A machine learning approach for the detection of firms infiltrated by organised crime in Italy

by P. Cariello, M. De Simoni and S. Iezzi[§]

Abstract

We develop a machine learning algorithm for detecting legally registered firms potentially at risk of being infiltrated by organized crime. To this end, we exploit a firm-level dataset for Italy by merging financial information from various sources, including, in particular, financial statements and data on loans. A sample of about 1,800 Italian firms known to be infiltrated almost with certainty is compared with probabilistic samples of alleged legal firms in order to train and test the model. The algorithm correctly classifies 86.5% of firms within matched testing samples including infiltrated and lawful firms. The main output of the algorithm is a risk score, which can be deployed at an operational level for anti-money laundering and prudential supervision.

JEL Classification: C53, C55, D22, K42.

Keywords: organized crime, infiltrated firms, money laundering, financial statements, machine learning.

§ Bank of Italy, Financial Intelligence Unit.

1. Introduction

Criminal organisations invest vast sums of money within the legal economies of many countries worldwide. The United Nations Office on Drugs and Crime assesses that in 2009 the revenues generated globally by OC amounted to 3.6% of the world's GDP (UNODC, 2011). According to sources of the European Council, in 2019 criminal revenues in the main criminal markets amounted to 1% of the EU's GDP, i.e. €139 billion. As for Italy, a study by Transcrime, in cooperation with the Italian Ministry of the Interior, shows that the proceeds of mafia groups is estimated to range between 1 and 2 per cent of GDP (Transcrime, 2015). One of the main concerns for the national and international authorities is the ever increasing investment of OC in the official economy through the infiltration of legitimate businesses.

Infiltrated businesses can be defined as firms that are legally registered and apparently engaged in lawful activities but are controlled by criminal organizations. Those firms differ from non-infiltrated firms in three main ways (Ravenda *et al.*, 2015;De Simoni, 2022). Generally, owners or managers are members of a criminal organization; funding comes partially or totally from illegal activities; in some cases criminal methods involving violence, intimidation or corruption might be used while doing business. Legal and illegal activities are therefore closely interconnected within infiltrated firms as the legal activities mostly serve to launder profits stemming from illegal ones.

Several scholars have recently engaged in explaining the effects of infiltration on firms' financial statements, with the aim of identifying differences in business management between infiltrated firms and lawful ones. These studies converge on many points and differ in others, but overall conclude that infiltrated firms, from a financial point of view, exhibit a peculiar financial statement, at least with regard to some of its dimensions. These relevant findings have given life to the development of statistical models with the aim to discriminate between infiltrated and non-infiltrated firms on the basis of financial reports and, ultimately, to detect apparently lawful firms, which are actually controlled by organized crime.

Our contribution to the literature is threefold. First, we build a unique firm-level dataset for Italy spanning from 2010 to 2020 by merging financial statement information provided by the National Official Business Register, data on firms' debts towards the banking and financial system provided by the Central Credit Register of Bank of Italy, and data on employment and payrolls provided by the National Institute of Social Security.

This highly varied source of data allows us to construct a large set of financial variables and indicators, which represent the cornerstone of our analysis.

Second, we use a unique sample of about 1,800 infiltrated firms by using a variety of public sources which make our study substantially more robust than the existing research on this topic: while many studies identify infiltrated firms by relying on an educated guess, our sample includes only firms whose infiltration by OC can be held to be almost certain.

Third, we resort to a machine learning approach with the aim to build a classifier capable to identify legally registered firms potentially infiltrated by organized crime. The most successful algorithm turned out to be the gradient-boosted decision trees (XGB) approach having a total accuracy of 86.5%, precision of 84% and recall of 81% on the testing set. Based on the confusion matrix on the testing set, type I error rate is 10% while type II error rate is 18.7%, i.e. the algorithm classifies legal companies as criminal in 10 cases out of 100 and classifies criminal firms to be reliable businesses in nearly 19 cases out of 100 cases. As far as we know there are no previous studies in the literature that try to develop a machine learning algorithm in order to detect legally registered firms at risk of being infiltrated by OC.

The main output of the algorithm is a risk score computed for a large portion of the universe of limited liabilities companies active in Italy between 2010 and 2020 for which we have complete data. The risk score can be deployed at an operational level for anti-money laundering and prudential supervision.

The paper is structured as follows. Section 2 briefly illustrates the main findings of the literature on the role of OC in the legitimate economy and outlines the motivation behind this work. In Section 3 we discuss the data. Section 4 illustrates the methodology and section 5 the results. In section 6 we discuss possible applications for anti-money laundering and prudential supervision of our methodology. Finally, Section 7 concludes.

2. Literature review

Many scholars tried to estimate how the presence of OC negatively affects the way the economy works, for instance, by hindering competition and the optimal allocation of resources, which in turn may lower the overall level of output (Peri, 2004; Barone and Mocetti, 2014; Pinotti, 2015). Literature has also focused on other costs associated to OC

presence, such as those arising from the deterioration of the quality of the political class (Daniele and Geys, 2015), the reduction of electoral competition (De Feo and De Luca, 2013) and the decrease in foreign investments (Daniele and Marani, 2011).

The question on how infiltrated firms operate in the economy, with particular focus on Italy, is widely debated in the literature. Indeed, several scholars have recently engaged in explaining the effects of infiltration on Italian firms' financial statements. By examining a list of companies subject to legal proceedings and located in central and northern Italy, Fabrizi *et al.* (2017a) show that criminal companies are larger, more indebted and hold more liquid assets than legal ones. Bianchi *et al.* (2020) analyse companies based in Lombardy that have representatives linked to organized crime, and show how criminal organizations "cannibalize" profits and drain resources, also through money laundering schemes. Mirenda *et al.* (2019) study the infiltration of 'Ndrangheta, a criminal organization headquartered in the Southern region of Calabria, into firms located in the Centre and North of Italy. They show that 'Ndrangheta tends to enter firms in economic and financial distress and those mostly relying on public sector procurement. De Simoni (2022) finds that infiltrated firms, despite their higher revenues, are less profitable and hold more cash assets. Investment decisions and evidence on financing costs depend on the type of infiltrated firms.

