Can we predict firms' innovativeness?

The identification of innovation performers in an Italian region through a supervised learning approach

Claudio Cozza (University of Naples Parthenope, Italy)
Ilaria Gandin (Area Science Park, Trieste, Italy)

New analytical tools and techniques for economic policymaking
Paris, April 16th 2019

Machine learning and (innovation) policymaking

- Over the very last years, a growing number of events and reports have supported the idea that policymaking should rely on new data, methodologies and techniques: Big data, Artificial Intelligence, Machine learning (among the first ones: OECD conference on AI, October 2017; EC-JRC report “Using new data sources for policymaking, 2017)

- The well established idea of evidence-based policies is being complemented by the use of non-parametric techniques

- This tendency is clear both for policymaking overall and for specific areas, including innovation studies and policies

The measurement of innovation…

… has always been problematic:

- **Output** (patents)?
  - Extensive use of patents as a measure of innovative activity (see Acs & Audretsch, 1989)…
  - … but limitations: due to firm size (large firms patenting more than SMEs) or sector/strategical reasons (“Not all inventions are patented. Firms sometimes protect their innovations with alternative methods, notably industrial secrecy”, Archibugi & Pianta, 1996)
- **Input** (R&D)?
  - Statistics on R&D expenditures since the 1960s are very frequent at country-level but less at micro-level, for confidentiality reasons
- **Innovation behaviour** (firm-level questionnaires)?

The Community Innovation Survey (CIS)…

- To overcome this limitation, since early 1990s many studies dealing with the direct measure of innovation in Europe have relied on the outcomes of the Community Innovation Survey (CIS) and to its statistical guidelines contained in the Oslo Manual

- High statistical costs → CIS has not a census approach for SMEs
  → it usually does not cover micro firms

- These issues especially affect EU countries/regions where a very large majority of firms is represented by SMEs and micro firms

… and its limitations

- In addition: CIS is meant for statistical purposes, thus implying its confidentiality

- As a result, stakeholders willing to know whether a specific firm is innovative or not…
  - **Policymakers** in terms of ex-ante analysis to optimally allocate public funding to innovation policy
  - and/or **Venture Capitalists** looking for support to an investment decision

- … are not allowed to explicitly know it

New perspective with machine learning techniques

- Successful examples can be found in health management, urban development, economic growth, educational system, public inspections

- The key idea is that machine learning algorithms provide predictions that, together with context-specific constraints and targets, help administrators in making decisions for a proper resource management

- Only some recent studies concern the economics of innovation and firm behaviour; none of them address our specific research questions: ***is it possible to predict firms' innovativeness?*** → i.e. ML classification problem

- We rely on the extensive theoretical and empirical literature (Griliches, Crepon-Duguet-Mairesse, etc.) on the drivers of firms' innovativeness

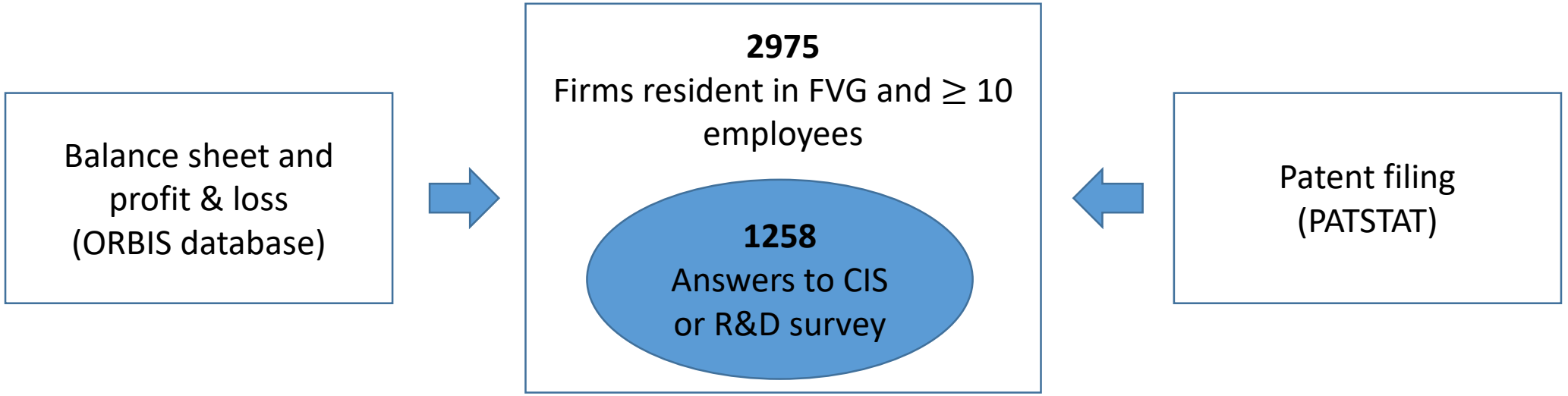
The starting point

- We had access to an ***intelligence tool*** aimed at ***supporting policymakers*** of an ***Italian NUTS2 region***, Friuli Venezia Giulia
- ***Large*** dataset in terms of variables, including information on all firms active in the region:
  - Administrative records
  - Balance sheet data
- ***Integrated*** with other information on firms' innovativeness:
  - Innovation, R&D, export and group belonging data, coming from the National Bureau of Statistics
  - Information on patent applications and EU projects participation (FP7 and H2020)
  - ISO certifications

