Machine Learning tools for the detection of state aid recipients

Proof of concept for the European Commission – Directorate general Internal Market, Industry, Entrepreneurship and SMEs and Directorate general Competition

Guglielmo Barone (University of Bologna), Marco Letta (Sapienza University of Rome)[1]

Final report, July 9th, 2021

Table of contents

[1] Guglielmo Barone coordinated the report and wrote Sections 1, 4, 6, 7, 8. Marco Letta wrote Sections 2, 3, 5. The authors want to thank Antoine Mathieu-Collin, Malwina Mejer, and Jose-Luis Varela-Irimia for their useful comments. The views expressed are those of the authors and do not represent necessarily those of the European Commission. Email: Guglielmo Barone: g.garone@unibo.it; Marco Letta: marco.letta@uniroma1.it.

Executive summary

- *A state aid is an economic aid (such as grants, tax exemptions, soft loans, public guarantees), granted from state resources, that delivers an economic advantage to a certain company, economic sector, or region. They are selective, meaning that they are likely to favour economic units receiving the aid. In this respect, they are different from other policies open to all enterprises.*
- *Within the European Union (EU), in light of its key objective of creating common markets for goods and services, there are strong concerns about state aid's anti-competitive effects. In principle, state aids are forbidden even if the European Commission has the power to approve certain aid measures if their beneficial effects are predicted to overcome distortions to competition. This set of rules level playing field among firms with respect to subsidies provided by the EU Member States.*
- *However, it may well happen that an undertaking operating in the internal market receives "foreign" subsidies, that is, financial contributions by a government (or other public bodies) of a non-EU State. Foreign subsidies are likely to distort the EU internal market, giving rise to an uneven playing field in which less efficient firms grow and increase market share at the expense of internal and more efficient operators. Unfortunately, there is no information on the actual number of foreign subsidies being granted to EU firms.*
- *The aim of this report is to assess whether recipients of state aids can be accurately predicted by applying machine learning tools applied to financial accounting data.*
- *The analysis combines data on financial statements taken from the Bureau van Dijk Orbis database with data on state aid recipients coming from the Transparency Award Module database, managed by the European Commission. On a more technical ground, the model predicts the status of being a state aid recipient or not using information on financial accounting data.*
- *Data on the UK, those on micro firms, as well as those on firms in industries with very specific sectoral rules or with limited relevance in the issue under investigation (such as agriculture, banking and finance, non-market service sectors) have been dropped. Because of data availability, the analysis focused on grants received in 2016 and 2017.*
- *The final dataset includes about 11 million observations; the potential predictor set is very large (around 190 variables). The following machine learning models have been tested: logistic lasso, classification tree, random forest, and evaluated against a simpler logit model as a benchmark.*
- *The best model is a classification tree, which is able to handle very well the missing data problem in the Orbis data, and whose predictive performance is very high. The prediction is based on five variables, combined in a non-linear fashion, out of about 190 potential predictors, thus drastically reducing the original dimensionality problem.*
- *According to this model, 13.2% of observations are classified as false positive, that is, firms that in a given year did not receive grants but are predicted to have done so.*
- *The descriptive analysis suggests that firm size, country, and sector are relevant factors associated with the false positive status.*
- *Thanks to the model, each firm can be labelled as suspect or not; this piece of information is a preliminary screening device than can be used as a base for further investigation.*

1. Introduction

- A state aid (SA) is a specific form of economic aid, granted from state resources, that delivers an economic advantage to a certain company, economic sector, or region. They are selective, meaning that they are likely to favour companies receiving the aid relative to their competitors. In this respect, they are different from other policies open to all enterprises.
- SAs can be provided in many ways, such as grants, tax exemptions, soft loans, public guarantees, etc. They can be usually classified as (i) horizontal aids; (ii) vertical (sectoral) aids, targeting a specific sector or an individual firm; (iii) regional aids that support lagging-behind regions.
- While, in principle, the correction of market failures and/or the reduction of inequality can provide their economic rationale, there are strong concerns about SAs' anti-competitive effects. Within the European Union (EU), in light of its key objective of creating common markets for goods and services, the control of SA has been of key interest.
- In principle, SAs are forbidden by the EU treaty because they are likely to distort competition and adversely affect trade. However, the European Commission (EC) has the power to approve certain aid measures if their beneficial effects are predicted to overcome distortions to competition. This set of rules level playing field among firms with respect to subsidies provided by the EU Member States.
- However, it may well happen that an undertaking operating in the internal market receives "foreign" subsidies, that is, financial contributions by a government (or other public bodies) of a non-EU State. This may be the case, for example, if the undertaking is ultimately owned or controlled by a non-EU government. Foreign subsidies are likely to distort the EU internal market, giving rise to an uneven playing field in which less efficient firms grow and increase market share at the expense of internal and more efficient operators.
- While EC can keep track of SAs granted according to the EU exemption rules, there is no information on the actual number of foreign subsidies being granted to EU firms.
- **The aim of the project is to assess whether recipients of SAs can be accurately predicted by applying machine learning (ML) tools applied to financial accounting data.**
- Consistently with the background outlined in the Introduction, **the ideal training dataset would have included data on both foreign aid recipients and on foreign aid non-recipients**. Unfortunately, such piece of information is not currently available.
- Hence, the analysis regards the prediction of internal SA recipients; as such, it should be considered as proxy analysis of the first best study sketched above.
- The assumption for the external validity of this second best exercise is that the data generating process for hidden European recipients is the same as that for hidden non-European recipients. In other terms, there are no systematic differences between these two groups of firms. This assumption is not empirically testable, given the current data availability.

2. Data

2.1. Data sources

- **Firm level data are taken from the Bureau van Dijk Orbis database** (Bajgar et al., 2020). They include financial accounting data (taken from the balance sheet and profit and loss account: e.g. operating revenue, total assets), as well as information on age and on 4-digit industry. Data are available for 28 EU countries and for the 2014-2018 period. Years before 2014 have been dropped for reasons stated below, while 2019 has been dropped because of incomplete data (missing firms may be non-random).

- A potentially interesting piece of information is about ownership. In fact, unrecorded grants might go to the final beneficiary from another firm in the same group by means of infra-group operations. Data on group structure are from the Orbis ownership module. It includes the Global Ultimate Owner (GUO) calculated by Bureau van Dijk following the ownership pyramid and according to a proprietary algorithm. Orbis data are available both at the single firm level and the group level.
- The EC provided us with the concordance table between firms and groups as well as with financial data when the GUO is non-European (about 2% of the cases). We keep track of this with a dummy variable for non-European GUOs.
- The complete original Orbis database includes 178 variables. We selected a subset taking into account the following criteria: (i) a large number of variables is fully consistent with the maximization of the performance of ML techniques, which are explicitly designed to deal with large databases; (ii) many variables have a large number of missing values and, at the same time, some ML methods do not handle missing values (just like traditional regression methods), while others do; (iii) a fully data-driven selection of the most predictive features, without any a priori selection made by the analyst, does not necessarily maximize the predictive accuracy and risks to obfuscate model interpretability, with repercussions in terms of transparency and communication of the results; (iv) a huge number of potential predictors makes file transfers and manipulation very cumbersome.
- Namely, we selected 48 variables that (i) are likely to account for the most part of firm heterogeneity, (ii) are not plagued by an excessive number of missing values, (iii) allow computing the most common ratios. In fact, we also computed 16 further variables (ratios, etc.), based on the initial 48 variables (Table A1, Panels A and B).
- **Data on SAs come from the Transparency Award Module (TAM) database** (European Commission, 2018), an IT application the Commission has developed to help Member States to fulfil their transparency obligations. It includes subsidies that have been granted from 1 July 2016 onwards and that exceed a certain threshold.[2] Data for Romania, Poland, and Spain, provided as separate files, have been appended after translation and/or unit conversions. We selected grants extended in 2016 and in 2017 only, because we need to put them in relationship with leads and lags of Orbis variable (see below).
- The variables selected from the TAM database (Table A1, Panel C) are: 'National Identification', 'Granted Aid Absolute EUR', 'Aid Award Instrument', 'Aid Award Granted Date'.
- All the data above have been provided by the EC, as well as the key variable used to merge the two datasets.

2.2. Data Processing

- Orbis data
    - Firms belonging to the following industries have been dropped because of very specific sectoral rules: (i) Agriculture (Nace Rev. 2 Section A: from 0111 to 0322); (ii) banking and finance (Nace Rev. 2 Section K: from 6411 to 6630 except for 6420 Activities of holding companies, and 6430 Trusts, funds and similar financial entities) Codes 6420 and 6430 might be relevant to pinpoint global ultimate owners (see below). We also dropped firms in non-market service sectors (Nace Rev. 2 Sections from P to U) because of their limited relevance in the issue under investigation.
    - The UK has been dropped because we are interested in the development of a prediction model to be employed on future data in the EU context.

[2] 500,000 euros; lower for firms belonging to the same group.

- Micro firms (as defined according to the EU thresholds) have been dropped too, as they have been considered to be less relevant for the purpose of this analysis.[3]
- The selection of years deserves attention. Orbis data are available at least from 2010 to 2019 (last year is incomplete), while TAM data are available from mid-2016 onwards. The idea is to use (symmetric) leads and lags of the Orbis variables to predict the SA recipient status (see below). Moreover, aids granted in 2020 should be excluded because of the temporary framework on SAs linked to the Covid 19 pandemic. It follows that we keep TAM data relative to 2016 and 2017 and Orbis data from 2014 to 2018, so that for SAs granted in 2016 we have Orbis data referred to 2014 and 2015 as lagged predictors and data for 2016 and 2017 as led predictors; analogously, for SAs granted in 2017 we have Orbis data referred to 2015 and 2016 as lagged predictors and data for 2017 and 2018 as led predictors.
- Following Kalemli-Özcan et al. (2019), we dropped firm-year observations from the Orbis dataset when: they report negative values for any measure of assets (fixed, current, total assets; tangible, or intangible fixed assets; other fixed assets or other current assets), negative stock, negative debtors, negative employment, negative operating revenue turnover, negative shareholder funds or other shareholder funds, negative capital, negative current, negative non-current or other current liabilities, negative loans, negative creditors, or if age (measured as years since incorporation) is negative.
- In the original Orbis database, for each firm-year there can be more than one entry for the following reasons:
  - Accounts reported in Orbis data come in five main types: unconsolidated accounts of companies for which consolidated accounts are not available (code U1). Unconsolidated accounts of companies for which consolidated accounts are available (code U2); codes U1 and U2 are mutually exclusive. Consolidated accounts of companies for which unconsolidated accounts are not available (code C1). Consolidated accounts of companies for which unconsolidated accounts are also available (code C2); codes C1 and C2 are mutually exclusive. Accounts with very limited financial information (code LF). LF statements have been dropped because of the huge number of missing values. Unconsolidated accounts are used to generate firm level features, while consolidated accounts are used separately to construct predictors at the group level.
  - Different filing types (i.e. the fact that a company in the Orbis database can be associated with more than one financial statement for the same accounting year) and different closing dates have been collapsed using the procedure borrowed from Kalemli-Özcan et al. (2019) (software code has been provided by the EC).

- TAM data
  - Whenever the variable 'Granted Aid Absolute EUR' was missing and only 'Granted Range' was available, we followed the EC's routine and imputed missing values with median values of the corresponding range.
  - Granted Aid amounts were collapsed at the firm-year level (after generating a 'year' variable from the 'Granted Date' provided variable). When collapsing data, the Aid Award Instrument associated with the largest grant received in that year was considered as the main Aid Award Instrument.

[3] Specifically, micro firms are those with a number of employees lower than or equal to ten, and with total assets lower than or equal to 350,000 euros or operating revenue turnover lower than or equal to 700,000 euros. See https://ec.europa.eu/info/law/accounting-rules-directive-2013-34-eu/implementation/guidance-implementation-and-interpretation-law_en.

- After brainstorming with the EC, we decided to have two definitions of a firm as a SA recipient in a certain year, which are the target variable of our machine learning models.
  - According to the first (broader) definition, a firm is a SA recipient if the total annual granted aid was higher than500,000 euros in that year. Consequently, firms receiving an annual amount lower than or equal to 500,000 euros are considered non-recipients. Consistently, we generated a 'SA recipient' variable, which takes value 1 if a firm received aids for a total annual amount higher than 500,000 euros in a given year, and 0 otherwise.
  - According to the second (stricter) definition, a firm is a SA recipient if the total annual granted aid was higher than1,500,000 euros in that year and zero otherwise. The resulting model is more apt to focus on "large" hidden recipients.

- Merge of the datasets. Overall, the whole process consists of the following steps.
  - We took the original Orbis dataset, applied preliminary cleaning (as described above), selected only the subsample with consolidation codes U1 or U2, created new derived variables, collapsed it at the firm-year level, and merged it with the GUO codes.
  - Then, we went back to the original Orbis dataset, now selecting only C1 or C2 consolidation codes, repeating exactly the same data cleaning processing, and eventually collapsing the data at the GUO-year level. Finally, the same data for non-European GUOs were appended to this GUO-year level dataset.
  - We merged (m:1 in the Stata parlance) the firm-year level Orbis dataset and GUO level Orbis, using GUO and year as merging variables. The output is a firm level Orbis dataset with the same set of firm and GUO level predictors. Years are from 2014 to 2018
  - We merged the firm-year Orbis dataset (enriched with GUO predictors) and the firm-year TAM dataset by employing the merging key provided by the EC. It includes firm-year observations for the period 2014-2018; TAM data are only for years 2016-2017.
  - The final Orbis-TAM merge resulted in 11,378 observations (for the years 2016 and 2017) imported from TAM to Orbis, of which 8,022 received a total annual granted aid above the 500,000 Euro cutoff. 8,388 TAM observations for the same years were not imported into the Orbis database, because: (i) some sectors are excluded (see above); ii) micro firms are excluded (iii) the key employed to merge the two data sources is not available for all observations; (iv) there may be different representativeness/unbalancing of the Orbis vs. TAM databases with respect to the universe of firms.