The literature on the analysis of criminal firms' balance sheet is sufficiently wide to provide a sound enough support to our idea. Our work mainly capitalize the findings of recent studies in order to build a highly varied set of financial variables and indicators to train the Machine Learning algorithm to identify infiltrated businesses. Nevertheless, as far as we know, there are no previous scientific studies seeking to develop a Machine Learning algorithm in order to detect criminal firms based on accounting and financial data. The only similar approach we found in the literature is provided by Ravenda *et al.* (2015) which develop a logistic regression model that contributes to the detection of legally mafia firms in Italy based on their financial statement characteristics. More broadly, there are some contributions providing machine learning applications in the field of financial fraud at large, which is weakly linked to the aim of our study (Chengwei et al., 2015; Maka et al., 2020; Sadgali et al., 2019; Sharma and Panigrahi, 2013; Wyrobek, 2020).

3. The data

We build a unique dataset by merging information from three different sources: 1) for financial statement information we use Cerved database, that is provided by the National Official Business Register and covers a very large portion of Italian small and medium-sized corporations, all of them limited liability companies; 2) for information on company's bank liabilities we use the Bank of Italy's Central Credit Register; 3) for data on employment and payrolls we exploit a database provided by the National Institute of Social Security (INPS). We extract firm-level data spanning over 11 years, from 2010 to 2020.

A crucial feature of our research is the appropriate selection of the variables to use in order to train the algorithm to detect infiltrated businesses. Based on most relevant works in the field, discussed in the previous section, we select a list of 25 variables/indicators, whose aim is defining the main dimensions of firms' financial profile. The analysed dimensions are six. The first one is firms' size, which is measured by total assets, revenues, equity, tangibles assets and short-term liabilities. A second group is composed by six equity and liquidity indicators. A third category is made of three indicators of indebtedness computed by combining financial budget variables with firms' bank liabilities coming from the Bank of Italy's Central Credit Register. Then we have a fifth group made of five profitability indicators, a sixth category made of investment and cost structure indicators, and the last group which includes three budget indicators computed per labour unit. To these wide set of variables and indicators we add two structural characteristics of the firms, that is, the province where the firm is located (in Italy there are 110 provinces distributed in 20 regions in 2020) and the main economic sector of activity identified by the 3-digit NACE code. Table 1 shows the complete list of variable used in our analysis and table 2 their main descriptive statistics. The dataset has been previously subjected to a basic cleansing treatment in order to spot and resolve potential data inconsistencies.

a. The list of infiltrated firms

In order to apply a supervised learning approach it is necessary to have a subset of firms labelled as criminal that we can compare with the remaining set of alleged legal firms. In order to determine whether a firm is infiltrated and the years where the firm is to be considered as such, we use a variety of sources. A first set of 223 firms (list A) is selected thanks to the cooperation with an Italian law enforcement unit specialised in the fight against OC and terrorism; this list includes businesses seized or confiscated by the judiciary as a result of criminal investigations in the decade 2007 to 2017. We expand this first core

by including a list of 340 seized businesses (list B), drawn by the archives of the ANBSC (Agenzia Nazionale per l'amministrazione e la destinazione dei Beni Sequestrati e Confiscati alla criminalità organizzata), an Italian governmental body in charge of administering all assets, including companies, seized to OC[1]. An additional set of 1,223 firms (list C) is extracted from a commercial database where we select all companies featuring stakeholders or administrators involved in OC-related investigation taking place in the same period. In our analysis we focus only on private limited liability companies, thus excluding partnerships and public limited liability companies, since for the former we have very limited data on financial accounts and for the letter we have only very few example cases of infiltration.

For firms of list A and B we do not know the starting year of infiltration but we do know the seizure year representing the moment where the firm stops to be infiltrated. Nevertheless, in order to prevent results from being influenced by seizure-related operations or anticipatory effects, we only use data for all years up to the second before the seizure. For the years following the seizure, the data are also excluded from the analysis since the firms are managed by judicial administrators, thus they cannot be considered as regular legal firms. For list C we assume that the starting year of infiltration is the year where colluded stakeholders or administrators take control of the firms. Since we do not have an ending year of infiltration for these cases, we use all data from the starting year of infiltration up to the last year of analysis, which is 2020. We discard data preceding the infiltration.

b. Missing data issues

Our dataset suffers from a relatively significant incomplete data issue; as shown in table 4 only 33.5 per cent of records for non-infiltrated firms have complete data and 32.4 per cent of records have more than 10 missing items out of 27 variables. Depending on the variable, the proportion of missing data can vary from 0.2 per cent for the variable

[1] The history of seizures of assets linked to OC groups in Italy traces back to the 1980s. Indeed, in 1982 the so-called Rognoni La Torre Act (Law 646/1982) introduced measures that directly attack the wealth of people affiliated to mafia groups. Moreover, since then, various other reforms were implemented aiming to prevent these firms from defaulting when run by the State. Overall, these instruments have been found to be very effective in the fight against OC. Moreover, seized and confiscated firms are a relevant phenomenon in the Italian economy. In the economic debate, it is acknowledged that a proper management of these assets can save jobs and help the State in defeating OC (Donato, Saporito and Scognamiglio, 2013). In 2010 a decree law established an Agency (*Agenzia Nazionale per l'Amministrazione e la Destinazione dei Beni sequestrati e Confiscati alla criminalità organizzata* - ANSBC), whose purpose is to manage seized and confiscated assets from OC. The Agency currently runs almost 2.900 firms and more than 18.000 real estates, and collects data on all seizures of mafia assets.

assets to about 44 per cent for the revenues over number of employees indicator (see table 2). The strategy we adopt for dealing with incomplete data in our analysis is twofold and differs between non-infiltrated and infiltrated firms.

For alleged legal firms, since we have a very large number of records, we decide to use a complete case analysis (CCA) approach by removing missing data using listwise deletion, i.e. deleting data for all cases that have missing data for any variable. In this way, we have a dataset that is complete for all firms included in it. A downside of this technique is that we end up with a sample of only 33.5 per cent of complete records and 44 per cent of firms (table 4). This means that our final sample could be biased because it may not adequately represent the population of legal firms. However, the full sample is not much different from the sample of complete cases according to the distribution of firms by sector and region: this evidence is not sufficient to say that missing data are completely at random, but gives a robust clue to this hypothesis, thus supporting the use of a complete case approach. Another downside of this approach is that the risk score cannot be computed for the whole population of firms but only for a portion of it, thus implying that there will be limitations in the scope of application for anti-money laundering (AML) and prudential supervision purposes.

