No 488

FEBRUARY 2022

CefES

Center for European Studies

PAPER SERIES

# Using Accounting Information to Predict Aggressive Tax Placement Decisions by European Groups

Matteo Borrotti, Michele Rabasco, and Alessandro Santoro

The Center for European Studies (CefES-DEMS) gathers scholars from different fields in Economics and Political Sciences with the objective of contributing to the empirical and theoretical debate on Europe.

DIPARTIMENTO DI ECONOMIA, METODI QUANTITATIVI E STRATEGIA DI IMPRESA

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

# Using Accounting Information to Predict Aggressive Tax Placement Decisions by European Groups

Matteo Borrotti[1], Michele Rabasco[1], and Alessandro Santoro[1]

[1]DEMS, CEFES & Datalab, University of Milano-Bicocca, Milan, Italy

**Abstract**

Aggressive tax planning (ATP) consists in taxpayers' reducing their tax liability through arrangements that may be legal but are in contradiction with the intent of the law. In particular, ATP by multinational groups (MNE) is a source of major concern. In this paper we consider the MNE's decision to locate or to maintain a company in a tax haven as a relevant symptom of ATP. The research question we want to address is whether this decision can be predicted using publicly available accounting information. We use ORBIS database and we focus on European MNEs. We observe that, in 2021, slightly less than 40% of European MNEs have a company located in a tax haven. Thus, for a tax authority it would be difficult, without a specific analysis, to identify riskier MNEs. We find that a random forest model that uses accounting information for years between 2015 and 2019 predicts reasonably well the decision to locate (or maintain) a company in a tax haven in 2021. Using this model in 2019, a tax authority could have identified almost 80% of European MNEs that were going to locate or maintain a company in a tax haven in 2021. We observe that the most important variables for prediction are those associated to the size of the group, to its positive profitability and to its financial structure, while individual time-invariant features are less relevant. We also find that the predictive performance of the model is maximized when the information is taken from the time subset 2017-2019 and that most important predictors for the risk of using tax

havens are also good predictors for the level of intensity of such a use, as measured by the share of subsidiaries located in tax havens. The main policy implication of these results is that (European) non-tax havens could effectively anticipate (and prevent) the decision to locate (or maintain) companies in tax havens, and shape their policies accordingly, with particular reference to cooperative compliance schemes. These policies are more credible in the context of renewed international cooperation in the design of corporate tax rules, and in particular, of the implementation of Pillar Two within the European Union.

**Keywords: Aggressive Tax Planning, European Multinationals, Machine Learning**

## 1 Introduction

In general, aggressive tax planning (ATP) consists in taxpayers' reducing their tax liability through arrangements that may be legal but are in contradiction with the intent of the law. According to the definition provided by the European Commission in its recommendation of December 2012, ATP consists "in taking advantage of the technicalities of a tax system or of mismatches between two or more tax systems for the purpose of reducing tax liability. It may result in double deductions (e.g. the same cost is deducted both in the state of source and residence) and double non-taxation (e.g. income which is not taxed in the source state is exempt in the state of residence").

In a paper which implements this definition (Meldgaard et al., 2015), the aforementioned "technicalities" are identified in a list of ATP indicators and, in turn, this list is used to assess to what extent a Member State can be thought to favour ATP. The differentiation between countries that can be classified as tax havens from the others, at a world level, is also at the heart of the *Missing Profits* project (Tørsløv et al., 2018) where profits shifted from non-tax havens to tax havens, and the associated loss of potential corporate tax revenues, are precisely quantified. Both these approaches suggest that the decision of a multinational to locate a subsidiary in a country offering a favourable tax regime (tax haven for the sake of simplicity hereinafter) is one of the main symptoms of ATP, albeit not the only one.

The aim of this paper is to analyze whether the decision to locate a company in a tax haven can be predicted using publicly available accounting information, namely that included

in the ORBIS database. ORBIS contains information on the performance and the ownership structure of more than 400 millions companies worldwide. However, it is well known (Tørsløv et al., 2018) that ORBIS this information is incomplete with respect to opaque jurisdictions, namely in tax havens. More precisely, ORBIS provides information about the location of a company in a tax haven, but not about the amount of profits that are diverted to tax havens.