2.3. Pooling the data

- Albeit recent applications to longitudinal data have recently emerged in the economics literature (especially for applications at the intersection between prediction and causal inference questions), **ML methods for predictive tasks are typically applied to cross-sectional data**. Hence, for the purpose of this project, we will consider 2016 and 2017 as pooled, without focusing on the longitudinal component.
- Since the estimation idea is to use changes in outcomes to predict the existence of SA (see below), the original panel structure has been changed as shown in Figure A1 (compare Panel A with Panel B): **the variable to be predicted is the SA 0/1 status (non-recipient/recipient) while potential predictors are leads and lags of Orbis variables**. Mind that we are conventionally treating Orbis variables measured in year t as a realization of the variable after the aid granted in t (because Orbis data are measured at the end of the year while the subsidy is granted during the year). This is purely conventional, but it helps in thinking symmetrically about leads and lags ($\pm$ 2 years). For the sake of simplicity, we averaged two lagged variables and two leaded variables so that for each Orbis variable

X we have a 'pre' value ($\frac{X_{t-1}+X_{t-2}}{2}$) and a 'delta' value computed as 'post' value ($\frac{X_t+X_{t+1}}{2}$) – 'pre' value ($\frac{X_{t-1}+X_{t-2}}{2}$).

3. Descriptive analysis

- After data processing, we have a dataset at the firm-year level for the years 2016-2017, which includes firm and GUO level predictors for both pre and post periods, generated from Orbis data spanning from 2014 to 2018. Here we provide some preliminary descriptive statistics on the characteristics of this dataset.
- The total number of observations in this dataset is 11,055,248 observations, 11,047,226 of which are SA non-recipients, and 8,022 are SA recipients (either in 2016 or in 2017; Table 1). This means that 99.93% of observations are referred to non-recipient firms, while SA recipients account only for 0.07%. Such values point to a **severe data imbalance problem** that will have important consequences on the analytical methodology and the success of the predictive task, which will be addressed below.

Table 1: Sample size by SA status

SA Recipient	Frequency	Percentage
0	11,047,226	99.93%
1	8,022	0.07%

Notes: The sample consists of firms receiving SAs in the years 2016 and 2017. Observations with an annual granted aid less than 500,000 euros are considered as non-recipients. Variables 'SA recipient' is a binary variable taking value 1 if the firm is a SA recipient and 0 otherwise.

- Table A2 reports summary statistics for some selected Orbis variables (namely, total assets, operating revenue turnover, number of employees, labour productivity, and a dummy for GUO data existence) and suggests that: (i) **SA recipients are much larger compared to the control group** (Panel A and Panel B, respectively). The availability of GUO level information is also substantially unbalanced between the two groups, with a much higher share of SA recipients having associated GUO data; (ii) **all variables display large residual variability**, despite having removed outliers for many of them by trimming the 1st and 99th percentiles of all the ORBIS variables.
- Table 2 presents basic descriptive statistics, by type of aid instrument, on the granted aids and the ratio between them and operating revenue turnover. Again, there is large variability in granted amounts, in addition to the dispersion of the Orbis variables highlighted above. The huge heterogeneity of the two statistics across different instrument types is particularly noteworthy.
- Figure 1 illustrates the distribution of the sample by geographic area and industries for two categories: all firms and SA recipients. The mode of firm location is in Western Europe (38%, Panel A), **SA recipients are also disproportionally located in Western Europe** (47.3%). As to the sectors (Panel B), **SA firms disproportionately belong to NACE sectors B** and C (47.8), whereas all other sectors are de-specialized.
- Figure 2 reports analogous histograms looking at the distribution by quartiles of total assets (Panel A) and of operating revenue turnover (Panel B). In both cases, we see that the distribution by size of

SA recipient firms is severely unbalanced, as more than 75 % of **SA recipients belong to the upper quartile of the distribution of both variables**.

Table 2: Summary statistics of granted aid amounts for SA recipients

Variable name	Unit	Mean	Median	SD	Min	Max	Obs
All							
Granted aid	€	3,515,054	1,072,369	2.70e+07	500,400	1.51e+09	8,022
(Granted aid/Operating revenue turnover)*100	%	126,458.4	6.788	6,220,797	0.00176	3.60e+08	4,738
Direct grants/Interest rate subsidies							
Granted aid	€	1,834,381	947,097.5	3,500,265	500,400	9.37e+07	3,864
(Granted aid/Operating revenue turnover)*100	%	246,412.9	9.983	8,728,775	0.00176	3.60e+08	2,406
Tax advantages							
Granted aid	€	3,386,758	1,500,000	4,575,182	500,490.8	4.02e+07	1,489
(Granted aid/Operating revenue turnover)*100	%	4.175	1.692	10.801	0.00464	222.455	959
Direct grants							
Granted aid	€	9,100,266	1,016,595	6.72e+07	500,739.8	1.51e+09	1,258
(Granted aid/Operating revenue turnover)*100	%	2043.98	13.049	32,737.83	0.00755	603,013.05	340
Loans							
Granted aid	€	1,583,733	994,963	1,577,694	500,465.4	1.20e+07	621
(Granted aid/Operating revenue turnover)*100	%	238.737	14.327	1749.339	0.00729	28,281.16	395
Others							
Granted aid	€	4,601,522	2,430,583	7,720,565	501,732	1.30e+08	790
(Granted aid/Operating revenue turnover)*100	%	1,074.364	50.406	20,051.93	0.00518	506,549.6	638

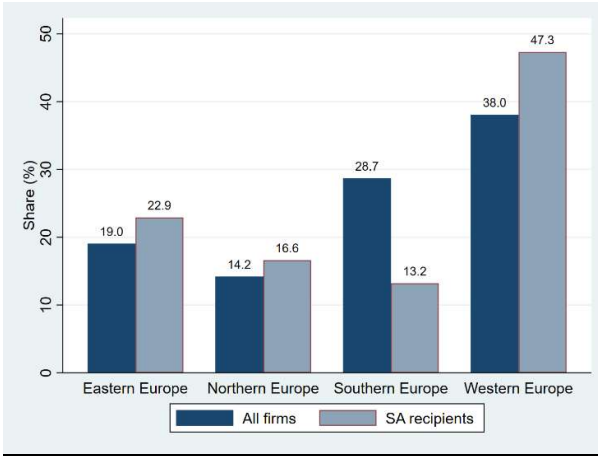
Notes: The sample consists of total annual aids granted to each recipient firm for the years 2016 and 2017. Observations with an annual granted aid less than 500,000 euros are considered as non-recipients. '(Granted aid/Operating revenue turnover)*100' is the ratio between granted aid amounts and operating revenue turnover of SA recipients, multiplied by 100.

- Similar insights emerge from the SA distribution by firm performance in Figure 3. Here we employ two measures of performance: Return on Assets (Panel A) and labour productivity (Panel B). Again, for both variables, **SA recipients are disproportionately distributed in the upper quartile,** albeit to a minor extent than for the firm size variables.
- Finally, Figure 4 shows the share of the number of SA recipients by aid instrument type. Bear in mind that, as several different aid instruments can be associated with different aids granted in a given year

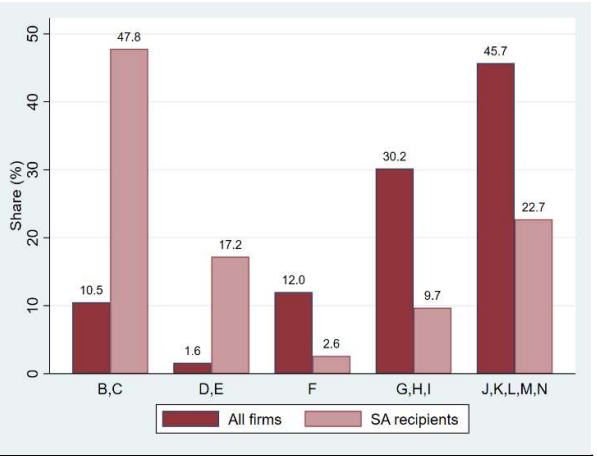
to a recipient firm, we selected, for each year, the type of instrument associated with the largest aid amount granted to the receiving firm in that year. The figure suggests that granted aids take predominantly the form of direct grants/interest rate subsidies (48.2%), followed by tax advantages (18.6%), direct grants (15.7%), and loans (7.7%).

Figure 1: Sample distribution by geographic area and industries

Panel A: Share of the number of firms by geographic area



Panel B: Share of the number of firms by NACE sectors



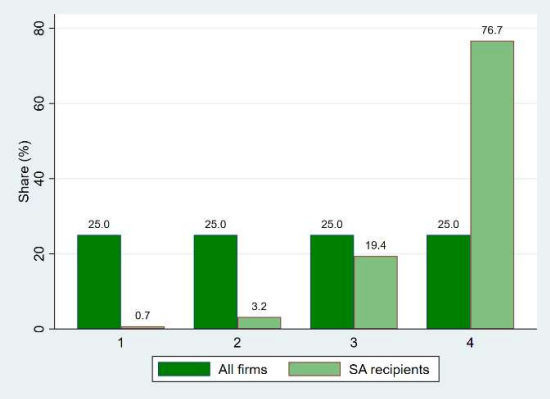
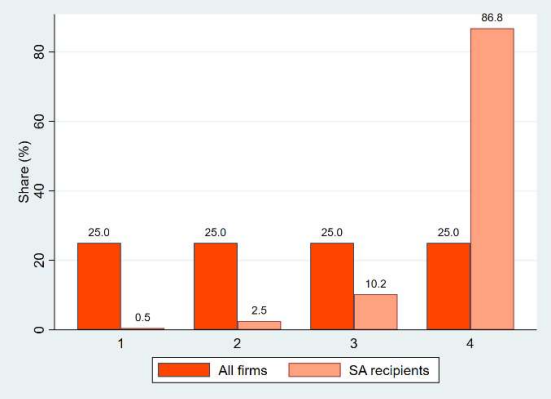
Notes: Panel A: The sample includes information for the years 2016 and 2017. European geographic areas are defined following the UN Geoscheme classification for geographic regions, except for Cyprus that is included in Southern Europe. Eastern Europe: Bulgaria, Czechia, Hungary, Poland, Romania, Slovakia. Northern Europe: Denmark, Estonia, Finland, Ireland, Latvia, Lithuania, Sweden, United Kingdom. Southern Europe: Cyprus, Croatia, Greece, Italy, Malta, Portugal, Slovenia, Spain. Western Europe: Austria, Belgium, France, Germany, Luxembourg, Netherlands (see https://unstats.un.org/unsd/methodology/m49/). Source: Authors' elaborations on Orbis and TAM data. Panel B: The sample includes information for the years 2016 and 2017. Sectors are as follows (NACE sections): B = Mining and quarrying; C = Manufacturing; D = Electricity, gas, steam and air-conditioning supply; E = Water supply, sewerage, waste management and remediation; F = Construction; G = Wholesale and retail trade, repair of motor vehicles and motorcycles; H = Transportation and storage; I = Accommodation and food service activities; J = Information and communication; K = Financial and insurance activities; L = Real estate activities; M = Professional, scientific and technical activities; N = Administrative and support service activities. In the G-M aggregate, sectors K and L (except for 6420 (Activities of holding companies), and 6430 (Trusts, funds and similar financial entities) are excluded. The list of NACE sectors is taken from here: https://ec.europa.eu/competition/mergers/cases/index/nace_all.html. Source: Authors' elaborations on Orbis and TAM data.

4. Methodology

4.1. Selection of candidate predictors

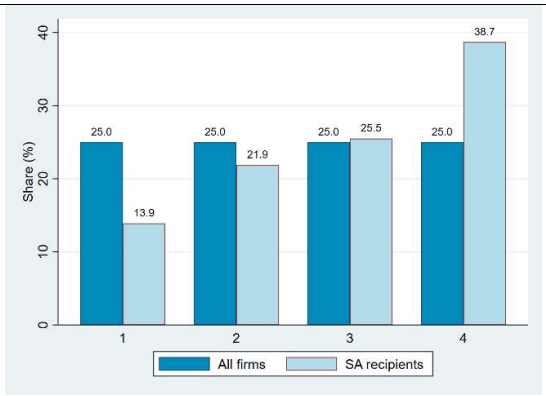
- When it comes to select the set of potential predictors, there are two possible strategies. The former is a fully data-driven approach, with no ex ante choice made by the analyst. It is fully in line with the ML spirit, but at the same time, it does not necessarily preserve interpretability and transparency. According to the latter strategy, the analyst makes an ex ante selection on the basis of one or more of the following criteria: (i) admissibility of the aid according to prevailing competition rules; (ii) accounting rules; (iii) economic effects of grants as highlighted in the settled literature. In what follows, we discuss the pros and cons of choices (i)-(iii).

Figure 2: Sample distribution by size

Panel A: Share of the number of firms by quartiles of operating revenue turnover	Panel B: Share of the number of firms by quartiles of total assets
	

Notes: The sample consists of SA recipients in 2016-2017. Firms with an annual granted aid less than 500,000 euros are not included among SA recipients. Source: Authors' elaborations on Orbis and TAM data.