For infiltrated firms, since we have a limited sample, we adopt a different strategy by applying an imputation procedure with the purpose of retaining most of the data. In particular, we previously remove records with missing data for the province and sector categorical variables or with more than 6 missing items out of 25 financial variables/indicators in order to limit the proportion of missing data to be imputed. This selection reduces the number of records from 9,294 to 6,294 (the number of firms drops from 2,293 to 1,786), although not altering the distribution according to sector and region. Given the general missing pattern in our dataset, we apply a fully conditional specification (FCS) method to impute the remaining missing values. The FCS method assumes the existence of a joint distribution for the variables (Brand, 1999; Van Buuren, 2007) and involves two phases: the preliminary filled-in phase followed by the imputation phase. At the filled-in phase, the missing values for all variables are filled in sequentially over the variables taken one at a time using a linear regression model with preceding variables serving as covariates. These filled-in values provide starting values for these missing values at the imputation phase. At the imputation phase, the missing values for each variable are imputed using the linear regression model and covariates at each iteration. After a specified

number of iterations, which we set at 200, the imputed values in each variable are used for the imputation. At each iteration, the missing values are imputed sequentially over the variables taken one at a time.

Our final dataset includes 6,294 records concerning 1,786 criminal firms and 3,224,204 records regarding 746,843 alleged legal firms. Table 4 briefly shows the structure of the dataset by year and table 5 reports means and standard deviations for the 25 financial variables/indicators for the final dataset; the final dataset seems to be broadly in line with the initial dataset (table 2) according to means and standard deviations.

c. Variables selection

We have studied the pairwise correlation of all the variables and indicators, and found that most of them have a low or null correlation. Only two couples of variables have a correlation index above 70%. Even most of the ML algorithms we use are robust to correlated factors, we choose to remove one of the variables for each couple. Therefore, we use 25 factors and one target variable. Figure 1 shows the pairwise correlation of variables selected for the model. Note also that the target variable has a low correlation with any of the factors.

4. Classification methodology

In order to cope with the problem of high imbalance between records for infiltrated and non-infiltrated firms, we employ an under-sampling strategy by applying a stratified random sampling of non-infiltrated firms. The strata are defined according to the combination of year, sector and region of activity of the firm. The list of non-infiltrated firms is sub-sampled in order to obtain a proportion of infiltrated firms of about 40% of the total.

We are aware that most likely only a minor part of criminal firms are labelled as infiltrated according to our sources. Nonetheless, considering the large population of alleged legal firms from which control samples have been selected in order to train the algorithm, we can assume a very low probability of a significant presence of criminal firms in those samples used for our training and test. Moreover, the goal of our model is to be able to identify those firms which are allegedly lawful but are under the control of criminal organizations.

After rebalancing the sample, we obtained a set made of 15,794 rows, each corresponding to annual data of a firm. Then we split the data sample in two sets with a proportion of 80/20, respectively for training and testing. Training set consists of 12,635 rows, of which 5,035 related to firm labelled as infiltrated. Testing set consists of 3,159 rows, of which 1,259 infiltrated.

The machine learning algorithm is iteratively trained on the training set. Since we have a small sample of infiltrated firms, we employ 5-fold cross-validation for the selection of the model and the calibration of hyper-parameters in order to avoid the creation of a validation set.

For model training we compare various ML algorithms: logistic regression, random forest model, neural networks and gradient boosted decision trees model. All the models are created in Python using the following libraries: scikit-learn, xgboost, TensorFlow and Keras. We also use SAS Viya to build a similar analysis by employing its visual interface Model Studio. The results are consistent with the ones obtained in Python (see. Fig. 2).

For gradient boosted decision trees we use gridsearchCV to find the optimal combination of hyperparameters. We use a grid of 480 points with recall as a variable score. We choose to maximize the recall because we are relatively confident about infiltrated firms, thus we consider a false negative less acceptable than a false positive.

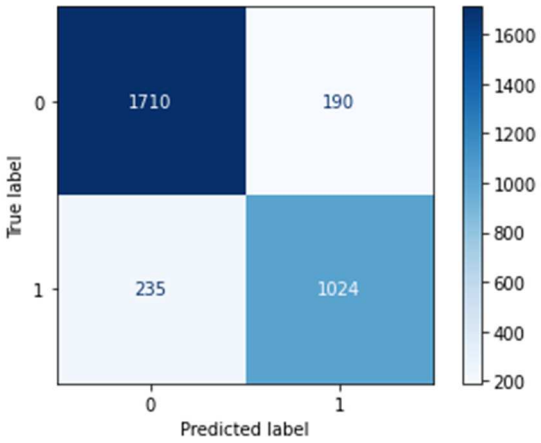
At the end of the search, we obtain optimal performances using this setting: 1,000 decision trees (n_estimators) with a maximum depth of 10 levels and learning rate set at 10%. Other parameters are left to default values.

5. Results

We use the test set to estimate the performance of the 'champion' model and to verify that it is generalizable to new data.

The most successful algorithm turned out to be the gradient-boosted decision trees (XGB) approach having a total accuracy of 86.5%, precision of 84% and recall of 81% on the testing set. Based on the confusion matrix on the testing set, type I error rate is 10% while type II error rate is 18.7%, i.e. the algorithm classifies legal companies as criminal in 10 cases out of 100 and classifies criminal firms to be reliable businesses in nearly 19 cases out of 100 cases.

The confusion matrix computed for the test set is the following:



The performance of our model is better than similar models found in the literature which use annual financial statements to identify unfair companies. In particular, Wyrobek J. (2020) finds that the best algorithm to maximize the recall is XGB; nevertheless, his model has slightly lower precision and recall compared to ours. Note, however, that any comparison should be considered with caution since the objectives of the analysis are different.

We also measure the relative importance of the independent variables in determining the final risk score (figure 3): equity ratios and the variables associated to employees turn out to bear the most weight on the indicator, whilst variables measuring a firm's indebtedness the least.

As a robustness check, instead of splitting the sample randomly into training and testing sets, we train the algorithm on a rolling subset of data spanning a 4-year period and test on the subsequent year. Results show that, although the number of positive cases of infiltrated firms employed in the analysis is reduced by more than 60%, the performance indicators deteriorate only moderately with respect to those obtained on the whole dataset (on average over the years precision and recall drop to 79% from 84% and 81%, respectively), thus suggesting that the algorithm can be effectively used for one-year ahead forecasting (see Fig. 4).

