmoil.

The industry life-cycle

Klepper (1996) and Gerosky(1995) empirically showed that the stage of life of a sector is the key determinant for explaining both entry and exit dynamics and innovativeness.

The Regional dimension

'Entrepreneurship is a regional event' (M. Feldman)

- regional policies;
- agglomeration economies;
- infrastructure;
- entrepreneurial atmosphere;
- amenities;
- user-producer interactions;
- universities;
- ...

Issue 1: Poor empirical evidence

Poor empirical evidence

Hyytinen et al. [2015] survey the literature and conclude for a mild evidence of positive effects on innovativeness. However, just to mention a few:

- Cefis and Marsili (2006) do not control for the sector;
- Colombelli (2016): small and significant effect for process innovation only;
- Helmer and Rogers (2010): very little significance at the industry level;

Issue 2: Measuring Innovation

Innovation Input variables

- R&D investment
- Cost of scientific personnel
- High-skilled workers

Innovation Output variables

- Process and product innovation
- Patent

Issues

- register data for costs and investments are not always reliable
- small firms do not have formal R&D
- the number of process and product innovation comes from self-reported survey (CIS)
- there is a huge variance among firms in the propensity to patent
- only a low percentage of patents is actually valuable

Issue 3: Business cycle as a confounding effect

Firms in times of crisis

New firms can prosper or fail for a large variety of factors which do not necessarily relate with economic or technological conditions at the micro level.
For instance, vulnerable firms might survive in a growing economy even if not profitable, while selection mechanisms become stricter in downturns.

Contribution

Ideas

In this paper we analyse survival and growth of innovative and non-innovative start-ups considering:

- the entire population of firms*
- a new empirical measure for innovativeness
- a period of crisis when constrains are more binding and economic and technological conditions are extremely important.

Methods

Our approach combines machine learning (predictive modeling) and econometrics (causal modeling)

A roadmap

Innovative start-ups according to the Italian Law 179/2012

Firms are innovative if they:

- are newly established or have been operational for less than 5 years in EU with at least a production site branch in Italy;
- have a yearly turnover lower than 5 million Euros;
- do not distribute profits;
- produce, develop and commercialise innovative goods or services of high technological value;
- are not the result of a merger, split-up or selling-off of a company or branch;
- show an innovative character, i.e. if:
    - at least 15% of the company's expenses can be attributed to R&D activities ;
    - at least 1/3 of the total workforce are PhD students, the holders of a PhD or researchers; alternatively, 2/3 of the total workforce must hold a Masters degree;
    - the enterprise is the holder, depositary or licensee of a patent or the owner of a program for original registered computers.

Solving an Issue

Law 179/2002

What are the benefits in using Law 179/2002 for the identification of innovative start-ups?

- We focus on small firms, which are very likely to be truly new entities and not subsidiaries or foreign green-field entrants.
- All innovative firms are focused on innovative goods or services.
- They need to have at least one of the usual proxy for innovative input and output, but not necessarily a specific one such as in previous works.

However...

The law has been coherently used only from 2013... not during the 2008 financial crisis!

Beyond just econometrics

Econometrics

Econometrics is a set of tools to highlight causal relations between variables. It evaluates uncertainty with statistical inference which imposes the use of simple models and specific assumptions. Low Power.

Supervised Machine Learning

SML is a set of tools to learn to classify observations in a pre-determined set of categories and make prediction about new data points. It evaluates uncertainty on a test-set and the complexity of the model has no boundaries. High power, no causality.

Unsupervised Machine Learning

UML is a set of tools for the creation of a partition of the data without any a-priori on the number and type of categories to be generate. Great hypothesis mining engine.

Data

AIDA dataset

Source: AIDA Bureau van Dijk, which contains information on Italian firms with the obligation to file financial statements*:

- 68,316 new firms (2013);
- a censored balanced panel of 65,088 new firms (2008-2018);
- 427 variables: identification codes and vital statistics activities and commodities sector legal and commercial information index, share, accounting and financial data shareholders, managers, company participation.

	2008	2013
Innovative	0	1,010
Not-innovative	65,088	67,306
Total	65,088	68,316
% All* Italian Start-ups	22.7%	24.7%
After MVA	39295	45576

The process

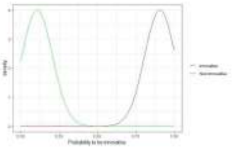
Data 2013

Training set
Learning Innovative firms

Test set
Evaluation

7 machine learning models

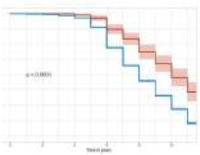
A mixture of two models (bagging and NN) with cut-offs for the clear case