Figure 3: Sample distribution by performance

Panel A: Share of the number of firms by ROA quartiles	Panel B: Share of the number of firms by quartiles of labour productivity
	

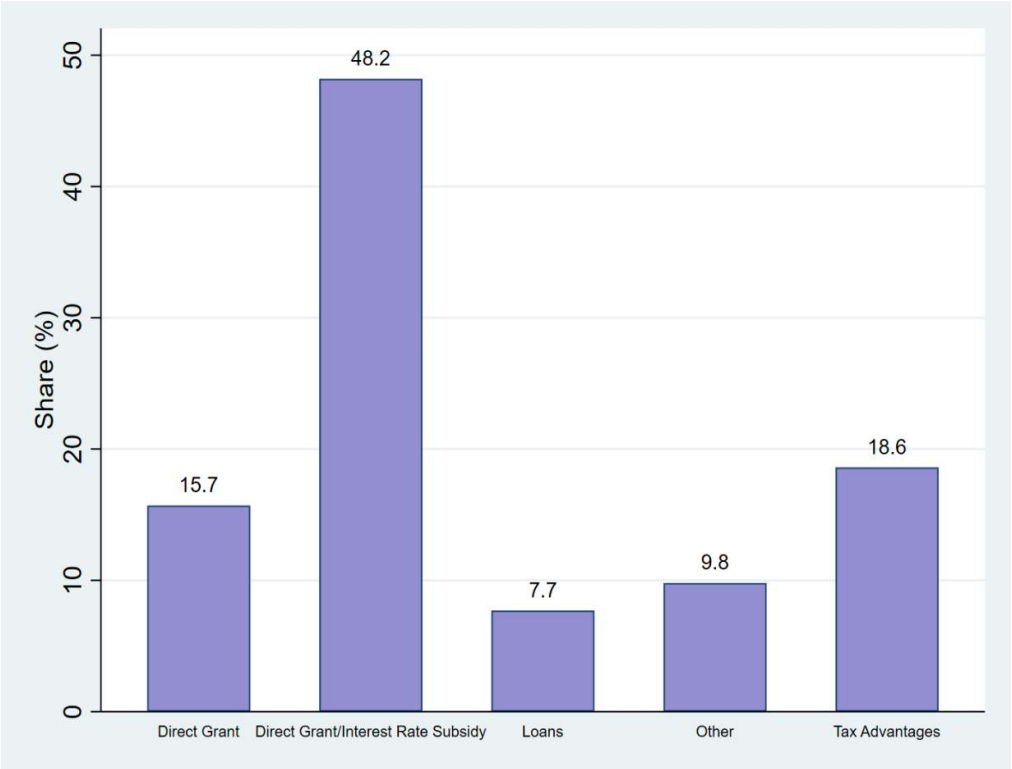
Notes: The sample consists of SA recipients in 2016-2017. Firms with an annual granted aid less than 500,000 euros are not included among SA recipients. In Panel A, Variable 'Return on Assets' is the ratio between net income and total assets, multiplied by 100. In Panel B, variable 'Labour productivity' is computed as value added per employee. Source: Authors' elaborations on Orbis and TAM data.

- Let us start from (*i*): admissibility of the aid according to prevailing competition rules. State aids are usually prohibited and are allowed only under special circumstances. Hence, if one knows the exception rules and is able to operationalize them thanks to available data, then that variable would be a powerful predictor for grants. For example, suppose that an exemption is given to all firms in a

given industry because of an idiosyncratic sectoral shock that hit all firms belonging to that industry (vertical aid). In such a case, a simple dummy variable for that industry would work very well as a predictor. The main problem with such an approach is that, by definition, it is not able to detect aids that are given without compliance with the existing set of rules, and contradicts the target of the project. A minor problem is that it can be difficult to have a good/satisfying operationalization of exemption rules. For these reasons, this approach will be discarded.

Figure 4: Share of SA recipients by type of aid instrument

Share of recipient firms by instrument type



Notes: The sample consists of SA recipients in 2016-2017. Firms with an annual granted aid less than 500,000 euros are not included among SA recipients. Source: Authors' elaborations on Orbis and TAM data.

- (ii) Accounting rules. The accounting standards usually outline outlines how to account for government grants. For example, an aid can increase turnover/revenues or can be deducted from the cost it alleviates (if any). This means that, in principle, if one knows the accounting rules and data on aids are detailed enough, one can map each aid in a balance sheet / financial statement item and use it as a predictor. However, this method has a number of shortcomings. First, there is no credible estimate of the degree of compliance of Orbis data with International Accounting Standards that would provide a unified framework; second, in many cases, the accounting principle does not prescribe a unique rule, as in the case mentioned above (the aid inflates turnover or decreases costs?) and the real application is discretionary. Third, this approach is potentially meaningless if undertakings receive undeclared aids by means of intragroup operations or of other "under the line" operations. For these reasons, this approach will be discarded.

- (iii) Economic effects of grants as highlighted in the settled literature. **The basic idea is that if subsidies have, say, a positive effect on outcome X_t (e.g. turnover), then observing an increase in X (that is a positive $\Delta X_t = X_t - X_{t-1}$) might signal the existence of an aid received in year t**. However, measuring the causal impact of state aid on improvements in firm performance is not a simple empirical exercise. The main inferential problem is that it is usually very difficult to have a satisfying control group: if treated firms are different from control ones in some unobserved characteristics correlated to the outcome, then differences in the outcome can hardly be imputed to the treatment. The ideal solution would be running a randomized control trial, where firms are randomized to receive the state aid, and changes in their performance are compared to that of a control group (similar firms that did not receive the aid). On the other hand, in many circumstances, this gold standard is unfeasible, and scholars had to resort to quasi-experimental methods like regression discontinuity design, propensity score matching, difference-in-differences (or their combinations).
- The literature review, summarized in Table A3, does not offer a sound and unambiguous guide for variable selection for a number of reasons: (i) results changes with the type of grants (for example R&D funds should have an impact on innovation while Rescue&Restructuring funds should shape long-term survival); (ii) even if most of the studies explicitly adopt a counterfactual approach, it is not obvious how to discriminate between papers of different quality and how to use such a discrimination; (iii) even within the same type of grant and the same type of empirical approach, meta-regression analysis shows a lack of conclusiveness; (iv) even if one was fully confident that the grant has a credible causal effect on outcome X, X may well be correlated with Y, Z, etc. (letting alone the causal effect of the aid on Y, Z, etc.): these regressors, in principle, should be included in the set of potential predictors; (v) on the other hand, it is far from being obvious that some economic effect in the counterfactual sense is necessarily the best potential predictor, given the predictive nature of ML tools; (vi) further limitations come from data availability: for example, survival analysis requires a quite long time span after receiving the grant and such long span is not currently available; R&D grants should shape Research & Development expenses that in Orbis has many missing values.
- It follows that, instead of reconstructing a complex mapping of each type of grant on various outcomes according to the literature review, it is much more promising to focus on the take-home message of the literature review: the more general idea that grants can improve firm performance, which in turn can be measured in many ways. More in detail, a well-balanced approach is the following two-step procedure that combines a data-driven method with an ex ante selection on potential predictors. **This approach is in the spirit of the machine learning literature in economics that emphasizes the need for a feature selection process based on a criterion of domain knowledge**. It is designed to be fully consistent with the ML spirit ("let the data speak"), but, at the same time, is intended to preserve some transparency. The steps are as follows:
    - Assume that funds granted in year t have an impact on the dynamics of X, Y, Z, … (e.g. turnover, employment, investments, value added per employee, Financial Expenses). Select a large number of outcomes so to make the ML algorithm work better. Then $X_{t-1}, X_t, Y_{t-1}, Y_t, Z_{t-1}, Z_t$, … are potential predictors. As stated above, at the current stage, we summarize all lagged values for each Orbis predictors X by means of the 'pre' transformation ($\frac{X_{t-1}+X_{t-2}}{2}$); analogously leaded values are summarized by means of the 'post' transformation ($\frac{X_t+X_{t+1}}{2}$). As stated above, we prefer to include the difference ($\frac{X_t+X_{t+1}}{2} - \frac{X_{t-1}+X_{t-2}}{2}$) as potential predictors, whose full list is in Table A1.
    - Run the ML algorithm on the variables selected above.

4.2. Dealing with GUO information

- As stated above, a potentially interesting piece of information is about ownership. In fact, unrecorded grants might go to the final beneficiary from another firm in the same group by means of infra-group operations.
- First, we review GUO data characteristics and, then we move to how to detect group-level channels.
- GUO data characteristics.
    - Bureau van Dijk's definition of Global ultimate owner (GUO): individual or entity at the top of the corporate ownership structure. From this definition, it stems that:
        - Being a GUO does not necessarily imply being either a controlling owner or a beneficiary owner. Bureau van Dijk has a proprietary algorithm that it uses to build a tree and to assign GUO using two levels of shares, i.e. >25% (minority link) and >50% (majority link). At the top of the tree, there can be an individual or entity. Individuals have no account related to them.
        - For the purpose of this project, the EC selected a firm as a corporate GUOs (an entity), following the majority (>50%) links between companies.
        - There is no detailed information on the tree structures. We only know the GUO of each individual company.
        - The requirement to publish consolidated accounts is determined by reporting obligation. For example, if a firm is a subsidiary listed on the stock exchange, it will have to publish a consolidated account along with an unconsolidated one. GUOs are assumed to consolidate across the whole group.
        - Consequently, we considered as GUOs all firms reporting consolidation codes equal to C1 or C2.
- Using GUO information.
    - There are two potential approaches to include GUO information in our setting: (i) run an additional machine learning analysis at the GUO level; (ii) **include GUO level data in the main Orbis-TAM dataset, and run only a single machine learning analysis by including both firm-level features and the GUO level ones as predictors**.
    - We opted for the latter solution, mainly because we thought it would be a better idea to see if the machine learning algorithms, when provided both with firm level predictors and GUO level ones, select the latter ones. This is more consistent with a data-driven approach: if GUO level variables are picked up by the algorithms, this would suggest 'hidden' infra-GUO operations.
    - This possibility will be tested in the subsequent analyses when we will feed the models with the full set of firm and GUO level predictors.
    - Consequently, the final dataset is augmented with the same sets of variables at the GUO level (see Figure A1, Panel B).

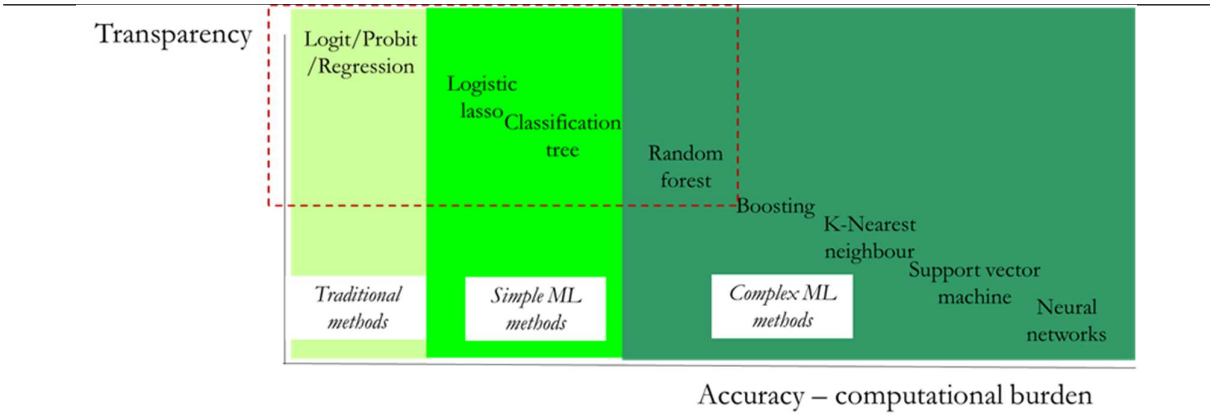
4.3. ML techniques

- ML techniques have been developed in computer science and statistical literature to deal with predictive tasks (Varian, 2014). In social sciences, they are rapidly becoming a key tool for the study of the so-called 'prediction policy problems' (Kleinberg et al., 2015). Their main focus is thus on the minimisation of the out-of-sample prediction error, with the ultimate aim of generalising well on future unseen data (Athey, 2018; Athey and Imbens, 2017; Mullainathan and Spiess, 2017).
- This project aims at: (i) developing a good predictive model for SA recipients; consequently, (ii) providing insights on the potential underreporting of SA recipients.
- As to goal (i) (developing a good predictive model for SA recipients), ML techniques allow us to provide a model able to effectively identify firms receiving SAs. **The core output of this model is**

a confusion matrix like that sketched in Figure A2. This 2x2 matrix crosses the actual SA recipient dichotomous status and the predicted one. **We want to maximise sensitivity, that is, the proportion of positive cases** (N_{22}, corresponding to SA recipient = 1) correctly identified (the so-called true positives). This will make us confident that the algorithm does a good job at predicting the recipient status: the higher the sensitivity, the higher the ability of the model to correctly classify SA recipients.

- As to goal (ii) (providing insights on the potential underreporting of SA recipients), the cell of main interest will become N_{21}, i.e., the false positives. These are firms that do not officially benefit from SAs but are predicted to do so by the algorithm because their characteristics are "similar" to those of firms with SAs. For example, suppose that the underlying data generation process is such that the SA implies a strong increase in turnover. Then a good prediction model should pinpoint those firms (i) which do not receive grants but (ii) whose strong increase in turnover is consistent with the SA recipient status.
- By definition, N_{21} will comprise a mix of prediction error and SA underreporting. **The lower the sensitivity, the higher the share of observations in N_{21} that are merely due to prediction error**, i.e., the scarce ability of the model to identify SA recipients, rather than to underreporting. On the opposite, the higher the sensitivity, the higher the share of observations in N_{21} that are due to underreporting. Therefore, maximizing sensitivity minimizes the component of N_{21} which is pure prediction error so that N_{21} can mainly be attributed to underreporting.
- After estimating a model with satisfying properties, N_{21} can be used to (i) estimate an upper bound to the overall size of the underreporting, (ii) use it as a preliminary screening device preceding further investigation, and (iii) describe the main characteristics of this type of firms.
- In principle, confusion matrix values can be estimated in a number of ways, from traditional logit/probit models to more advanced ML models.
- **The use of ML techniques comes with a key concern regarding the trade-off between accuracy and interpretability**. While simpler techniques tend to be more transparent and easier to understand, they might result in inferior performance compared to more complex algorithms. On the opposite, more 'black-box' methods tend to be more accurate but less, if at all, interpretable. So, the choice of the most appropriate technique depends on the problem under scrutiny. In our case, in which ML is used in the service of public policies, we prefer to take into consideration communication and accountability aspects. In Figure 5, we plot a stylized scatterplot (adapted from Hastie et al., 2009) in which each method is placed along the spectrum of the accuracy-transparency trade-off. The algorithm we employ below for the analysis, the classification trees, are suited for applications in which the decision rule needs to be transparent (Lantz, 2019), such as when the output of the model must be shared in order to facilitate decision making (Andini et al., 2018).
- **The standard ML routine consists of randomly splitting the sample into two separate subsets, a training set and a testing (hold-out) sample**. This routine stems from the need to apply a so-called 'firewall' principle: none of the data involved in fitting the prediction function is used to evaluate the prediction function that is produced (Mullainathan and Spiess, 2017). The split has to be necessarily random so to avoid including systematic differences between the two separate sets (Lantz, 2019). Also, the predictive model has to be evaluated on the testing set, because measures of performance evaluated on the training set tend to be typically overoptimistic with respect to the true model performance, as the algorithm is evaluating the model on data it has already learned from. The better practice is thus to evaluate a model's performance on data it has not yet seen. As for the random split, we employ the conventional choice of using 2/3 of the dataset as the training set and the remaining 1/3 of observations as the testing set (Hastie et al., 2009). We use the training set to train and tune our algorithms, and the testing set to estimate their future performance.