6. Applications for anti-money laundering and prudential supervision

The risk score is computed for 472,539 firms, for which we have complete records in the most recent years, 2018-2020. Table 6 shows the frequency distribution of the estimated risk score: 89.6 per cent of firms are to be considered with a low risk profile, having a risk score of less than 0.5; the remaining 10.4 per cent of firms are labelled as

risky businesses according to our model and, in particular, 1 per cent of them are labelled as very high-risk firms having a risk score of more than 0.99. Table 7 shows that more than 70 per cent of risky firms, i.e. with a risk score higher than 0.5, are located in the South of Italy although only 28.6 per cent of all firms in the sample are sited in southern regions.

The risk score associated to the Italian registered limited liability companies has several potential applications for anti-money laundering (AML) purposes alike. The risk score may be used in order to prioritise work within the central AML authority, as it may signal the involvement of high-risk companies in the financial conducts that are reported as suspicious by AML obliged entities. The score can be also computed as an aggregate risk indicator both at a geographical or sectoral level, which may provide interesting insights, for instance, within the periodical exercise of the National Money Laundering Risk Assessment.

As for prudential supervision ends, a potential application of the indicator can be obtained by computing the financial exposure of each banking institution towards risky companies (i.e., the one likeliest to be infiltrated). In order to show its potential, we have merged at firm level our risk score with the 2021 data of the Bank of Italy's Central Credit Register regarding the loans granted from each financial intermediary to their customers.[2] This allows us to compute two exposure indicators for each intermediary: 1) the percentage of risky businesses over all business customers to which the intermediary has granted loans; 2) the percentage of loans granted to risky businesses over all granted loans. Figure 5 shows the frequency distribution of the two indicators across the Italian intermediaries, and reveals that none of the intermediary is exposed for more than 20 per cent in terms of number of firms. Nevertheless, 1.9 per cent of intermediaries have an exposure to firms at risk of being infiltrated by organized crime for more than 30 per cent in terms of amount of granted loans. This implies that, even if the number of firms at risk of being infiltrated is limited over all, the financial exposure is not negligible for the whole financial system since financial debt of risky firms is higher than non-risky firms' on average.

[2] There is a reporting threshold: a customer is reported if the sum to be repaid to the intermediary is equal to or over €30,000; this threshold is lowered to €250 if the customer has a bad debt.

7. Conclusions

In this study we develop a Machine Learning algorithm in order to detect legally registered firms potentially at risk of being infiltrated by organized crime. To this end, we employ a highly varied list of financial and budget indicators and variables, drawn by the most recent literature on criminal infiltration in real economy and computed for Italy by using data from three different sources: the National Official Business Register, the Central Credit Register of Bank of Italy and the National Institute of Social Security. A sample of almost 1,800 Italian firms known to be infiltrated almost with certainty is compared with stratified random samples of alleged legal firms in order to train and test the model. The main output of the algorithm is a risk score computed for a portion of the population of registered firms for which we have complete data for all variables and indicators.

The ML procedure can be deployed at an operational level for anti-money laundering and prudential supervision. In particular, a high probability score resulting from the algorithm could be used as a further selection criterion of firms to be regularly inspected in order to unmask illegal activities and as a red flag strengthening existing evidence of Mafia activities. Indeed, because of its limitations, the algorithm cannot by itself support allegations of criminal infiltrations within a firm without additional proofs.

As for prudential supervision ends, a potential application of the indicator can be obtained by computing the financial exposure of each banking institution towards companies potentially at risk of being infiltrated by organized crime.

We propose several opportunities for future research. First of all, we could expand the sample of infiltrated firms by resorting to other sources in order to evaluate potential improvements of the algorithm in terms of capacity to detect criminal firms among the alleged legal ones. Second, additional financial and non-financial information from other sources may be considered to improve the predictive power of the model. In particular, we could explore the use of indicators measuring the degree of opacity in firms' ownership structure. Third, other Machine Learning techniques, such as neural networks in particular, could be tested in order to find out how they perform in comparison with our algorithm. Finally, we could apply multiple imputation techniques for alleged legal firms in order to compute the risk score for a larger portion of Italian registered limited liability companies, thus widening the scope of application for anti-money laundering (AML) and prudential supervision purposes.

Table 1. List of variables/indicators

Category	Variable	Source
Sector of activity	3-digit NACE code	Central business register / National Statistical Institute
Location	Province of location	
Size	Assets	Central business register
	Revenues	
	Equity	
	Tangibles	
	Short term liabilities	
Equity and liquidity ratios	Cash over assets	Central business register
	Equity over assets	
	Equity over tangibles	
	Short-term assets over short-term liabilities	
	Revenues over assets	
	Working capital over assets	
Indebtedness	Leverage (granted loans over equity)	Central business register / Central Credit Register
	Granted loans over revenues	
	Net debt (granted loans - cash) over EBITDA	
Profitability	EBITDA over revenues	Central business register
	EBITDA over assets	
	ROI	
	ROE	
	ROA	
Investment (internal vs external resources) and cost structure	Tangibles over assets	Central business register
	Cost of rents and leases over revenues	
	Net purchases over revenues	
Employment	Cost of labour over number of employees	Central business register / National Institute for Social Security database
	Revenues over number of employees	
	Added value over number of employees	