Data

Cohort of firms with at least 10 employees and resident in "**Friuli-Venezia Giulia**"
Focus on four blocks of data:

- Chambers of Commerce registry for **administrative records**
- Bureau van Dijk ORBIS database for **balance sheet data**
- Data concerning **innovation performance** are:
  - ISTAT **CIS, years 2012** and **2014**
  - ISTAT **R&D survey, years 2012** and **2014**
- PATSTAT for an additional check on **patents**



Data (2)

Binary outcome variable: innovators (INN) *vs* non-innovators (N-INN)

Y= **INN**, if a firm reports innovation expenditures >0

N-INN, otherwise

Predictors:

- *Pavitt class*
- *No. of employees*
- *Turnover*
- *Turnover on cost of employees*
- *Profit–loss on cost of employees*
- *Creditors turnover ratio*
- *Employee average cost*
- *Share of intangibles on fixed assets*
- *Share of fixed assets on total assets*
- *ROS*
- *ROI*
- *ROE*
- *Leverage*
- *Net long term debt*
- *Net short term debt*

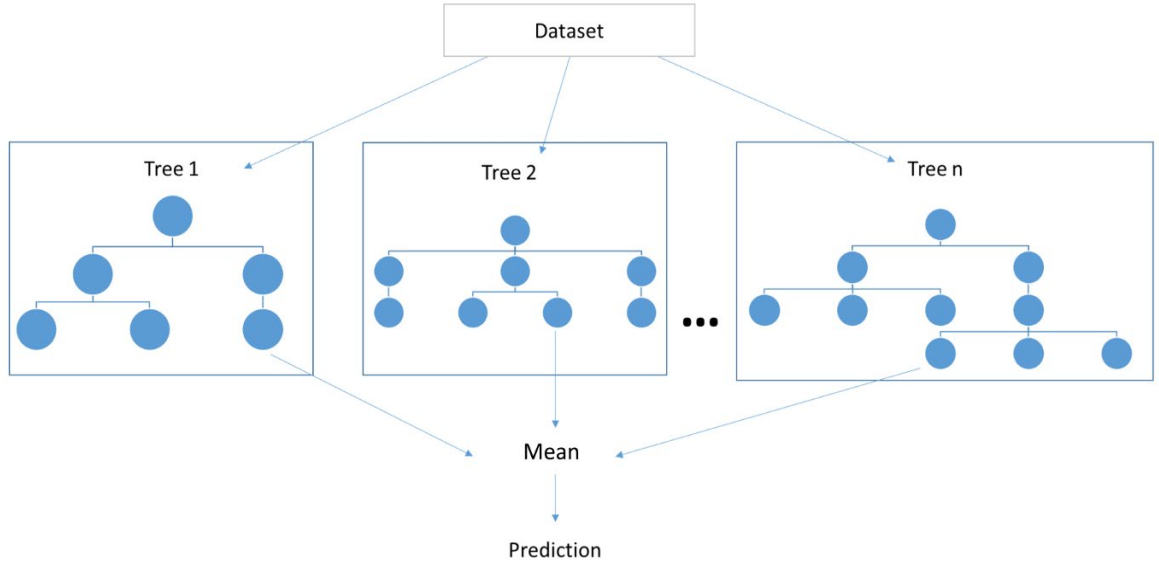
Methods

Classification problem:

1) a **classification model** based on companies' features is calculated for respondents (**survey sample**)

2) the model is applied to the remaining companies (**out-of-survey firms**) to predict their class

Random Forest algorithm



- Able to integrate non-linear effects and suitable when data are not abundant
- Ensemble method based on decision trees
- Class probability estimate obtained averaging the votes of all trees forming the forest
- Out-Of-Bag data can be used to perform an internal validation
- Default hyper-parameters of the RandomForest R package
- Model's performance is measured based on the mean squared error (MSE)