Figure 5: Trade-off between accuracy and transparency



- **We focus on the following ML methods: logistic lasso, classification tree, random forest**. The other methods have been excluded because to preserve a minimum amount of transparency and to avoid excessive computational burden that, in the case under scrutiny, might be too cumbersome. ML approaches will be contrasted to a benchmark logit estimation to see whether the differences in performance are significant.
- The descriptive statistics presented in Table 1 suggest the existence of a **severe data imbalance problem**: SA firms account for a very small part of the sample. Imbalanced datasets can fatally disrupt the performance of any algorithm applied to a predictive exercise, be it an ML method or not. This is because, in the case of imbalanced datasets, predictive algorithms run into the so-called "accuracy paradox": they provide predictions featured by a very high out-of-sample overall accuracy (even greater than 90%), but totally useless for practical purposes, because simply always predict the overrepresented label (SA = 0, in our case). Using the ML jargon, predictive exercises on imbalanced datasets result in a very high specificity (i.e., true negatives) but an extremely low, if not null, sensitivity (true positives, i.e. SA recipients, our key category of interest). Therefore, before performing our classification task, we need to tackle the challenge stemming from our highly imbalanced dataset. We employ a well-known solution: rebalancing the training set. Specifically, **we make use of the Synthetic Minority Oversampling Technique (SMOTE) routine** developed by Chawla et al. (2002) to rebalance the two classes in our training sample. SMOTE is an algorithm that oversamples the under-represented cases and undersamples the majority class, leading to a much smaller but rebalanced dataset.
- Crucially, we implement the SMOTE algorithm only on the training subsample, i.e., the set on which when train our model, leaving the testing sample, i.e., the set on which we evaluate its future performance on unseen data, untouched. This means that the training dataset is artificially balanced over the two outcomes, while the prediction is tested on the original skewed sample, i.e., on real-world data.
- After rebalancing our training dataset, the two outcomes are perfectly balanced between the two classes, and the sample size is sharply reduced due to the undersampling of the majority class. On these rebalanced data, we then apply our algorithms, whose results are provided in the next Section.

Finally, a key point to underline is the missing data problem, which is pervasive in the Orbis data (nor missing data can be found elsewhere). As it will be clearer in Section 5, within this project, **the**

classification tree is the best method because it automatically handles missing data, thanks to the use of surrogate splits in the case of missing observations. Conversely, all the other methods suffer from a drastic reduction in sample size.[4]

5. Results

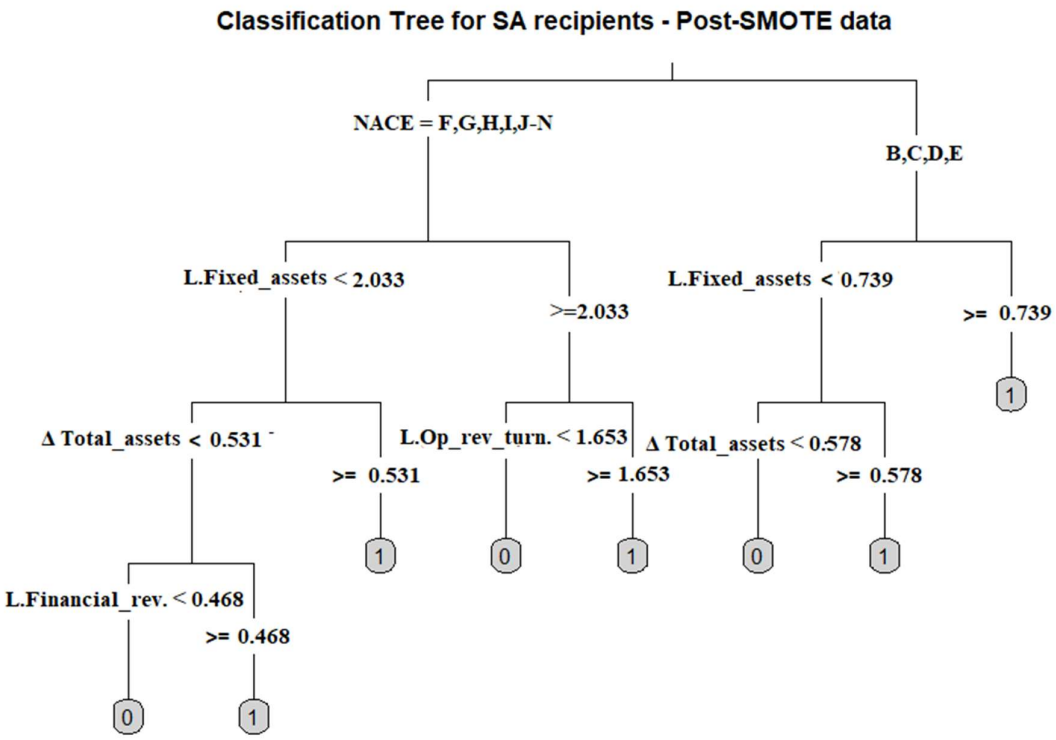
- In what follows, we present results stemming from the following models: classification tree, logit, logistic lasso, random forest. Each model is estimated with two different cutoffs separating SA recipients from other firms: 500,000 euros ("all recipients sample") and 1,500,000 euros ("large recipients" sample"). We fed the algorithms with two separate sets of firm and GUO level features, both including the quantitative Orbis variables in their 'pre' and 'delta' forms.[5] Finally, we also included dummies for four geographic areas and five aggregate NACE sectors, consistently with those reported in Figure 1. As stated above, outliers are excluded by trimming the sample at the 1st and 99th percentiles.
- Before looking at the results, we clarify some aspects involving the methodological choices implemented. The classification tree splits the data into smaller and smaller subsets to identify important patterns that can be employed for predicting a qualitative outcome. It is an extremely flexible method because it can easily capture non-linearities and interactions among predictors through the sequence of splits. Albeit one could grow a very complex tree, large enough so that no observation is misclassified, in practice, a high number of levels in a tree is likely to result in high variance and to overfit the data, leading to a predictive model with poor out-of-sample performance. This is why regularization, via the so-called tree 'pruning' procedure, is used to tune the algorithm and prevent the risk of in-sample overfitting. **Pruning means reducing the complexity of the tree by setting a penalization cost for flexibility**; this cost takes the name of 'complexity parameter' (*cp*). In order to select the optimal value of *cp*, which maximizes the out-of-sample accuracy of our model, we employ 10-fold cross-validation for model selection in the training dataset, compare the ten resulting cross-validation errors, and pick up the complexity parameter associated with the lowest cross-validation error. This *cp* is then selected for the model used to predict unseen observations belonging to the testing set and evaluate its performance on the unseen held-out data.
- Please also remind that all the algorithms whose performance we are about to discuss have been trained and tuned on the 'smoted' training data, i.e. after applying the rebalancing SMOTE algorithm discussed in the methodological Section. Without applying SMOTE on the training data, all these algorithms would simply always predict the overrepresented class (SA=0), resulting in predictive performances with a very high overall accuracy but null sensitivity, making them useless for the purpose at hand.
- We start from **the classification tree, which will be our core model** for reasons that will be clear soon. Figure 6 shows the estimated tree for the 'all recipients' model (that is, the cutoff for SA recipients is set at 500,000 euros). Out of about 190 potential predictors, the model selects just five variables (sector, 'pre' values for fixed assets, operating revenue turnover, and financial revenues, 'delta' value for total assets, combined in a non-linear manner. No GUO level variable is selected, probably also because these variables display many missing values, resulting in scarce predictive

[4] In principle, the random forest could also overcome the missing data issue via preliminary imputation through a proximity-based random forest algorithm (*rfImpute* package in R) that generates imputed data via a nearest-neighbor approach. After applying this routine, one would then have to run the standard random forest predictive algorithm. We tried to use the *rfImpute* routine, but the system crashed, due to the extremely computationally-cumbersome task of imputing millions of missing values for many different predictors.

[5] Unfortunately, as it is explained below, the missing data issue when including also GUO level features was so severe that only the classification tree was able to run, while, in order to run logit, LASSO, and the random forest, the only viable solution has been to exclude GUO level predictors.

power. The corresponding confusion matrix in reported in Table 3: out of more than 3.6 million of form-year observations in the testing sample, near 87% are correctly classified. Nicely, **the sensitivity of the model (the percentage of positives that are correctly predicted) is very high (81%). 13.2% of observations are false positives: they are not SA recipients but are predicted to be so**. Mind that such false positive subsample captures both 'true' hidden recipients and the prediction error, and that there is no way to disentangle between these two components. However, the very good predictive performance of the model in terms of both accuracy and sensitivity reassures that the latter component is 'small'.

Figure 6: Classification tree – All recipients



Notes: Visual output of a classification tree generated on the full combined Orbis-TAM databases. Such dataset was divided into a training set consisting of 2/3 of the observations and a testing set consisting of the remaining 1/3. Next, the original unbalanced training set was rebalanced using the SMOTE algorithm. The model was trained and tuned (via 10-fold cross-validation) on the artificially rebalanced training set, and its predictive performance was evaluated on the original (unbalanced) testing set. The sample consists of total annual aids granted to each recipient firm for the years 2016 and 2017. SA recipient is a binary variable taking value 1 if the firm is a State Aid recipient and 0 otherwise. Predictors include two separate sets of firm and GUO level features, both including all the quantitative Orbis variables. These Orbis variables are included in two forms: pre-SA averages of the two years $(t-1, t-2)$ before the SA was received; differences (i.e. deltas) between the pre-SA averages of the two years $(t-1, t-2)$ before the SA was received - and post-SA averages of the values for the year in which the SA was granted and the year after that $(t, t+1)$. Before the random splitting between training and testing sets, the sample was trimmed at the 1st and 99th percentiles to drop outliers in these Orbis predictors. Dummies for four geographic areas and five aggregate NACE sectors are also included. . Variable 'L.Fixed_assets' refers to the average value of Fixed Assets for the two years before the SA was granted. Variable 'ΔTotal Assets (post)' refers to the difference in the values of variable Total Assets between the pre-SA averages of the two years $(t-1, t-2)$ before the SA was received - and post-SA averages of the values for the year in which the SA was granted and the year after that $(t, t+1)$. Variable 'L.Op _rev_turn.' refers to the average value of Operating Revenue Turnover for the two years before the SA was granted. Variable 'L.Financial_rev' refers to the average value of Financial Revenue for the two years before the SA was granted. All the values are expressed in millions of euros.

Table 3: Classification tree out-of-sample performance – All recipients

		Real status		
		SA recipient = 0	**SA recipient = 1**	**Total**
Predicted status	**SA recipient = 0**	3,161,501	510	3,162,011
	SA recipient = 1	481,423	2,174	483,598
	Total	3,642,924	2,684	3,645,608
	Correctly predicted	**86.8%**	**81%**	**86.8%**

Notes: Output of a classification tree generated on the full combined Orbis-TAM databases. Such dataset was divided into a training set consisting of 2/3 of the observations and a testing set consisting of the remaining 1/3. Next, the original unbalanced training set was rebalanced using the SMOTE algorithm. The model was trained and tuned (via 10-fold cross-validation) on the artificially rebalanced training set, and its predictive performance was evaluated on the original (unbalanced) testing set. The sample consists of total annual aids granted to each recipient firm for the years 2016 and 2017. SA recipient is a binary variable taking value 1 if the firm is a State Aid recipient and 0 otherwise. Predictors include two separate sets of firm and GUO level features, both including the quantitative Orbis variables. These Orbis variables are included in two forms: pre-SA averages of the two years $(t-1, t-2)$ before the SA was received; differences (i.e. deltas) between the pre-SA averages of the two years $(t-1, t-2)$ before the SA was received - and post-SA averages of the values for the year in which the SA was granted and the year after that $(t, t+1)$. Before the random splitting between training and testing sets, the sample was trimmed at the 1st and 99th percentiles to drop outliers in these Orbis predictors. Dummies for four geographic areas and five aggregate NACE sectors are also included.