Table 2. Descriptive statistics of raw dataset

	Infiltrated firms			*Non-infiltrated firms*		
	Mean	**S.D.**	**Percentage of missing data**	**Mean**	**S.D.**	**Percentage of missing data**
Sector of activity	-	-	0.7	-	-	0.4
Province	-	-	0.7	-	-	0.7
Assets	84,335	2,260,281	0.2	3,327	141,537	0.4
Revenues	27,913	584,948	20.3	2,956	87,075	16.3
Equity	25,189	670,848	1.0	975	184,156	1.2
Tangibles	81,854	2,149,639	10.8	1,801	105,632	10.6
Short term liabilities	18,452	445,134	3.4	1,603	192,114	3.0
Cash over assets	0.137	0.207	10.9	0.167	0.223	11.1
Equity over assets	0.114	0.926	1.1	0.221	0.533	2.4
Equity over tangibles	3.764	15.801	11.6	3.697	10.663	13.2
Short-term assets over short-term liabilities	4.099	15.578	4.4	2.739	6.178	5.7
Revenues over assets	1.143	1.347	20.3	1.134	1.039	18.0
Working capital over assets	0.026	0.720	4.4	0.103	0.481	5.7
Leverage (granted loans over equity)	4.029	72.123	1.0	3.709	84.262	1.2
Granted loans over revenues	2.959	64.402	20.3	2.925	104.487	16.3
Net debt (granted loans - cash) over EBITDA	1.822	94.030	24.9	1.090	205.800	24.9
EBITDA over revenues	-0.007	0.823	28.4	0.052	0.426	27.1
EBITDA over assets	0.032	0.282	19.4	0.058	0.219	20.2
ROI	0.197	1.164	28.4	0.218	0.694	29.1
ROE	0.096	1.367	9.3	0.104	0.779	10.8
ROA	-0.054	0.381	8.7	-0.025	0.220	10.4
Tangibles over assets	0.393	0.414	32.9	0.438	0.424	30.2
Cost of rents and leases over revenues	0.500	0.862	21.5	0.438	0.460	18.9
Net purchases over revenues	0.430	0.489	36.9	0.390	0.323	35.4
Cost of labour over number of employees	28.573	27.997	43.1	29.620	15.081	46.2
Revenues over number of employees	370.500	923.917	44.0	240.178	351.739	46.4
Added value over number of employees	9.589	87.151	43.3	9.832	36.366	46.6

Statistics are computed over all cases. Data have been subjected to a basic cleansing treatment in order to spot and resolve potential data inconsistencies. For infiltrated firms N=9,234; for non-infiltrated firms N=9,629,044.

Table 3. Percentage of records according to the frequency of missing items

Number of missing items	Infiltrated firms	Non-infiltrated firms
Zero: complete cases	36.5	33.5
1 to 5	31.1	34.1
5 to 10	15.6	17.1
11 to 15	11.4	10.5
16 to 20	5.0	4.2
20 to 27	0.4	0.6
All	**100.0**	**100.0**

Table 4. Structure of the final dataset

Year	Infiltrated firms	Non-infiltrated firms
	Number of records	
2010	800	265,245
2011	773	271,678
2012	693	271,643
2013	597	272,655
2014	548	287,951
2015	583	307,054
2016	553	356,404
2017	573	370,876
2018	565	384,994
2019	557	389,319
2020	52	46,385
Total	**6,294**	**3,224,204**
	Number of firms	
Total	**1.786**	**746,843**

Table 5. Descriptive statistics of the final dataset

	Infiltrated firms		*Non-infiltrated firms*	
	Mean	**S.D.**	**Mean**	**S.D.**
Assets	122,822	2,734,295	3,291	110,595
Revenues	32,573	632,478	2,856	46,871
Equity	36,487	808,443	1,131	61,245
Tangibles	106,775	2,458,220	1,472	87,084
Short term liabilities	25,878	529,665	1,462	37,045
Cash over assets	0.122	0.172	0.137	0.167
Equity over assets	0.158	0.532	0.232	0.282
Equity over tangibles	3.764	14.874	3.389	8.972
Short-term assets over short-term liabilities	2.309	7.936	1.767	2.417
Revenues over assets	1.239	1.341	1.368	0.939
Working capital over assets	0.052	0.493	0.144	0.335
Leverage (granted loans over equity)	2.841	24.782	2.955	13.642
Granted loans over revenues	1.247	12.937	0.391	1.196
Net debt (granted loans - cash) over EBITDA	2.261	87.316	1.954	56.859
EBITDA over revenues	-0.003	0.778	0.065	0.218
EBITDA over assets	0.061	0.246	0.091	0.163
ROI	0.230	1.148	0.233	0.581
ROE	0.140	1.331	0.126	0.685
ROA	-0.019	0.289	0.013	0.137
Tangibles over assets	0.399	0.897	0.394	0.384
Cost of rents and leases over revenues	0.472	0.801	0.323	0.237
Net purchases over revenues	0.416	0.604	0.378	0.279
Cost of labour over number of employees	29.016	26.861	29.613	13.894
Revenues over number of employees	374.119	899.867	225.633	296.937
Added value over number of employees	10.999	88.548	10.571	29.712

For infiltrated firms records with missing data for the province and sector categorical variables or with more than 6 missing items have been removed; a fully conditional specification (FCS) procedure has been applied to impute the remaining missing data.
For non-infiltrated firms statistics are computed only over complete cases.