V. Kumar, "Random forests and decision trees from scratch in python", towardsdatascience.com

Results

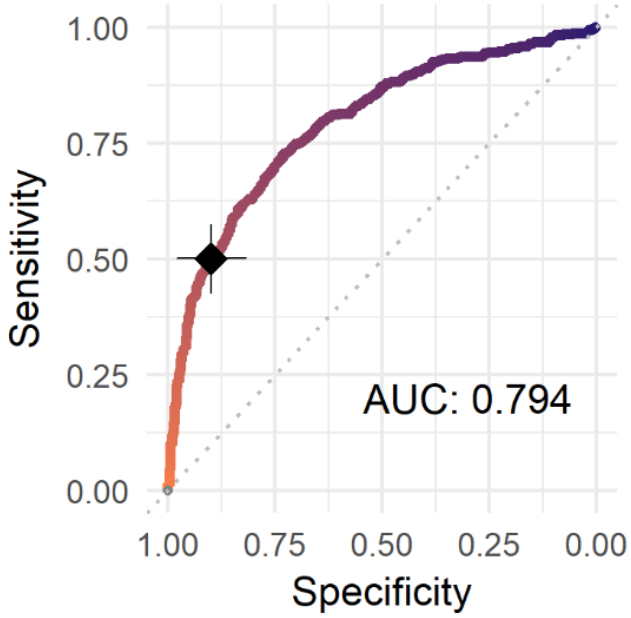
Classes are slightly unbalanced: INN 44.91% and N-INN the 55.09%. In our framework of policymaking support, the desirable model should perform well in terms of precision.

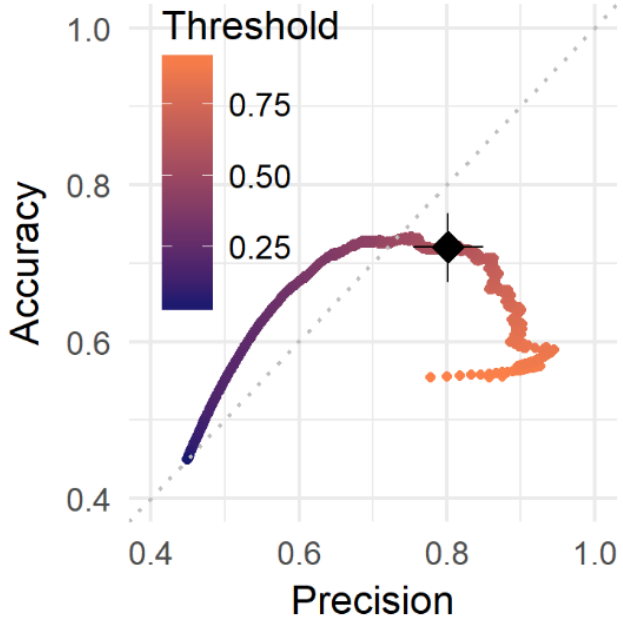
Operating point: thresholds within class probability that splits observations into INN and N-INN classes

ROC curve: specificity and sensitivity calculated at different threshold.

Performance: AUC 0.794

Accuracy-precision trade-off: suggests operating point = 0.589 as a good choice aiming to obtain 80.17% of precision, 72.02% of accuracy, 89.90% of specificity and 50.09% of sensitivity