- Tables A4, A5, A6 reports the confusion matrix for the logit, lasso, and random forest model, respectively. It is worth noting that the size of the testing sample is drastically reduced with respect to Table 3 (to around 1/25). This is due to the fact that because of the huge number of missing values in the Orbis data, which logit, lasso, and random forest models cannot manage (the missing observations are simply dropped). The classification tree, on the contrary, uses a sophisticated algorithm that allows overcoming the missing data problem. As reported in the methodological Section, a similarly sophisticated algorithm is available for the random forest too, but its practical implementation with the (big) data at hand turns out to be too cumbersome from a computational viewpoint. Due to the severe missing data issue, GUO level predictors were excluded from the analysis for these three algorithms, because including them would result in such a low number of complete observations that the outcome variable had only 1 level, making it impossible for them to run. **Importantly, note that such a high proportion of missing values is due to different reporting requirements by Member States as well as to the different scope of reporting by firm size, and that such limitations are intrinsic to accounting data.** The dramatic fall in the number of observations makes the performance analysis much less important. In any case. all the models in Tables A4, A5, A6 show much lower sensitivity than the one reached by the classification tree. For all these reasons, the classification tree is the core model of the ML analysis.
- Figure A3 replicates the classification tree analysis for the 'large recipients' sample (that is, the cutoff for SA recipients is equal to 1,500,000 euros). There are four selected predictors: sector, 'pre' values for operating revenue turnover, tangible fixed assets, operating profit). Table A7 shows the classification tree. As expected, the sample size does not suffer from the missing value problem. The sensitivity parameter is 72.6%, quite lower than that in Table 3. This drop should not be understated because, in order to minimize the prediction error component of the 'false positive' cell, reaching high sensitivity is critical. Our subjective assessment is that the sensitivity threshold to be reasonably confident that observations included in the 'false positive' cell point to 'suspect' firms rather than to

model's mistakes, should be 80 %. For these reasons, **the preferred specification is that one obtains by setting the cutoff at the lower level (500,000 euros)**. Having said that, conditional on reaching a sufficiently high sensitivity (>=80%), the final choice of the sample used to train the classification tree depends on the focus of the policy interest. For the sake of completeness, Tables A8, A9, A10 reports the confusion matrix for the logit, lasso, and random forest model, all trained with the 'large recipients' sample: in all cases, missing values drastically limit the analysis. As for the analysis of all recipients, GUO level predictors are excluded from the logit, lasso, and random forest models.

6. Which firms are most likely to be false positive?

- Section 5 suggests that the best model is a classification tree on the 'all recipients' sample (Figure 6), whose corresponding confusion matrix (Table 3) indicates that 13.2% of observations are classified as false positives (or 'suspect' recipients). This Section provides a descriptive analysis of them, compared to the non-false positives. The covariates are: (i) country dummies, (ii) Nace 2-digit sector dummies, (iii) a dummy for large firms (those with total assets greater than the median); (iv) dummies for European vs. non-European GUOs. Variables (i)-(iv) are likely to capture a vast portion of firm heterogeneity?
- Univariate analysis is shown in Figures A4-A7. Each figure depicts the false positive status after conditioning to a particular value of the regressor. Results in Figure A4 suggest a huge variability across countries. In a number of cases, the incidence of suspect recipients is near or above 20% (e.g. Cyprus, Greece, Latvia, Germany), much larger than the average value (13.2%); on the other hand, units from Estonia, Hungary, Slovenia, Bulgaria show a lower share of false positive. The heterogeneity is even larger across sectors (Figure A5). On average, firms in industrial sectors are much more likely to be predicted as hidden recipients. In some cases, the incidence is larger than 60%. Service sectors tend less to be predicted in the 'false positive' cell. Firm size is also highly correlated to the 'suspect recipient' status: the incidence is 28.6% for large firms, nearly zero for the other ones (Figure A6). Finally, Figure A7 illustrates the false positive status by GUOs type: firms with non-missing GUOs are more likely to be hidden recipients (strongly for non-European GUOs), but the very small sample size underlying this evidence suggests to be very cautious in using GUOs nationality for any descriptive analysis of false positives.
- The main limitation of the above-described univariate descriptive analysis is that it does not take into account composition effects. For example, if large firms are disproportionally located in a given country, that country will also rank in top positions not for the 'country' effect but simply because of size. To overcome this difficulty, Table A11 shows the output of a multivariate logit regression in which the covariates are dummies for size, countries, and sectors (GUO's nationality is not included not to drastically reduce sample size). The reported coefficients are marginal effects capturing the correlation between each covariate and the probability of being a false positive, while controlling for the other regressors. To escape the collinearity trap, 'Spain' and 'Accommodation' are the excluded country and sector, respectively. They have been chosen because, according to the univariate analysis shown above, in both cases, the incidence of false positives is near the average: Spain = 14.1%, Accommodation (Nace 55) = 13.0%, average value = 13.2%.
- **The multivariate analysis confirms the predominant role of size (Table A11)**: large firms (those whose total assets are larger than the median) display a probability of being hidden recipients that is 27 percentage points (pp) larger than small firms. Heterogeneity across countries persists. **The top 5 countries showing the largest estimated marginal effect are Cyprus (in this case, the probability of being a false positive increases by 13.5 pp, all else being equal), Luxembourg**

(9.5 pp), Austria (8.2 pp), Poland (7.1 pp), Lithuania (5.8 pp). No country exhibits large negative differences with respect to Spain. Mind that in the univariate analysis, Spain was near the average while now it ranks low, signaling that correcting for composition effects is very important. **The top 5 sectors showing the largest estimated marginal effect are 'Water collection, treatment and supply' (Nace 36, 24.4 pp), 'Electricity, gas, steam and air conditioning supply' (Nace 35, 20.1 pp), 'Manufacture of tobacco products' (Nace 12, 20.1 pp), 'Manufacture of basic pharmaceutical products and pharmaceutical preparations' (Nace 21.0 pp), 'Manufacture of beverages' (Nace 11, 19.8 pp)**. Table A11 also shows that a large number of sectors, other than the top 5, are associated with significantly different probabilities to be in the 'false positive' group. On the other hand, firms in other sectors such as, for example, 'Veterinary activities' (Nace 75), 'Legal and accounting activities' (Nace 69), 'Food and beverage service activities' (Nace 56), are significantly less likely to be suspect.

7. Potential future steps

- The final output of the classification tree is a label for suspect observations: firms that in 2016 or in 2017 did not receive any grant from an EU country but whose observable characteristics were very close to those of SA recipient firms. In the best model estimated above, this label is switched on for 13.2% of the testing sample. This Section briefly discusses potential future steps for the analysis.
- A first line of development starts from the fact that the total number of suspect recipients is 'high': the label can be considered as a preliminary screening device. The next step is to carry out a more focused inspection by means of quasi-manual work involving soft information, other sources, further micro-level analysis of financial accounts, etc. This requires that the size of the hidden recipients must be drastically reduced. Such reduction could be achieved as follows:
    - Random sampling on the false-positive group.
    - Estimate different intensities of being predicted as false positive by means of a probabilistic classification tree, if feasible, and use them to prioritize further investigations.
    - Use of some crude rules such as: (i) excluding from them suspect sample firms that received small (below 500,000 euros) grant, and/or are small enough not to be a concern for competition in the internal market, etc.; (ii) keep only firms involved in M&A operations, taken from the Zephyr database (managed by Bureau Van Dijk); (iii) …
- A second line has to do with the original data problem underlying this study: ideally, one would want to have had data on firms receiving public grants from abroad. Training the model with firms taking grants from countries belonging to the internal market is a second best, that is reasonable under the assumption that are no systematic differences between these two groups of firms. This assumption is not empirically testable, given the current data availability. In order to overcome this data problem, it would be possible to run a small ad hoc survey, for example, by means of national statistical institutes, which usually run surveys on firms.

8. Conclusions

- This report analyses whether and how recipients of EU SAs can be accurately predicted by applying machine learning (ML) tools to financial accounting data.
- The empirical analysis exploits the combination of financial accounting data taken from the Orbis database and data on SA recipients from the TAM database, managed by the EC.

- Micro firms, as well as firms located in the UK or in some heavily regulated sectors (such as agriculture, banking and finance, non-market service sectors), have been excluded from the analysis.
- The best ML model (a classification tree) predicts that 13.2% of cases can be classified as suspect recipients, i.e. firms that in a certain year did not receive public grants, but are predicted to have done so. The descriptive analysis of these firms shows some clear patterns in terms of size, country, and sector.
- This ML analysis represents the first step for further investigations.

References

Andini M, Ciani E, de Blasio G, D'Ignazio A, Salvestrini V (2018), Targeting with machine learning: An application to a tax rebate program in Italy, *Journal of Economic Behavior & Organization*, 156, 86-102.

Athey S (2018), The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda* (A Agrawal, J Gans, A Goldfarb editors), University of Chicago Press.

Athey S, Imbens G W (2019), Machine learning methods that economists should know about, *Annual Review of Economics*, 11, 685-725.

Bajgar M, Berlingieri G, Calligaris S, Criscuolo C, Timmis J (2020), Coverage and representativeness of Orbis data, *OECD Science, Technology and Industry Working Papers*, 2020/06.

Becker B (2014), Public R&D Policies and private R&D investment: a survey of the empirical evidence, *Journal of Economic Surveys*, 29, 917–942.

Bronzini R, Piselli P (2016), The impact of R&D subsidies on firm innovation, *Research Policy*, 45, 442-457.

Buts C, Jegers M (2013), The Effect of 'State Aid' on Market Shares: An Empirical Investigation in an EU Member State, *Journal of Industry, Competition and Trade*, 13, 89-100.

Cerqua A, Pellegrini G (2014), Do subsidies to private capital boost firms' growth? A multiple regression discontinuity design approach, *Journal of Public Economics*, 109, 114-126.

Chawla N V, Bowyer K W, Hall L O, Kegelmeyer W P (2002), SMOTE: synthetic minority over-sampling technique, *Journal of artificial intelligence research*, 16, 321-357.

Colombo M G, Croce A, Guerini M (2013), The effect of public subsidies on firms' investment–cash flow sensitivity: Transient or persistent?, *Research Policy*, 42, 1605-1623.

Criscuolo C, Martin R, Overman H G, Van Reenen J (2019), Some Causal Effects of an Industrial Policy, *American Economic Review*, 109, 48-85.

Czarnitzki D, Lopes-Bento C (2013), Value for money? New microeconometric evidence on public R&D grants in Flanders, *Research Policy*, 42, 76-89.

Dimos C, Pugh G (2016), The effectiveness of R&D subsidies: A meta-regression analysis of the evaluation literature, *Research Policy*, 45, 797-815.

Duso T, Nardotto M, Seldeslachts J (2021), A Retrospective Study of State Aid Control in the German Broadband Market, *CEPR Discussion Paper Series*, DP15779.

European Commission (2018), Commission's staff paper: encoding information in the Transparency Award Module for State aid.

European Commission (2020), White paper on levelling the playing field as regards foreign subsidies.

Grigolon L, Leheyda N, Verboven F (2016), Scrapping subsidies during the financial crisis – Evidence from Europe, *International Journal of Industrial Organization*, 44, 41-59.

Hastie T, Tibshirani R., Friedman J (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

Heim S, Hüschelrath K, Schmidt-Dengler P, Strazzeri M (2017), The impact of state aid on the survival and financial viability of aided firms, *European Economic Review*, 100, 193-214.

Hyytinen A, Toivanen O (2005), Do financial constraints hold back innovation and growth? Evidence on the role of public policy, *Research Policy*, 34, 1385-1403.

Kalemli-Özcan Ş, Sørensen B, Villegas-Sanchez C, Volosovych V, Yeşiltaş S (2019), How to Construct Nationally Representative Firm Level Data from the Orbis Global Database, *National Bureau of Economic Research Working Paper*, 21558.

Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z (2015), Prediction policy problems, *American Economic Review*, 105, 491-95.

Lantz B (2019), *Machine learning with R: expert techniques for predictive modelling*, Packt Publishing Ltd.

Mullainathan S, Spiess J (2017), Machine learning: an applied econometric approach, *Journal of Economic Perspectives*, 31, 87-106.

Sergant I, Van Cayseele P (2019), Financial Constraints: State Aid to the Rescue? Empirical Evidence from Belgian Firm-Level Data, *Journal of Industry, Competition and Trade*, 19, 33-67.

Szücs F (2020), Do research subsidies crowd out private R&D of large firms? Evidence from European Framework Programmes, *Research Policy*, 49, 103923.

Varian H R (2014), Big data: New tricks for econometrics, *Journal of Economic Perspectives*, 28, 3-28.