Panel 2008-2018

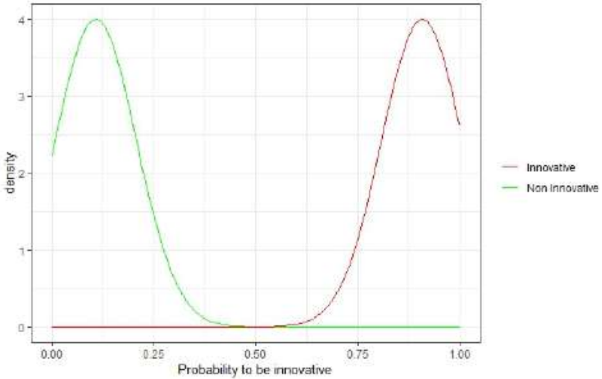
Prediction:
Innovative firms

Econometrics:
Survival analysis with Dummy Innovative, industry and province as exogenous covariates

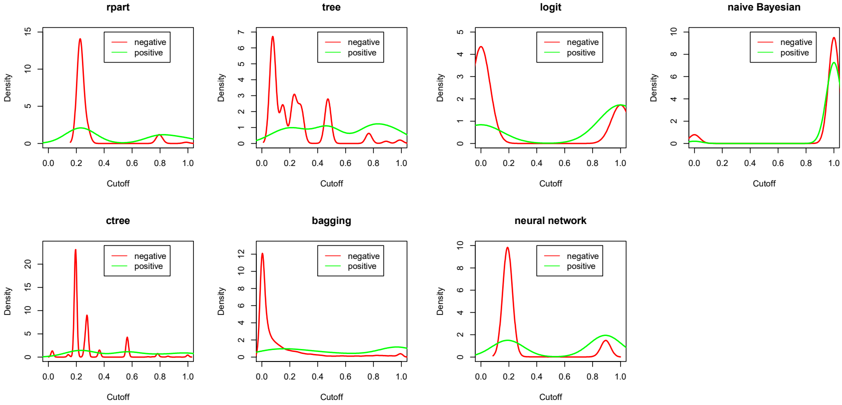


NB: industry and geographical variable are NOT used in the analysis

A well behaved model

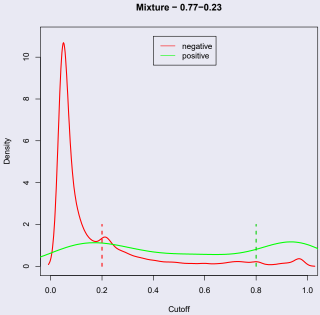


Learning and test



Learning and test

What is the best performing model?



- The selected model is mixture of two models. Weights minimize overlapping (0.77 and 0.23)
- Two cut-offs. We compare firms with either a very high or a very low probability to be innovative .

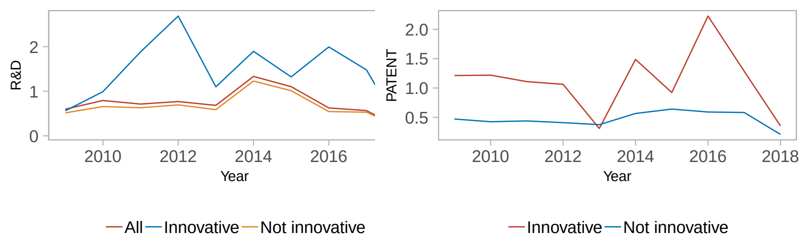
Prediction

Table: B-NN Mixture classification of not innovative (predicted probability $\leq$ 0.2) and innovative (predicted probability $\geq$ 0.8) start-ups on the 2008 sample

	predicted	probability	Total		%
	$\leq$0.2	$\geq$ 0.8		$\leq$0.2	$\geq$ 0.8
2008	34487	763	35250	87.8%	1.9%

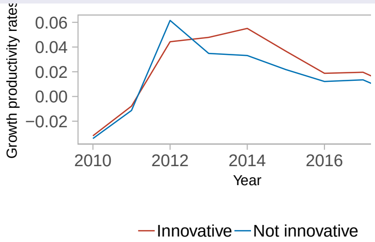
Robustness

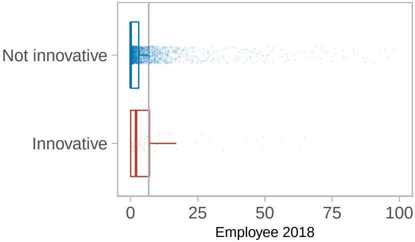
R&D and Patent



Other Statistics

Productivity and Employment





A roadmap

Survival 1

Kaplan and Meier estimator

The survival at time t is:

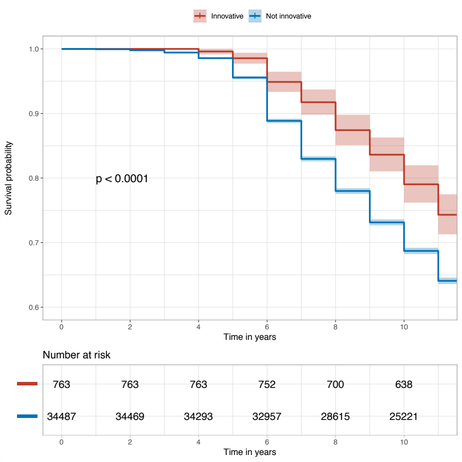
$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{r_i}\right) \tag{1}$$

Confidence interval

$$\hat{S}(t) \exp\left\{ \pm \frac{z_{1-\alpha/2}\hat{\sigma}(t)}{\hat{S}(t) \ln \hat{S}(t)} \right\} \tag{2}$$

Survival 1

Survival curve



Survival 2

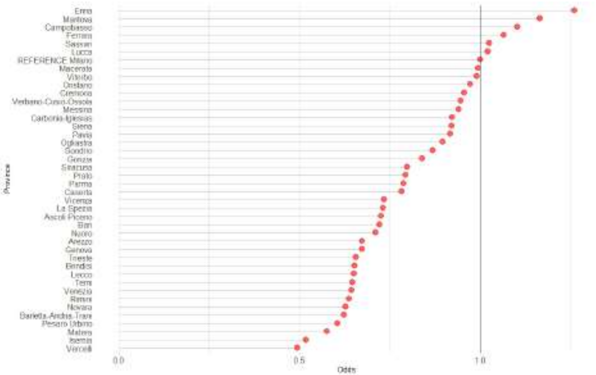
Table: Cox regression

	Dependent variable:				
	Hazard				
	(1)	(2)	(3)	(4)	(5)
Innovative	-0.428^{***}	-0.459^{***}	-0.438^{***}	-0.512^{***}	-0.122
	(0.072)	(0.072)	(0.072)	(0.198)	(0.246)
Industry Controls		YES		YES	
Province Controls			YES		YES
Interaction with Industry				YES	
Interaction with Province					YES
Observations	35,250	35,212	35,250	35,212	35,250
Note:			*p$<$0.1; **p$<$0.05; ***p$<$0.01		

Survival curve

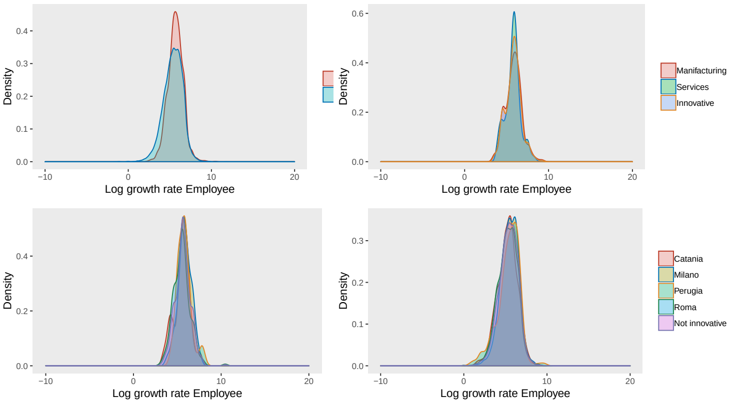
Innovative firms have a hazard ratio of $e^{-0.428} = 0.65$ i.e. at any given time innovative firms almost double their chance of survival. We can compute the same for the interaction which is the survival premium (or curse) for innovative firms in a specific sector or geography.

Interaction effect

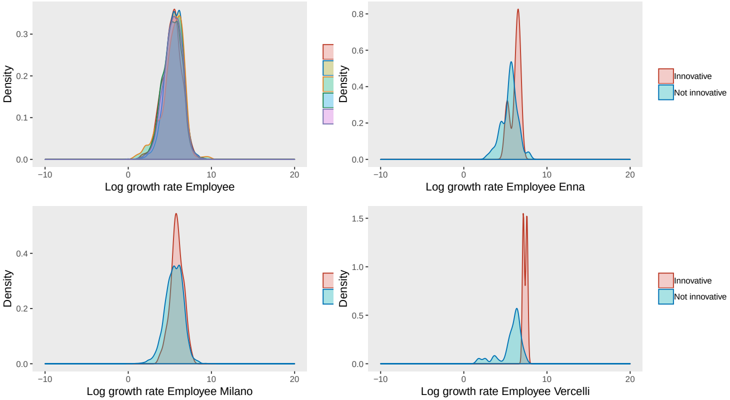


The effect of being innovative within a specific province. Values showed if significant, Milan is the reference

Growth 1: density distribution



Growth 2: regional focus



Conclusion

Policy

Innovativeness is a crucial factor for survival and growth of new firms but only in the right place and in the right industry.

Methodology

The combination of machine learning and econometrics allows to explore causal and non-causal effects when data quality is initially low

[marco.guerzoni@unito.it]