Figure 1. Pairwise linear correlation of the variables

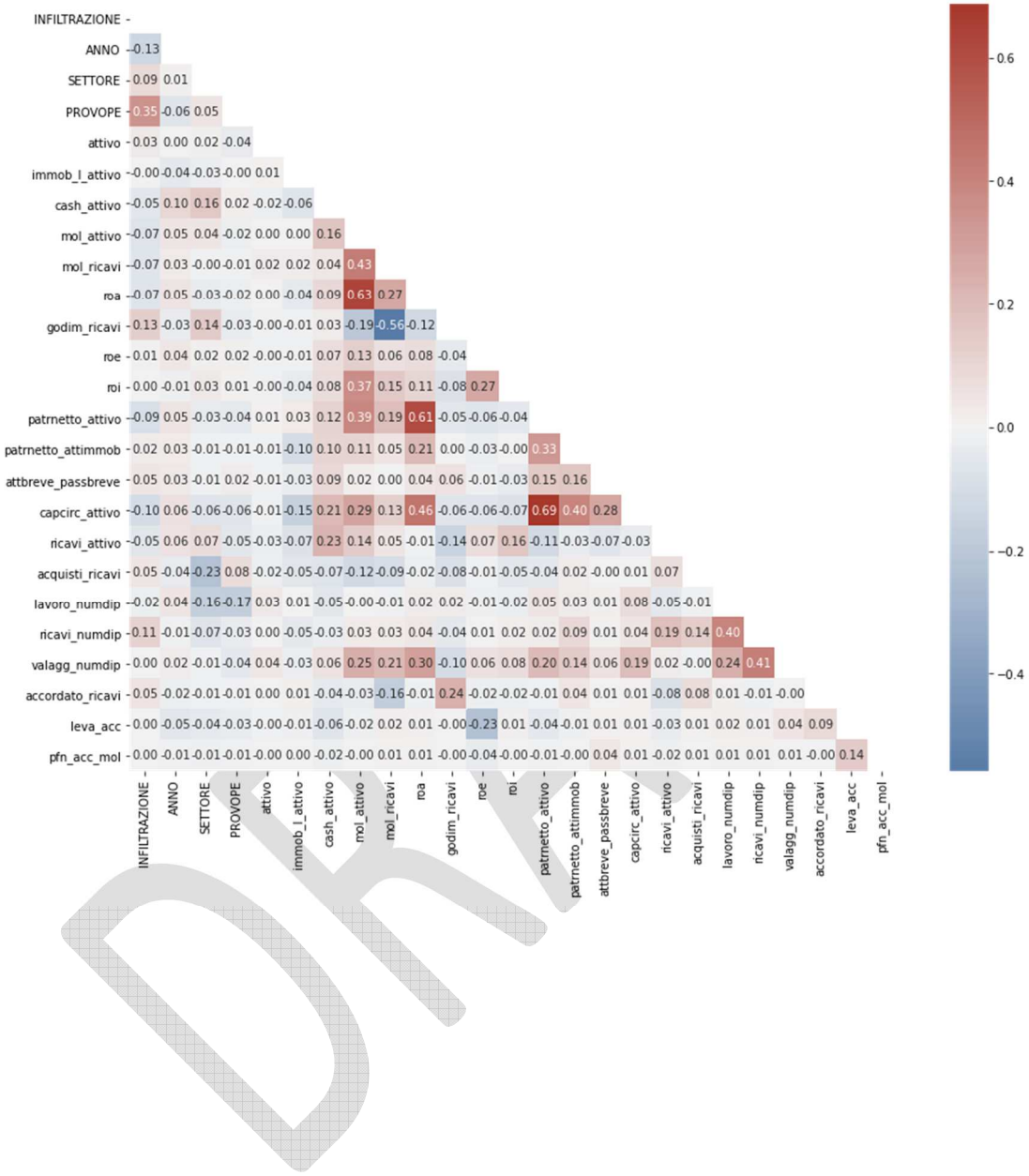


Figure 2. Model comparison
(SAS VIYA® model studio)

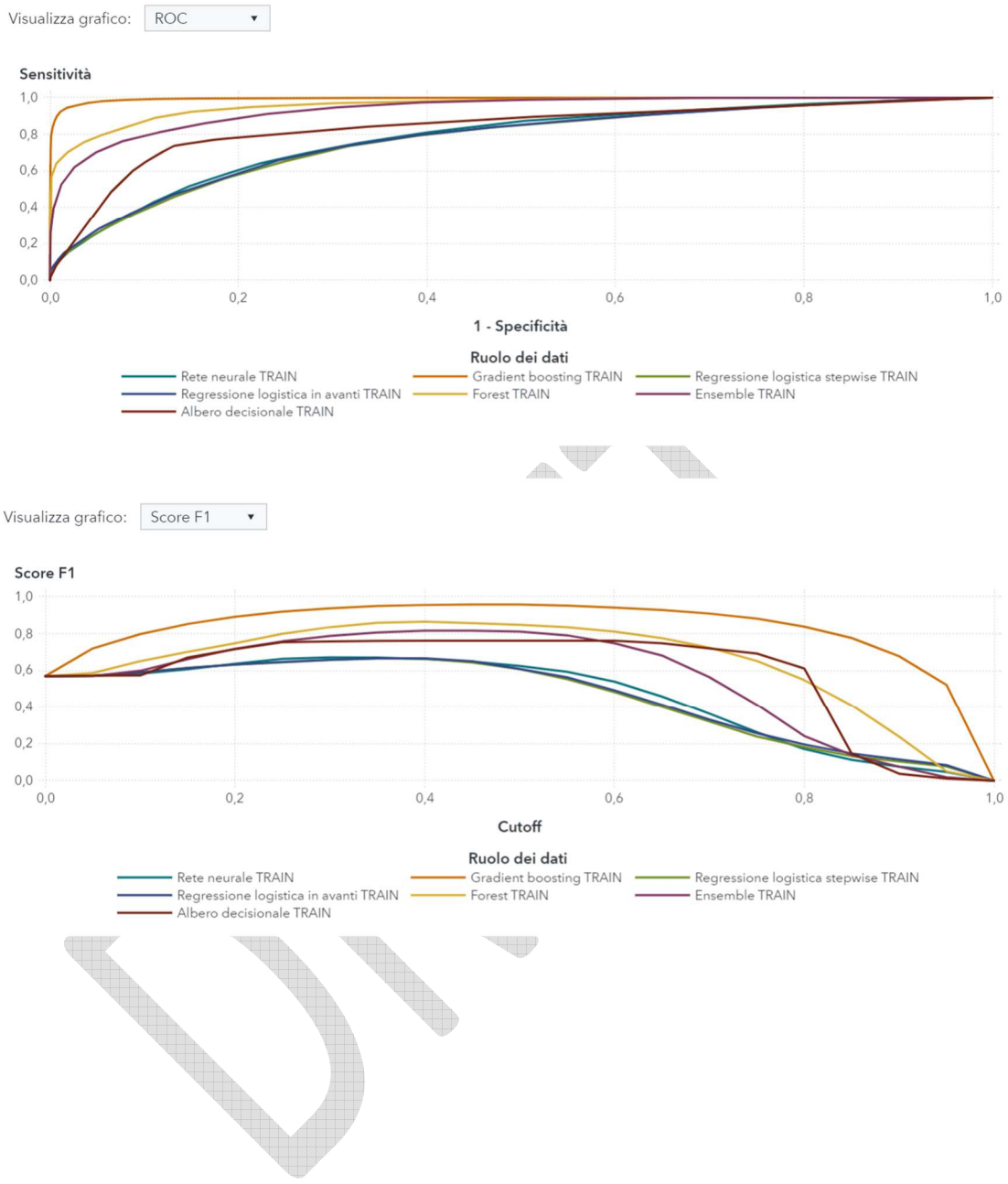
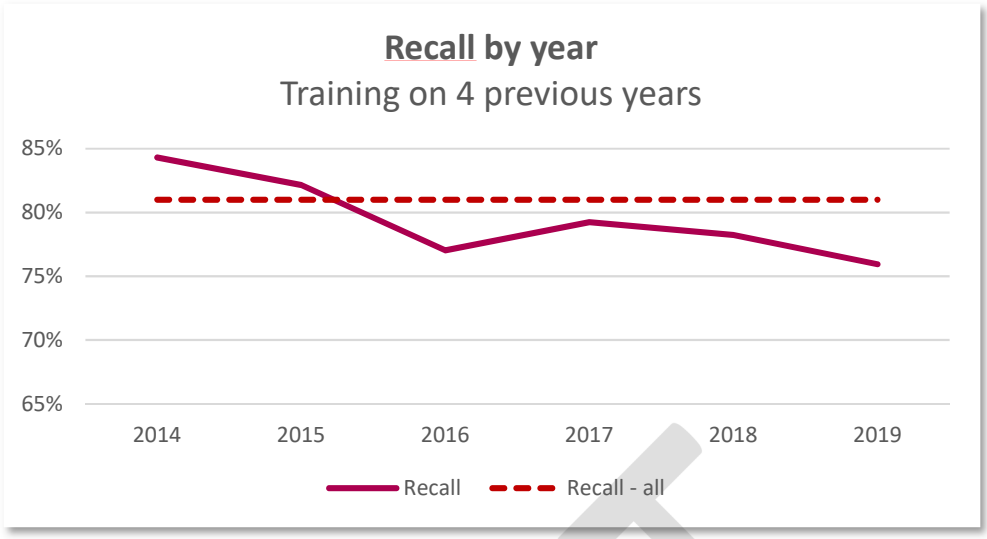