Appendix: additional tables and figures

Table A1: Orbis and TAM variables

Panel A: Orbis variables provided by the EC			
ADDED_VALUE	CURRENT_LIABILITIES	MATERIAL_COSTS	ORIGINAL_UNITS
BVD_ID_NUMBER	DATE_OF_INCORPORATION	METROPOLITAN_AREA	OTHER_FIXED_ASSETS
CAPITAL	DEBTORS	NACE_CORE_CODE.	P_L_AFTER_TAX
CASH_FLOW	DEPRECIATION_AND_AMORTIZATION	NAME_INTERNAT	P_L_BEFORE_TAX
CITY	EBITDA	NAME_NATIVE	P_L_FOR_PERIOD_NET_INCOME
CLOSING_DATE	FILING_TYPE	NATIONAL_LEGAL_FORM	REGION_IN_COUNTRY
CONSOLIDATION_CODE	FINANCIAL_EXPENSES	NON_CURRENT_LIABILITIES	SHAREHOLDERS_FUNDS
COSTS_OF_EMPLOYEES	FINANCIAL_P_L	NUMBER_OF_EMPLOYEES	STANDARDISED_LEGAL_FORM
COUNTRY	FIXED_ASSETS	NUMBER_OF_MONTHS	STATE_OR_PROVINCE
COUNTRY_ISO_CODE	INTANGIBLE_FIXED_ASSETS	OPERATING_P_L_EBIT	STOCK
CREDITORS	INTEREST_PAID	OPERATING_REVENUE_TURNOVER	TANGIBLE_FIXED_ASSETS
CURRENT_ASSETS	LOANS	ORIGINAL_CURRENCY	TAXATION
Panel B: Other Orbis variables (computed ex post)			
ACE	FIRE	ROA	SOLR
CURR	GEAR	ROE	TOAS
EBMA	LIQR	RSHF	TSHF
ETMA	PRMA	SOLL	ValueAddedperemployee
Panel C: TAM variables provided by the EC			
AidAwardCreatedDate	AidAwardObjective	BeneficiaryNameEnglish	GrantedRangeEUR
AidAwardGAEnglish	AidAwardObjectiveOtherEngli	BeneficiaryRegion	IsCoFinance
AidAwardGAOriginal	AidAwardPublishedDate	BeneficiarySector	MainProcedureTypeCode
AidAwardGrantedDate	AidAwardReference	BeneficiaryType	NationalIdentification
AidAwardInstrument	AidAwardStatus	CaseReference	NationalIdentificationType
AidAwardInstrumentOtherEngl	BeneficiaryCountry	CaseTitleOriginal	NominalAidAbsoluteEUR
AidAwardNutsCode.	BeneficiaryName	GrantedAidAbsoluteEUR	

Table A2: Main Orbis variables by SA status

Panel A: SA recipients

Variable name	Unit	Mean	Median	SD	Min	Max	Obs
Total assets	Millions of euros	14.757	6.225	44.724	0	2,232.276	3,625
Operating revenue turnover	Millions of euros	29.204	6.103	57.298	0	555.052	3,081
Return on Assets	%	3.757	3.714	13.963	-538.672	93.439	5,492
Number of employees	Units	98.654	56.5	109.431	0	687.750	3,291
Labour productivity	Euros	103,358.572	69,430.22	162,685.739	-34,270.469	3,337,145.750	3,714
Dummy for GUO data exist.	-	0.276	0	0.447	0	1	8,022

Panel B: SA non-recipients

Variable name	Unit	Mean	Median	SD	Min	Max	Obs
Total assets	Millions of euros	2.042	0.471	9.949	0	2,167.542	6,060,988
Operating revenue turnover	Millions of euros	2.449	0.337	10.768	0	1,498.298	3,552,225
Return on Assets	%	-0.024	1.985	50.788	-6,134.047	1,812.907	5,773,130
Number of employees	Units	25.155	13	42.148	0	1,454.500	1,820,772
Labour productivity	Euros	58,935.760	40,578.6	137,616.715	-381,537.375	36,792,640	1,468,645
Dummy for GUO data exist.	-	0.0355	0	0.185	0	1	11,047,226

Notes: Panel A includes information for firms receiving SAs in 2016 and 2017. Observations with an annual granted aid less than 500,000 euros are considered as non-recipients. Panel B includes information for all the other firms in the same years. Variable 'Return on Assets' is the ratio between net income and total assets, multiplied by 100. 'Dummy for GUO data existence' is a binary variable taking value 1 if GUO level information is available and 0 otherwise. Variable 'Labour productivity' is added value per employee.

Table A3: Literature review

Paper	Grant type	Outcome	Sample	Method	Result
Colombo et al. (2013)	R&D	Investment rate / investment–cash flow sensitivity	Italian firms 1994-2008	Error correction model; GMM	expected only for small firms
Bronzini, Piselli (2016)	R&D	# patent applications / probability of submission	Italian firms 2004-2005	Regression discontinuity design	Positive, larger effects for small firms
Czarnitzki, Lopes-Bento (2013)	R&D	R&D investment / R&D employment	Belgian firms 2002-2008	Propensity score matching	Positive, not very significant
Dimos, Pugh	R&D	R&D investment	52 micro-level studies on R&D	Meta-analysis	No evidence of substantial additionality
Hyytinen, Toivanen (2005)	R&D	Private R&D / Firm growth	SMEs in Finland 2002	Difference-indifferences type	Positive on both outcomes
Szucs (2020)	R&D	Private R&D spending	Very large firms in 55 countries 2003-2017	Matching and difference-in-differences	Positive for smaller firms as well as for more R&D-intensive firms
Becker (2014)	R&D	Private R&D spending	Existing literature on R&D incentives	Survey	Mixed, also depending on the type of subsidy (e.g. direct, tax credit, …)
Duso et al. (2021)	Market Competition	Incentives for broadband coverage	German municipalities	Difference-in-differences	Positive effect on coverage, no harm to competition
Cerqua Pellegrini (2014)	Investment subsidies	Employment, investment, turnover, productivity	Italian firms 1995-2004	Regression discontinuity design	Positive effect on employment / investment / turnover; no effect on productivity
Grigolon et al. (2016)	Car scrapping subsidies	Car sales	Country-year data on 8 European countries 1998-2011	Difference-in-differences	Positive effect on sales; some crowding out with schemes targeted to low emission vehicles
Heim et al. (2017)	Recovery and restructuring aid	Long term survival	European firms 2004-2013	Propensity score matching	Positive effect on survival and on Z score
Sergant, Van Cayseele (2019)	State aid	TFP	Belgian firms 2003-2012	Regression	State aid enhances productivity
Criscuolo et al. (2019)	Regional policy – Regional Selective Assistance (UK) 1972-1980s	Employment	Ward areas	IV-type	Positive only for small firms
Buts Jegers (2013)	All state aids	Market shares	Large Belgian firms 2005-2008	Regression	Positive effect

Table A4: Logit out-of-sample performance – All recipients

		Real status		
		SA recipient = 0	**SA recipient = 1**	**Total**
Predicted status	**SA recipient = 0**	119,539	24	119,563
	SA recipient = 1	16,983	41	17,024
	Total	136,522	65	136,587
	Correctly predicted	**87.6%**	**63.1%**	**87.6%**

Notes: Output of a logit model estimated on the full combined Orbis-TAM databases. Such dataset was divided into a training set consisting of 2/3 of the observations and a testing set consisting of the remaining 1/3. Next, the original unbalanced training set was rebalanced using the SMOTE algorithm. The model was trained and tuned (via 10-fold cross-validation) on the artificially rebalanced training set, and its predictive performance was evaluated on the original (unbalanced) testing set. The sample consists of total annual aids granted to each recipient firm for the years 2016 and 2017. SA recipient is a binary variable taking value 1 if the firm is a State Aid recipient and 0 otherwise. Predictors include the set of firm-level quantitative Orbis variables. These Orbis variables are included in two forms: pre-SA averages of the two years $(t-1, t-2)$ before the SA was received; differences (i.e. deltas) between the pre-SA averages of the two years $(t-1, t-2)$ before the SA was received - and post-SA averages of the values for the year in which the SA was granted and the year after that $(t, t+1)$. Before the random splitting between training and testing sets, the sample was trimmed at the 1st and 99th percentiles to drop outliers in these Orbis predictors. The conventional probability threshold of 0.5 was used to assign observations into one of the two classes.

Table A5: LASSO out-of-sample performance – All recipients

		Real status		
		SA recipient = 0	**SA recipient = 1**	**Total**
Predicted status	**SA recipient = 0**	128,814	26	128,840
	SA recipient = 1	7,048	39	7,087
	Total	135,862	65	135,927
	Correctly predicted	**94.8%**	**60%**	**94.8%**

Notes: Output of a LASSO model generated on the full combined Orbis-TAM databases. Such dataset was divided into a training set consisting of 2/3 of the observations and a testing set consisting of the remaining 1/3. Next, the original unbalanced training set was rebalanced using the SMOTE algorithm. The model was trained and tuned (via 10-fold cross-validation) on the artificially rebalanced training set, and its predictive performance was evaluated on the original (unbalanced) testing set. The sample consists of total annual aids granted to each recipient firm for the years 2016 and 2017. SA recipient is a binary variable taking value 1 if the firm is a State Aid recipient and 0 otherwise. Predictors include the set of firm-level quantitative Orbis variables. These Orbis variables are included in two forms: pre-SA averages of the two years $(t-1, t-2)$ before the SA was received; differences (i.e. deltas) between the pre-SA averages of the two years $(t-1, t-2)$ before the SA was received - and post-SA averages of the values for the year in which the SA was granted and the year after that $(t, t+1)$. Before the random splitting between training and testing sets, the sample was trimmed at the 1st and 99th percentiles to drop outliers in these Orbis predictors. Dummies for four geographic areas and five aggregate NACE sectors are also included.

Table A6: Random forest out-of-sample performance – All recipients

		Real status		
		SA recipient = 0	**SA recipient = 1**	**Total**
Predicted status	**SA recipient = 0**	128,038	20	128,058
	SA recipient = 1	7,824	45	7,869
	Total	135,862	65	135,927
	Correctly predicted	**94.2%**	**69.2%**	**94.2%**

Notes: Output of a random forest (500 trees) generated on the full combined Orbis-TAM databases. Such dataset was divided into a training set consisting of 2/3 of the observations and a testing set consisting of the remaining 1/3. Next, the original unbalanced training set was rebalanced using the SMOTE algorithm. The model was trained and tuned (via 10-fold cross-validation) on the artificially rebalanced training set, and its predictive performance was evaluated on the original (unbalanced) testing set. The sample consists of total annual aids granted to each recipient firm for the years 2016 and 2017. SA recipient is a binary variable taking value 1 if the firm is a State Aid recipient and 0 otherwise. Predictors include the set of firm-level quantitative Orbis variables. These Orbis variables are included in two forms: pre-SA averages of the two years $(t-1, t-2)$ before the SA was received; differences (i.e. deltas) between the pre-SA averages of the two years $(t-1, t-2)$ before the SA was received - and post-SA averages of the values for the year in which the SA was granted and the year after that $(t, t+1)$. Before the random splitting between training and testing sets, the sample was trimmed at the 1st and 99th percentiles to drop outliers in these Orbis predictors. Dummies for four geographic areas and five aggregate NACE sectors are also included.

Table A7: Classification tree out-of-sample performance – Large recipients only

		Real status		
		SA recipient = 0	**SA recipient = 1**	**Total**
Predicted status	**SA recipient = 0**	3,416,325	264	3,416,589
	SA recipient = 1	228,318	701	229,019
	Total	3,644,643	965	3,645,608
	Correctly predicted	**93.7%**	**72.6%**	**93.7%**

Notes: Output of a classification tree generated on the full combined Orbis-TAM databases. Such dataset was divided into a training set consisting of 2/3 of the observations and a testing set consisting of the remaining 1/3. Next, the original unbalanced training set was rebalanced using the SMOTE algorithm. The model was trained and tuned (via 10-fold cross-validation) on the artificially rebalanced training set, and its predictive performance was evaluated on the original (unbalanced) testing set. The sample consists of total annual aids granted to each recipient firm for the years 2016 and 2017. SA recipient is a binary variable taking value 1 if the firm received total state aids larger than 1.5 million Euros in a given year and 0 otherwise. Predictors include two separate sets of firm and GUO level features, both including the quantitative Orbis variables. These Orbis variables are included in two forms: pre-SA averages of the two years $(t-1, t-2)$ before the SA was received; differences (i.e. deltas) between the pre-SA averages of the two years $(t-1, t-2)$ before the SA was received - and post-SA averages of the values for the year in which the SA was granted and the year after that $(t, t+1)$. Before the random splitting between training and testing sets, the sample was trimmed at the 1st and 99th percentiles to drop outliers in these Orbis predictors. Dummies for four geographic areas and five aggregate NACE sectors are also included.

Table A8: Logit out-of-sample performance – Large recipients only

		Real status		
		SA recipient = 0	**SA recipient = 1**	**Total**
Predicted status	**SA recipient = 0**	121,120	3	121,123
	SA recipient = 1	15,458	6	15,464
	Total	136,578	9	136,587
	Correctly predicted	**88.7%**	**66.7%**	**88.7%**

Notes: Output of a logit model estimated on the full combined Orbis-TAM databases. Such dataset was divided into a training set consisting of 2/3 of the observations and a testing set consisting of the remaining 1/3. Next, the original unbalanced training set was rebalanced using the SMOTE algorithm. The model was trained and tuned (via 10-fold cross-validation) on the artificially rebalanced training set, and its predictive performance was evaluated on the original (unbalanced) testing set. The sample consists of total annual aids granted to each recipient firm for the years 2016 and 2017. SA recipient is a binary variable taking value 1 if the firm received total state aids larger than 1.5 million Euros in a given year and 0 otherwise. Predictors include the set of firm-level quantitative Orbis variables. These Orbis variables are included in two forms: pre-SA averages of the two years $(t-1, t-2)$ before the SA was received; differences (i.e. deltas) between the pre-SA averages of the two years $(t-1, t-2)$ before the SA was received - and post-SA averages of the values for the year in which the SA was granted and the year after that $(t, t+1)$. Before the random splitting between training and testing sets, the sample was trimmed at the 1st and 99th percentiles to drop outliers in these Orbis predictors. The conventional probability threshold of 0.5 was used to assign observations into one of the two classes.

Table A9: LASSO out-of-sample performance – Large recipients only

		Real status		
		SA recipient = 0	**SA recipient = 1**	**Total**
Predicted status	**SA recipient = 0**	133,927	3	133,930
	SA recipient = 1	1,991	6	1,997
	Total	135,918	9	135,927
	Correctly predicted	**98.5%**	**66.7%**	**98.5%**

Notes: Output of a LASSO model generated on the full combined Orbis-TAM databases. Such dataset was divided into a training set consisting of 2/3 of the observations and a testing set consisting of the remaining 1/3. Next, the original unbalanced training set was rebalanced using the SMOTE algorithm. The model was trained and tuned (via 10-fold cross-validation) on the artificially rebalanced training set, and its predictive performance was evaluated on the original (unbalanced) testing set. The sample consists of total annual aids granted to each recipient firm for the years 2016 and 2017. SA recipient is a binary variable taking value 1 if the firm received total state aids larger than 1.5 million Euros in a given year and 0 otherwise. Predictors include the set of firm-level quantitative Orbis variables. These Orbis variables are included in two forms: pre-SA averages of the two years $(t-1, t-2)$ before the SA was received; differences (i.e. deltas) between the pre-SA averages of the two years $(t-1, t-2)$ before the SA was received - and post-SA averages of the values for the year in which the SA was granted and the year after that $(t, t+1)$. Before the random splitting between training and testing sets, the sample was trimmed at the 1st and 99th percentiles to drop outliers in these Orbis predictors. Dummies for four geographic areas and five aggregate NACE sectors are also included.