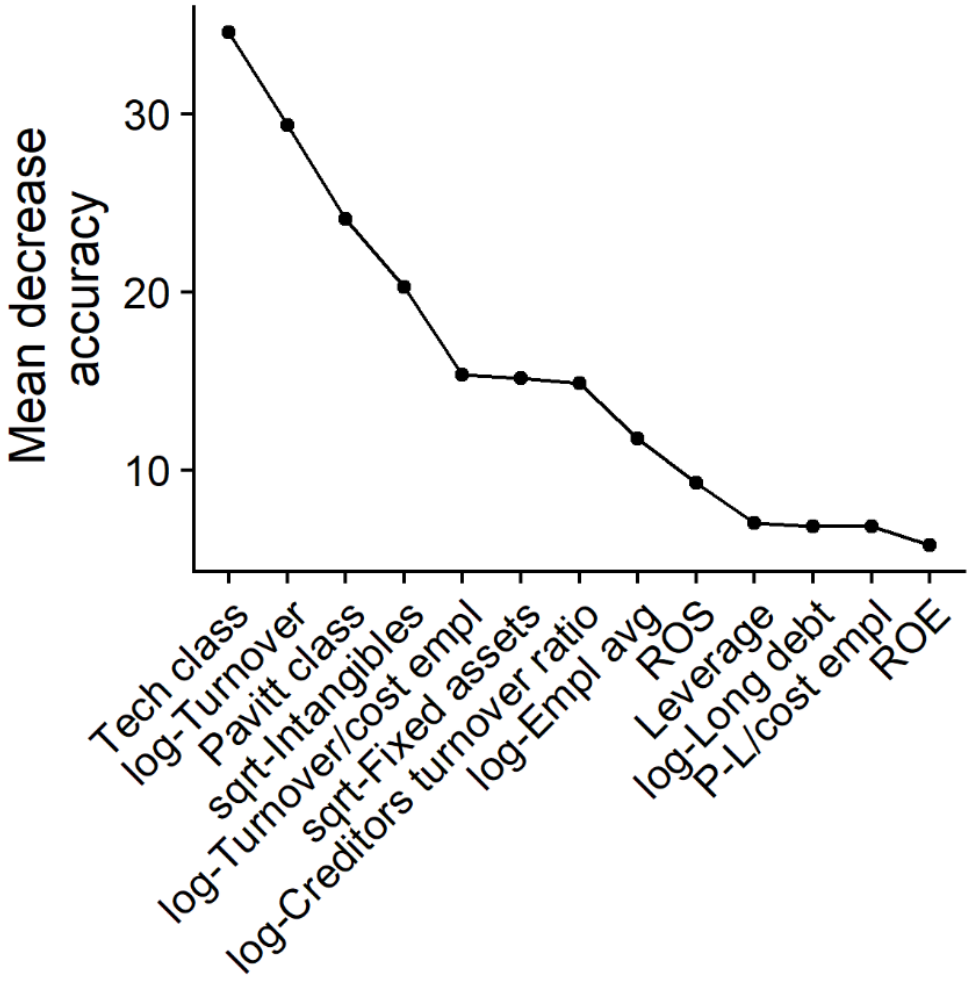


Results (2)

Variable importance: obtained permuting the values of each variable and measuring how much classification accuracy decreases

4 most important:

- Tech class
- log-Turnover
- Pavitt class
- Sqrt-Intangibles

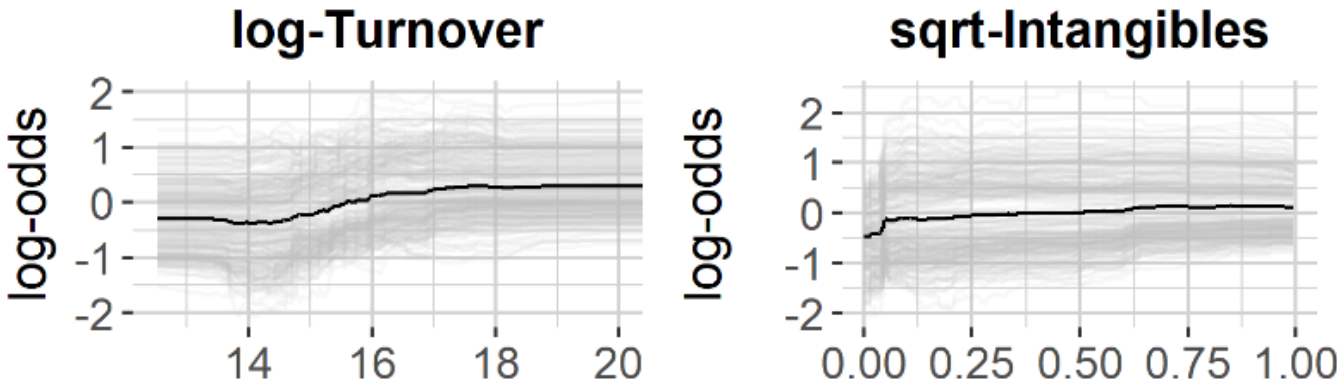


Results (3)

Partial Dependent Plots: The idea of PDPs is to focus on one independent variable at a time and analyse the behaviour of a function approximating its marginal distribution.

Individual Conditional Expectation plots: a group of curves are graphed for individual observations (refinement of PDPs)

Positive contribution: however, in the case of intangibles the increase is concentrated in a small range of values suggesting the presence of a threshold effect



Results (4)

Prediction on out-of-survey firms. Comparison between model predictions and innovation outputs can be done considering information about patent filing:

- Focus on patents deposited within the survey year (2012 and 2014)
- compare model predictions and companies having any patent either to the Italian Patent Office (UIBM) or directly to the European patent office (EPO)

- **Observations with at least one patent:** 41 of them (71.9%) are actually predicted as innovative

- **Patenting rate:**
  - 5.7% for INN
  - 0.9% for N-INN

	N-INN	INN	N-INN prop.	INN prop.
UIBM	9	29	0.237	0.763
EPO	7	15	0.318	0.682
any	16	41	0.281	0.719

Conclusions: limitations and *further steps*

- Main limitation of our pilot study concerns the number of observations

- As a consequence, we obtain high precision (shifting the operating point) but low sensibility → high proportion of false negatives
  - *We are currently extending the study to country-level data, in order to handle high heterogeneity*

- CIS is not the only tool to measure innovation
  - *We are currently applying machine learning techniques to other research and innovation data (patents, publications, EU projects participation), thus retrieving more predictors* and obtaining an advanced profiling of innovators

- The overall goal is to prove the effectiveness of machine learning techniques in supporting innovation policymaking