Figure 3. Stability test
(Recall & Precision)



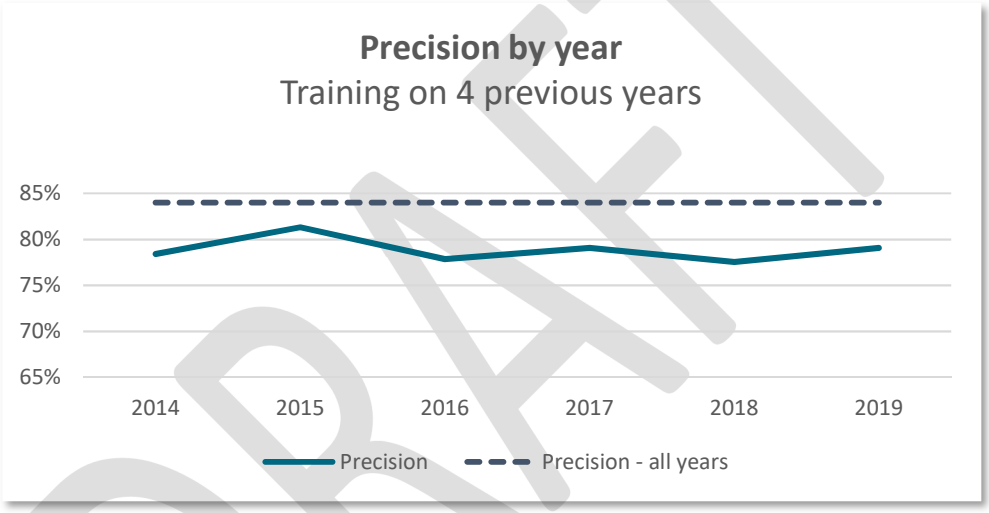


Figure 4. Variables relative relevance in the computation of the indicators

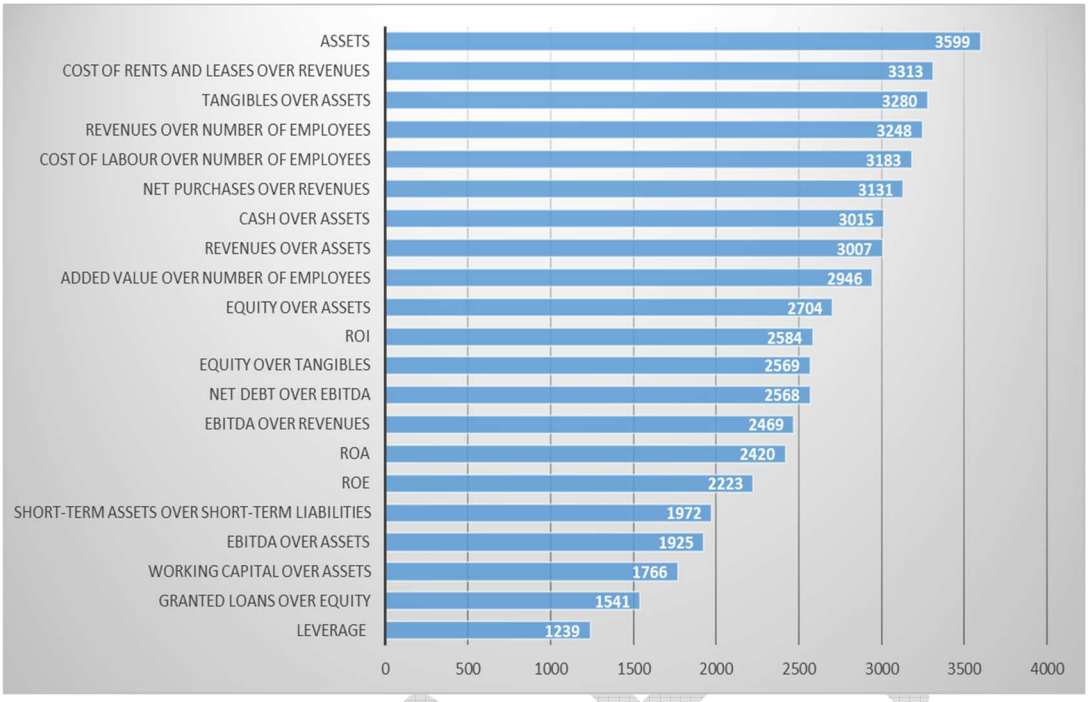


Table 6. Frequency distribution of estimated risk score – years 2018-2020

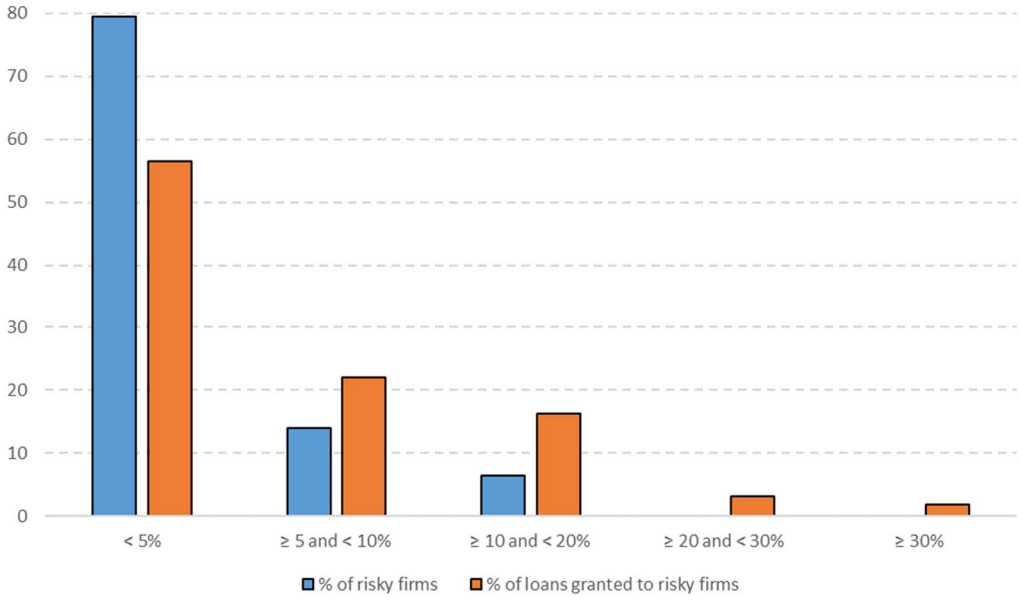
Risk score	N	%
Up to 0.5	423,360	89.6
From 0.5 to 0.8	21,983	4.7
From 0.8 to 0.95	14,571	3.1
From 0.95 to 0.99	7,797	1.7
Over 0.99	4,828	1.0
Total	**472,539**	**100.0**

Table 7. Frequency distribution of estimated risk score by geographic area

	Firms at risk of being infiltrated (risk score > 0.5)		Total firms in the sample		Percentage of firms at risk of being infiltrated
	N	%	N	%	%
North-West	4,830	9.8	128,986	27.3	3.7
North-East	3,282	6.7	98,990	20.9	3.3
Center	6,540	13.3	109,196	23.1	6.0
South and Islands	34,527	70.2	135,367	28.6	25.5
Total	**49,179**	**100.0**	**472,539**	**100.0**	**10.4**

Figure 5. Financial exposure of intermediaries to firms at risk of being infiltrated

(percentage of intermediaries)



References

Barone G. and S. Mocetti (2014), *"Natural Disasters, Growth and Institutions: A Tale of Two Earthquakes."* Journal of Urban Economics, 84(C), pp. 52-66.

Bianchi P., A. Marra, D. Masciandaro, N. Pecchiari (2020), OC and Firms' financial statements: Evidence from Criminal Investigation in Italy. *Bocconi Legal Studies Research Paper* No. 2017-59.

Brand, J. P. L. (1999). "Development, Implementation, and Evaluation of Multiple Imputation strategies for the Statistical Analysis of Incomplete Data Sets." Ph.D. thesis, Erasmus University.

Chengwei L., Yixiang c., Kazmi S.H.A. and Hao, F. (2015), Financial Fraud Detection Model Based on Random Forest. *International Journal of Economics and Finance*, vol. 7, no. 7,

Daniele, G. and B. Geys (2015), *"OC, Institutions and Political Quality: Empirical Evidence from Italian Municipalities."* Economic Journal 125(586), pp. 233-255.

Daniele V. and U. Marani. (2011), *"OC, the quality of local institutions and FDI in Italy: A panel data analysis."* European Journal of Political Economy, 27(1), pp. 132-142.

De Feo G. and G. De Luca (2013), "Mafia in the ballot box." *DEM Working Paper Series* No. 57.