Table A10: Random forest out-of-sample performance – Large recipients only

		Real status		
		SA recipient = 0	**SA recipient = 1**	**Total**
Predicted status	**SA recipient = 0**	133,073	3	133,076
	SA recipient = 1	2,845	6	2,851
	Total	135,918	9	135,927
	Correctly predicted	**97.9%**	**66.7%**	**97.9%**

Notes: Output of a random forest (500 trees) generated on the full combined Orbis-TAM databases. Such dataset was divided into a training set consisting of 2/3 of the observations and a testing set consisting of the remaining 1/3. Next, the original unbalanced training set was rebalanced using the SMOTE algorithm. The model was trained and tuned (via 10-fold cross-validation) on the artificially rebalanced training set, and its predictive performance was evaluated on the original (unbalanced) testing set. The sample consists of total annual aids granted to each recipient firm for the years 2016 and 2017. SA recipient is a binary variable taking value 1 if the firm received total state aids larger than 1.5 million Euros in a given year and 0 otherwise. Predictors include the set of firm-level quantitative Orbis variables. These Orbis variables are included in two forms: pre-SA averages of the two years $(t-1, t-2)$ before the SA was received; differences (i.e. deltas) between the pre-SA averages of the two years $(t-1, t-2)$ before the SA was received - and post-SA averages of the values for the year in which the SA was granted and the year after that $(t, t+1)$. Before the random splitting between training and testing sets, the sample was trimmed at the 1st and 99th percentiles to drop outliers in these Orbis predictors. Dummies for four geographic areas and five aggregate NACE sectors are also included.

Table A11: Logistic regression for the false positive status (marginal effects)

Independent variables	Marginal effect on false positive
Large size	0.267***
	(0.000410)
Austria	0.0825***
	(0.00130)
Belgium	0.0186***
	(0.00100)
Bulgaria	-0.00714***
	(0.00182)
Cyprus	0.135***
	(0.0151)
Czech Republic	0.00454***
	(0.00148)
Germany	0.0281***
	(0.000810)
Denmark	0.0539***
	(0.00121)
Estonia	0.00570**
	(0.00269)
Finland	0.000705
	(0.00163)
France	-0.00831***
	(0.000809)
Greece	0.0386***
	(0.00249)
Croatia	0.00961***
	(0.00240)
Hungary	0.0124***
	(0.00161)
Ireland	0.0382***
	(0.00184)
Italy	-0.00800***
	(0.000723)
Lituania	0.0584***
	(0.00426)
Luxembourg	0.0945***
	(0.00346)
Latvia	0.0168***
	(0.00357)
Malta	0.0179***
	(0.00608)
Netherlands	0.00685***
	(0.000864)
Poland	0.0711***
	(0.00147)

Table A11: Logistic regression for the false positive status (marginal effects) – continued

Independent variables	Marginal effect on false positive
Portugal	0.0115***
	(0.00131)
Romania	0.0267***
	(0.00170)
Sweden	0.0165***
	(0.00115)
Slovenia	0.0140***
	(0.00271)
Slovakia	0.00617***
	(0.00194)
5.NACE_2 digits	0.144***
	(0.0361)
6.NACE_2 digits	0.151***
	(0.0185)
7.NACE_2 digits	0.177***
	(0.0246)
8.NACE_2 digits	0.174***
	(0.00446)
9.NACE_2 digits	0.148***
	(0.0126)
10.NACE_2 digits	0.128***
	(0.00235)
11.NACE_2 digits	0.198***
	(0.00406)
12.NACE_2 digits	0.205***
	(0.0261)
13.NACE_2 digits	0.0905***
	(0.00374)
14.NACE_2 digits	0.0364***
	(0.00372)
15.NACE_2 digits	0.0489***
	(0.00443)
16.NACE_2 digits	0.0564***
	(0.00311)
17.NACE_2 digits	0.149***
	(0.00442)
18.NACE_2 digits	0.0510***
	(0.00341)
19.NACE_2 digits	0.197***
	(0.0129)
20.NACE_2 digits	0.161***
	(0.00342)
21.NACE_2 digits	0.200***
	(0.00609)

Table A11: Logistic regression for the false positive status (marginal effects) – continued

Independent variables	Marginal effect on false positive
22.NACE_2 digits	0.126***
	(0.00284)
23.NACE_2 digits	0.119***
	(0.00300)
24.NACE_2 digits	0.160***
	(0.00432)
25.NACE_2 digits	0.0634***
	(0.00203)
26.NACE_2 digits	0.0857***
	(0.00338)
27.NACE_2 digits	0.0908***
	(0.00337)
28.NACE_2 digits	0.0975***
	(0.00235)
29.NACE_2 digits	0.155***
	(0.00424)
30.NACE_2 digits	0.115***
	(0.00560)
31.NACE_2 digits	0.0590***
	(0.00341)
32.NACE_2 digits	0.0270***
	(0.00336)
33.NACE_2 digits	0.00127
	(0.00288)
35.NACE_2 digits	0.207***
	(0.00227)
36.NACE_2 digits	0.244***
	(0.00603)
37.NACE_2 digits	0.133***
	(0.00803)
38.NACE_2 digits	0.136***
	(0.00336)
39.NACE_2 digits	0.0607***
	(0.0103)
41.NACE_2 digits	-0.0427***
	(0.00174)
42.NACE_2 digits	-0.000229
	(0.00270)
43.NACE_2 digits	-0.0764***
	(0.00171)
45.NACE_2 digits	-0.0262***
	(0.00190)
46.NACE_2 digits	-0.0217***
	(0.00166)

Table A11: Logistic regression for the false positive status (marginal effects) – continued

Independent variables	Marginal effect on false positive
47.NACE_2 digits	-0.0658***
	(0.00169)
49.NACE_2 digits	-0.0226***
	(0.00197)
50.NACE_2 digits	0.110***
	(0.00476)
51.NACE_2 digits	0.0691***
	(0.0106)
52.NACE_2 digits	0.00307
	(0.00231)
53.NACE_2 digits	-0.0447***
	(0.00672)
56.NACE_2 digits	-0.0842***
	(0.00194)
58.NACE_2 digits	-0.0145***
	(0.00324)
59.NACE_2 digits	-0.0205***
	(0.00350)
60.NACE_2 digits	-0.00915
	(0.00723)
61.NACE_2 digits	0.0297***
	(0.00482)
62.NACE_2 digits	-0.0168***
	(0.00208)
63.NACE_2 digits	-0.0286***
	(0.00360)
64.NACE_2 digits	-0.000369
	(0.00175)
68.NACE_2 digits	-0.0325***
	(0.00166)
69.NACE_2 digits	-0.0821***
	(0.00202)
70.NACE_2 digits	-0.0162***
	(0.00181)
71.NACE_2 digits	-0.0418***
	(0.00200)
72.NACE_2 digits	0.0117***
	(0.00381)
73.NACE_2 digits	-0.0513***
	(0.00245)
74.NACE_2 digits	-0.0292***
	(0.00271)
75.NACE_2 digits	-0.106***
	(0.00508)

Table A11: Logistic regression for the false positive status (marginal effects) – continued

Independent variables	Marginal effect on false positive
77.NACE_2 digits	0.0205***
	(0.00267)
78.NACE_2 digits	-0.0320***
	(0.00305)
79.NACE_2 digits	-0.0400***
	(0.00329)
80.NACE_2 digits	-0.0392***
	(0.00436)
81.NACE_2 digits	-0.0658***
	(0.00241)
82.NACE_2 digits	-0.00668***
	(0.00213)
84.NACE_2 digits	0.180**
	(0.0827)

Notes: Multivariate logistic regression of the false positive status on a set of country dummies NACE 2-digit sectoral dummies, and a size dummy called 'Large size' which takes value 1 if a firm has above median total assets and 0 otherwise. False positive status is indicated by a dummy taking value 1 if the tree a non-recipient firm as recipient and 0 otherwise. Standard errors are derived from asymptotic theory for models fit with maximum likelihood estimation. Excluded dummies are Spain for countries and NACE-2 no.55 for the two-digit sectors. These were chosen as they have an average value of the share of false positives closest to the sample average. *** p<0.01, ** p<0.05, * p<0.1.

Figure A1: From panel data to cross sectional data with leads and lags

Panel A: Original panel data

year	firm_id	x1	…	xk	SA
2014	1				.
…	2				.
…	…		X_2014		.
2014	n				.
2015	1				.
…	2				.
…	…		X_2015		.
2015	n				.
2016	1				0
…	2				1
…	…		X_2016		0
2016	n				1
2017	1				1
…	2				0
…	…		X_2017		0
2017	n				0
2018	1				.
…	2				.
…	…		X_2018		.
2018	n				.

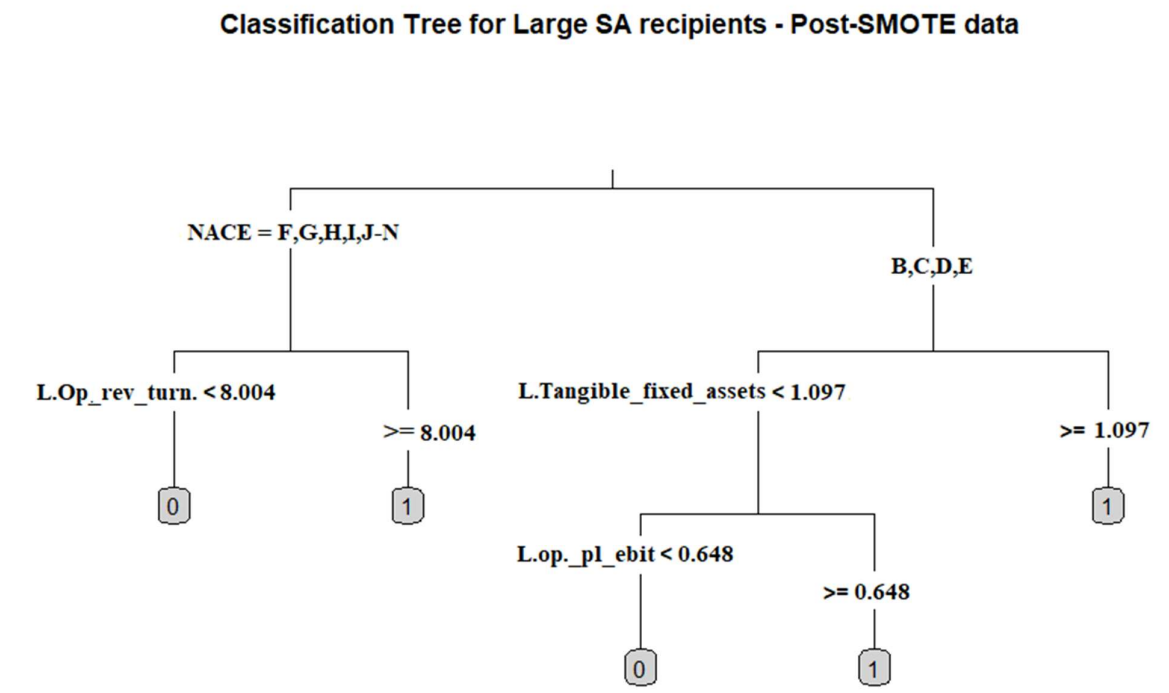
Panel B: Transformed data structure for ML analysis

year	SA	firm_id	x1_pre	…	xk_pre	x1_delta	…	xk_delta	same x-type variables but at the GUO level
2016	0	1							
…	1	2							
…	0	…	X1_avg(2014, 2015)	…	Xk_avg(2014, 2015)	X1_avg(2016, 2017)- X1_avg(2014, 2015)	…	Xk_avg(2016, 2017)- Xk_avg(2014, 2015)	…
2016	1	n							
2017	1	1							
…	0	2							
…	0	…	X1_avg(2015, 2016)	…	Xk_avg(2015, 2016)	X1_avg(2018, 2019)- X1_avg(2016, 2017)		Xk_avg(2018, 2019)- Xk_avg(2016, 2017)	…
2017	0	n							

Figure A2: Stylized confusion matrix

		Real Status	
		SA recipient = 0	SA recipient = 1
Predicted status	SA recipient = 0	N_{11}	N_{12}
	SA recipient = 1	N_{21}	N_{22}