In this paper, we look only at the location decision, therefore we rely on ORBIS, which offers the advantage to be a publicly available database. Clearly, our approach and methodology can be applied to cases where the tax authority can use more detailed information, such as that provided by the Country-by-Country reporting system. We focus on the European context, for two reasons. First, there is a inter-European competitive dimension, namely European countries that attract tax bases and tax revenues from MNEs headquartered in another European country by applying reduced tax rates and favourable tax regimes. Second, and related, the European Union has been particularly active in the implementation of policies designed to reduce tax competition, such as cooperative compliance and, more recently (December 2021) the implementation of a minimum effective tax rate.

In 2021, slightly less than 40% of European groups have a company located in a tax haven.

To run our prediction exercise we test two types of machine learning models, decision trees and random forests, that allow the analysis of the variable importance. Our approach is data-driven, so that we are not testing an economic model. In this context, the possibility to quantify the variable importance allows to gain some insights on the determinants of the ATP activity. We find that a random forest model that uses accounting information for years between 2015 and 2019 predicts reasonably well the decision to locate (or maintain) a company in a tax haven in 2021. In particular, after tuning the probability threshold at 40%, the best model has accuracy of 77%, recall of 75% and a F1 score of 70%. We observe that the most important variables for prediction are those associated to the size of the group, to its positive profitability and to its financial structure, while individual time-invariant features are less relevant. We also find that the predicted performance of the model is maximized when the information is taken from the time subset 2017-2019. Overall, these results suggest that the use of tax havens is associated with the MNE's ability to join a given size and profitability. Clearly, the non-causal nature of our analysis prevents us from assessing whether ATP is more a cause or a consequence

of increased size and profitability. Both explanations could hold. ATP could either be a 'stage of evolution' of MNE, i.e. a consequence of the profitability and of the size of the activity, or 'a driver', i.e. a cause of the pattern observed in size and profitability.

However, our main interest here is to contribute to the design of the policies to prevent ATP.The main policy implication of our results is that (European) non-tax havens could effectively anticipate (and prevent) the decision to locate (or maintain) companies in tax havens. This allows tax administrations to increase the efficiency of their approach to the management of ATP risk. More in particular, our results suggests that the prediction of location decision can be useful in the design and implementation of Cooperative Compliance (CoCo hereinafter) schemes.

The term cooperative compliance was proposed by OECD (2013) as an evolution of the "enhanced relationship" between tax authorities and taxpayers advocated by the Forum on Tax Administrations (FTA) in 2008. CoCo schemes cover a wide range of different policies adopted by 21 FTA member countries to increase voluntary tax compliance by large businesses, and especially by multinationals. In particular, CoCo schemes are aimed to reduce BEPS practices, i.e. base erosion -from high- taxed companies- and profit shifting -towards low taxed companies- that are accomplished within multinationals using a variety of schemes (OECD, 2013). Also, CoCo schemes may be interpreted as ways to address the issue of inequality in the distribution of the tax burden between large multinational companies and small domestic ones. When a tax authority and a company decide to enter in a CoCo agreement they commit to a mutual information disclosure. On the one hand, businesses disclose information about the schemes they use to manage tax-related issues within their organization. On the other hand, tax authorities disclose their views about the legitimacy of these schemes and grant some benefits, such as a privileged access to ruling and a reduction of the probability to be audited. The CoCo agreement can then be summarized as "transparency in exchange of certainty".

Now, by applying our results to Italy, we show that CoCo schemes could be designed from an ex-ante perspective, i.e not only conditioning the benefits for MNEs to the disclosure of future schemes, but also to the presence of a high level of risk to locate or maintain subsidiaries in tax havens.

## 2 Related literature

International ATP by multinationals (MNEs) has been extensively studied by the accounting and economics literature. In particular, the paper whose aim was closest to ours is that by Newberry and Dhaliwal (2001).They define the placement decision as the decision by a US multinational to issue a bond through a subsidiary located in a different country rather than through the US parent itself, and they try to explain this decision on the basis of a number of tax and non-tax variables. Since then, the literature has somewhat departed away from the placement decision and it has rather focussed on the determinants of income shifting by US multinationals. Papers surveyed by De Simone et al. (2019) have looked at the impact of specific financial incentives or constraints, at income shifting between specific countries or between affiliates of a given profitability level, but they all suffered from limited information about the dependent variable: actual income shifting can be measured only using IRS data on intercompany payments.