Donato L., A. Saporito, A. Scognamiglio (2013), "Aziende Sequestrate Alla Criminalità Organizzata: Le Relazioni Con Il Sistema Bancario." *Bank of Italy Occasional Paper* No. 202.

De Simoni M. (2022), The financial profile of firms infiltrated by organised crime in Italy. Forthcoming in *UIF, Quaderni dell'antiriciclaggio, Collana Analisi e Studi.*

Fabrizi M., P. Malaspina, A. Parbonetti (2017), Caratteristiche e modalità di gestione delle aziende criminali. *Rivista di studi e ricerche sulla criminalità organizzata*, 3(1), pp. 47-66.

Maka K., S. Pazhanirajan and S. Mallapur (2020), *Selection of most significant variables to detect fraud in financial statements*. Materials Today: Proceedings

Mirenda L., S. Mocetti, L. Rizzica (2019), *The real effects of 'ndrangheta: firm-level evidence*, *Temi di Discussione, Banca d'Italia* N. 1235.

Peri G. (2004), "*Socio-Cultural Variables and Economic Success: Evidence from Italian Provinces 1951-1991."* B.E. Journal of Macroeconomics, 4(1), pp. 1-36.

Pinotti P. (2015), *"The economic costs of OC: Evidence from Southern Italy*." Economic Journal 125(586), pp. 203-232.

Ravenda, D., Argilés-Bosch, J.M. and Valencia-Silva, M.M. (2015), Detection Model of Legally Registered Mafia Firms in Italy. *European Management Review*, 12: 23-39.

Sadgali, I., N. Sael and F. Benabbo (2019), *Performance of machine learning techniques in the detection of financial frauds*. Procedia Computer Science.

Sharma, A. and P .Panigrahi (2013), A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. *International Journal of Computer Applications* vol. 39, n.1

Transcrime (2015), Gli investimenti delle mafie. Progetto PON sicurezza, 2007-2013. Transcrime e Università Cattolica del Sacro Cuore.

UNODC (2011). Estimating illicit financial flows resulting from drug trafficking and other transnational organized crimes. Research report, United Nations Office on Drugs and Crime.

Van Buuren, S. (2007). "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification." Statistical Methods in Medical Research 16:219–242.

Wyrobek J. (2020), Application of machine learning models and artificial intelligence to analyze annual financial statements to identify companies with unfair corporate culture. *Procedia Computer Science* 176, 3037–3046.