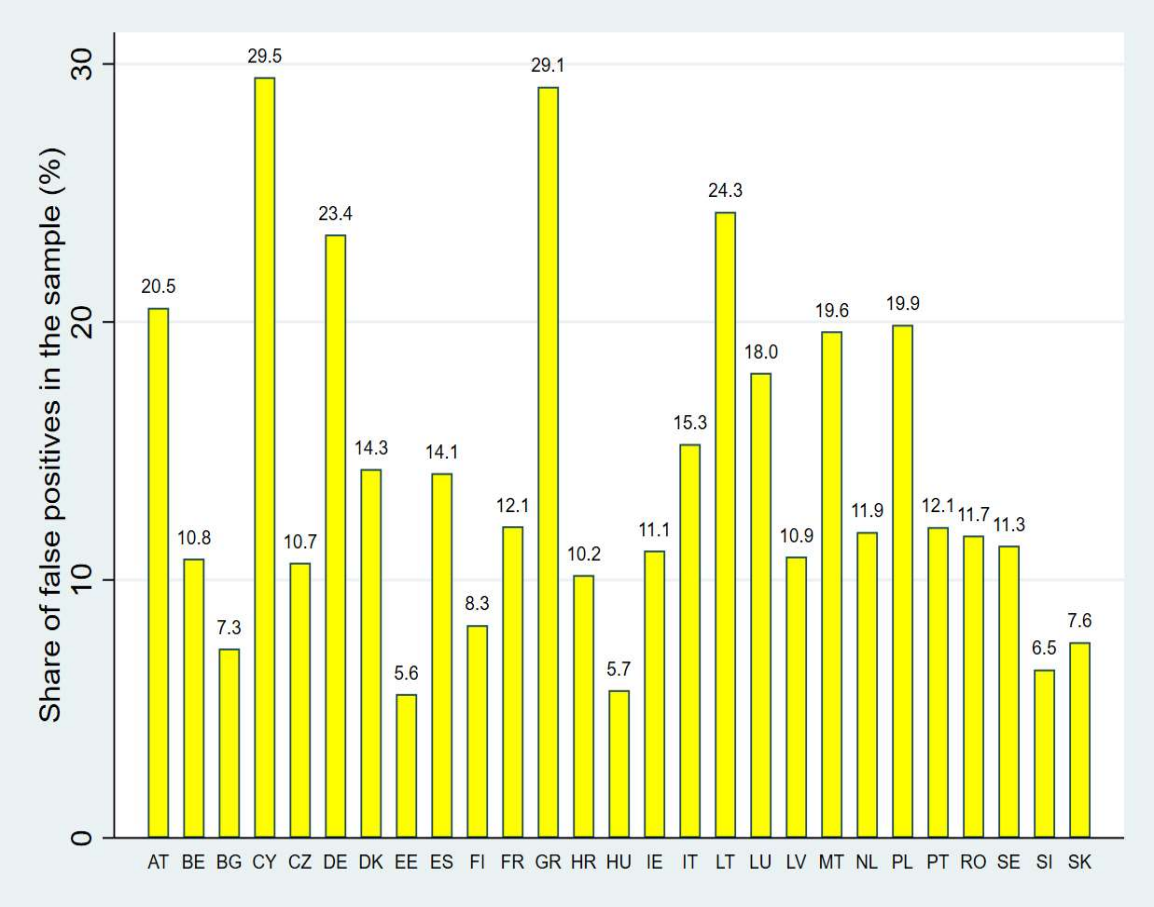
Figure A3: Classification tree – Large recipients only



Notes: Visual output of a classification tree generated on the full combined Orbis-TAM databases. Such dataset was divided into a training set consisting of 2/3 of the observations and a testing set consisting of the remaining 1/3. Next, the original unbalanced training set was rebalanced using the SMOTE algorithm. The model was trained and tuned (via 10-fold cross-validation) on the artificially rebalanced training set, and its predictive performance was evaluated on the original (unbalanced) testing set. The sample consists of total annual aids granted to each recipient firm for the years 2016 and 2017. SA recipient is a binary variable taking value 1 if the firm received total state aids larger than 1.5 million Euros in a given year and 0 otherwise. Predictors include two separate sets of firm and GUO level features, both including all the quantitative Orbis variables. These Orbis variables are included in two forms: pre-SA averages of the two years $(t-1, t-2)$ before the SA was received; differences (i.e. deltas) between the pre-SA averages of the two years $(t-1, t-2)$ before the SA was received - and post-SA averages of the values for the year in which the SA was granted and the year after that $(t, t+1)$. Before the random splitting between training and testing sets, the sample was trimmed at the 1st and 99th percentiles to drop outliers in these Orbis predictors. Dummies for four geographic areas and five aggregate NACE sectors are also included. Variable 'L.Tangible_fixed_assets' refers to the average value of Tangible Fixed Assets for the two years before the SA was granted. Variable 'L.Op _rev_turn.' refers to the average value of Operating Revenue Turnover for the two years before the SA was granted. Variable 'L.op.pl_ebit.' refers to the average value of Operating_pl_ebit for the two years before the SA was granted. All the values are expressed in millions of euros.

Figure A4: Descriptive statistics of false positives by country

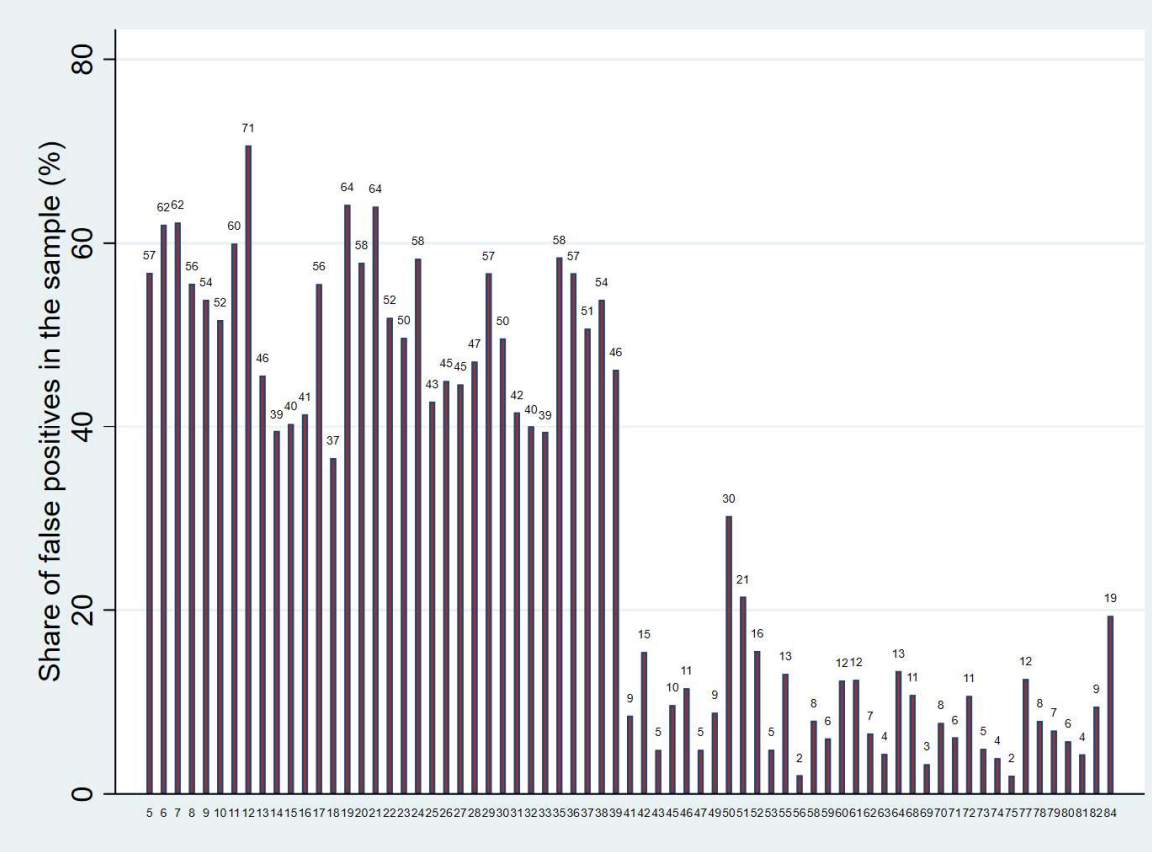
Share of false positives by country – Testing set, all recipients



Notes: False positive status is indicated by a dummy taking value 1 if the firm is a false positive and 0 otherwise. Source. Source: Authors' elaborations on Orbis and TAM data.

Figure A5: Descriptive statistics of false positives by sector

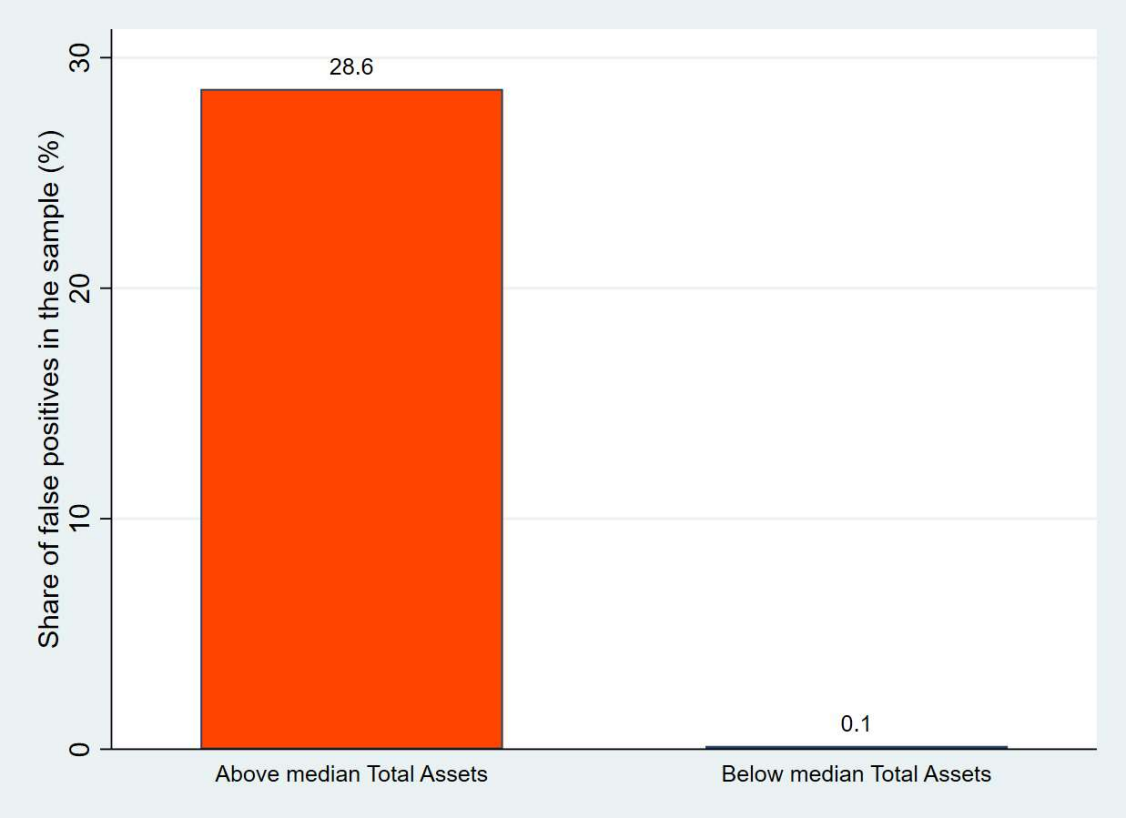
Share of false positives by NACE 2-digit sector – Testing set, all recipients



Notes: False positive status is indicated by a dummy taking value 1 if the firm is a false positive and 0 otherwise. Source: Authors' elaborations on Orbis and TAM data.

Figure A6: Descriptive statistics of false positives by firm size

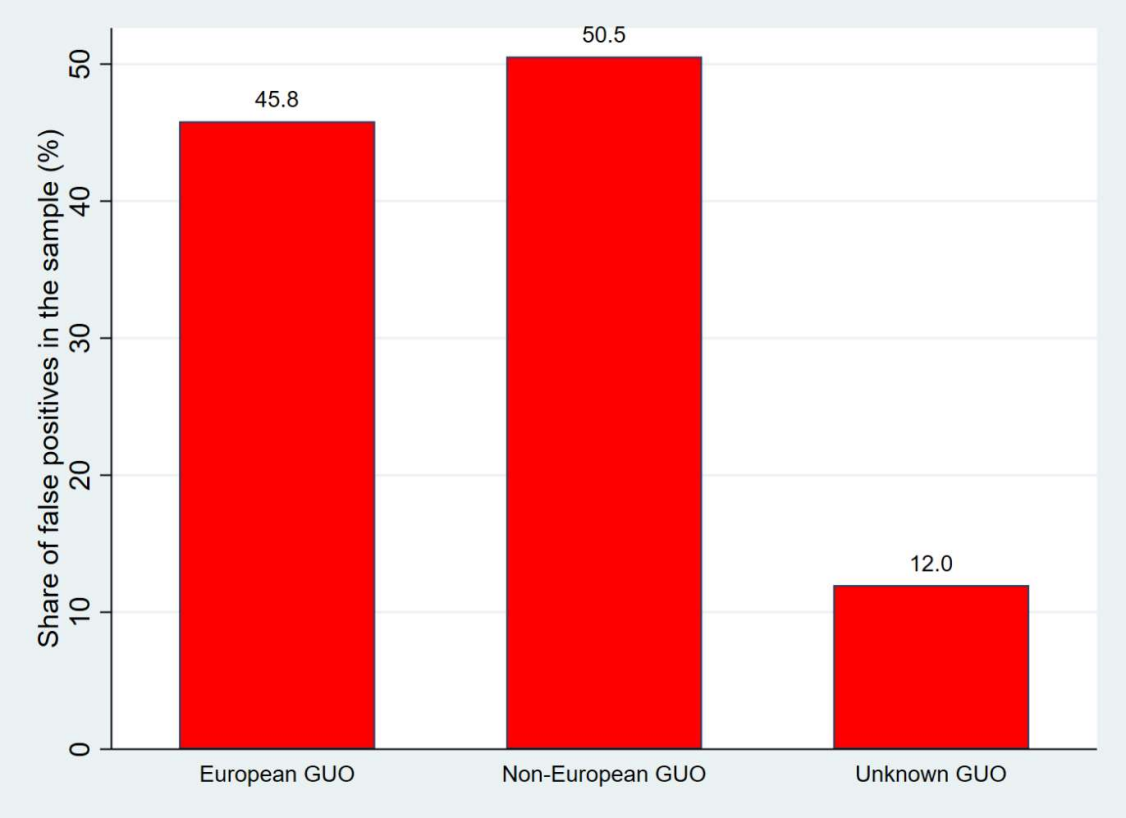
Share of false positives by size – Testing set, all recipients



Notes: False positive status is indicated by a dummy taking value 1 if the firm is a false positive and 0 otherwise. Source. Source: Authors' elaborations on Orbis and TAM data.

Figure A7: Descriptive statistics of false positives by GUO

Share of false positives by GUO – Testing set, all recipients



Notes: False positive status is indicated by a dummy taking value 1 if the firm is a false positive and 0 otherwise. Source: Authors' elaborations on Orbis and TAM data.