The importance of the decision to locate a company in countries which provide favourable tax treatments is now very clear. For example the (in)famous *Double Irish-Dutch Sandwich Scheme* can essentially be summarized as the creation of "two Irish affiliates and a Dutch shell company squeezed in between" (Zucman, 2014). The first Irish company is created to transfer the property of intangibles from the US to Bermuda, the second Irish company is used to sell the licensing rights to subsidiaries located everywhere in the world and the Dutch company is created so that the royalty paid by the second Irish company to the former is not taxed. Clearly, the exact amount of different transactions should be known to estimate the loss of tax revenues suffered by the US. As Tørsløv et al. (2018) show, the information about the magnitude of capital flows which is available in ORBIS is highly unreliable. However, in the *Double Irish-Dutch Sandwich Scheme* the threat to the US revenue is signalled by the very creation of the three subsidiaries in the three tax havens (Bermuda, Ireland and the Netherlands), which is a necessary condition for the scheme to be implemented.

Therefore, it is fundamental for every country which is under the threat of tax competition to be able to predict ex-ante, and not only to assess ex post, the probability that a 'national' MNE (i.e a MNE whose fiscal residence is within that country) diverts some of its tax base to

tax havens through creation of affiliates. Likely, a tax administration can have different sources to predict such a risk. However, in this paper, we use a publicly available database, ORBIS, that, for reasons discussed in the Introduction, can be considered as a reliable source for the location decision.

There are three main differences between our approach and that followed by the previous literature on aggressive tax planning.

First, we focus on European groups (i.e. groups to which companies listed in a stock market and resident in EU-27 belong) and we look mainly at within-Europe tax competition. Both these features of our research design are a novelty in the literature, which is mainly focused on US multinationals and on threats posed by non-European tax havens.

Second, our dependent variables are designed to capture the location decision and we do not conjecture that such a location is associated with a specific type of tax planning, such as the one investigated by Newberry and Dhaliwal (2001). To construct a list of tax havens we rely on Meldgaard et al. (2015), Tørsløv et al. (2018) and European Council (2021 [Online]).

Third we use a machine learning approach and, rather than pre-selecting explanatory variables that should be associated to ATP according to a given theory or approach and then test the significance of their correlation, we let the data reveal which are the best predictors of the decision to locate or maintain a company in a tax haven. To keep the results interpretable, we use decision trees and random forest models.

## 3 Dataset

We use data from the ORBIS database [1] which provide firm-level data on over 400 million companies and entities worldwide, along with detailed information on company structure.

The starting point of the analysis is the identification of the active listed companies, resident in one of the 27 EU countries[2]. Among them, we consider for the analysis only those ORBIS identifies as corporate companies. These companies are labelled as LISTED, because they are listed in any stock market, not necessarily one in the European Union.

[1] ORBIS is a commercial database provided to the OECD by the electronic publishing firm, Bureau Van Dijk.

[2] These countries are: Austria, Belgium, Bulgaria, Cyprus, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain and Sweden.

The second step is the identification of the group to which every LISTED company belongs. In turn, this requires a) the identification of the Global Ultimate Owner (GUO) and b) the identification of all companies belonging to the same UO (subsidiaries, in short). Both these steps are accomplished using the data available in the spring 2021.

ORBIS identifies UOs analysing the shareholding structure of a company. It looks for the shareholder with the highest or total (direct plus indirect) percentage of ownership: the minimum percentage of control in the path from a subject company to its Ultimate Owner must be 50.01% [3]. A company is considered Ultimate Owner (UO) if it has no identified shareholders or if its shareholder's percentages are not known.

We refer to a Global UO (GUO) to distinguish it from a Domestic UO (DUO), that is the highest company in the path between a subject company and its Global UO located in the same country as the subject company. Starting from the LISTED companies, we are able to identify 4,031 GUOs subdivided as follows: 2,654 OWN_GUOs, if the GUO is any LISTED company, 815 OTHER_COMPANY_GUOs, if the GUO is not a LISTED company and 562 OTHER_INSTITUTIONS_GUOs, if the GUO is not a company (partnership, family, single person).

We then identify all companies that have the same GUO [4] and obtain a set of 127,827 subsidiaries. The composition of the final dataset of 131858 companies which data are updated to at least 2015 is shown in table 1

[3] The Orbis guide specifies that the procedure for the Ultimate Owner identification "intends to track control relationships rather than relationships that do not allow the shareholder to take a decision in the company; when there are 2 categories of shares split into Voting/Non-voting shares, the percentages that are recorded are the ones attached to the Voting shares category".

[4] With reference to the subsets of OWN_GUOs and OTHER_COMPANY_GUOs, ORBIS allows to extract the whole group of companies affiliated with them; this corporate group is composed of all the subsidiaries that are ultimately owned by the subject company's GUO. Unfortunately, ORBIS does not provide the same information for the set of OTHER_INSTITUTIONS_GUOs, because they are not companies. For the latter, the group reconstruction process is shown in the Appendix **??**.

**Table 1:** Dataset composition-GUOs and subsidiaries

| category | Freq | % |
|---|---|---|
| OWN_GUO | 2,654 | 65.8 |
| OTHER_COMPANY_GUO | 562 | 14.0 |
| OTHER_INSTITUTIONS_GUO | 815 | 20.2 |
| All GUOs | 4,031 | 100 |
| SUBS_OWN_GUO | 84,066 | 65.8 |
| SUBS_OTHER_COMPANY_GUO | 29,406 | 22.8 |
| SUBS_OTHER_INSTITUTIONS_GUO | 14,355 | 11.3 |
| All SUBS | 127,827 | 100 |

Companies belonging to groups whose GUO is a LISTED company represent approximately 2/3 of the dataset. Groups with a non-company GUO are relatively more frequent but less populated than groups with non listed GUOs. More details on the numerosity of the groups are provided by table 2. The table also illustrates considerable variability in group numerosity; our dataset is composed of both GUOs for which we cannot identify any company within the group that has financial data updated to at least 2015 (group numerosity = 1, the GUO itself) and groups of considerable numerosity (up to 2,114 companies) [5].

**Table 2:** Group numerosity for GUOs category

| category | min | q1 | median | mean | q3 | max |
|---|---|---|---|---|---|---|
| OWN_GUO | 1 | 2 | 6 | 32.68 | 20 | 2,114 |
| OTHER_COMPANY_GUO | 1 | 5 | 15 | 53.33 | 50 | 1,542 |
| OTHER_INSTITUTIONS_GUO | 1 | 1 | 5 | 18.62 | 15 | 689 |

## 4 Definition of variables and choice of prediction methods

In our prediction exercises, the target variables are the dichotomic information about a company belonging to the group (either the GUO itself or a subsidiary) in a tax haven in 2021 and the share of total companies which is located in a tax haven.

For each company, we observe the country of fiscal residence in 2021 and therefore we can

[5]However, it should be noted that the lower average numerosity of groups having an OTHER_ISTITUTION_GUO may be the result of the different logic used to construct these groups, made necessary by the impossibility of extracting the corporate composition directly, in ORBIS.

classify it as as a subsidiary located in a tax haven or in a non-tax haven. More precisely, the list of countries with a high risk of aggressive tax planning, henceforth (ATPC), is constructed as follows. The starting point is the list of 12 Countries identified as EU list of non-cooperative jurisdictions for tax purposes by the European Council in the official journal of the organization (European Council, 2021 [Online]). These countries are: American Samoa, Anguilla, Barbados, Fiji, Guam, Palau, Panama, Samoa, Seychelles, Trinidad and Tobago, U.S. Virgin Islands, and Vanuatu. To these, we add the Cayman Islands and Oman, which were present until the previous revision. To these 14 countries, we add those countries for which at least 13 of the 33 indicators of aggressive tax planning used by the European Commission Meldgaard et al. (2015) are present. This leads to the selection of the following 7 countries: Belgium, Cyprus, Hungary, Luxembourg, Latvia, Malta, and the Netherlands. Finally, we add the countries present in the list of prepared by Tørsløv et al. (2018), not yet present in our list namely Switzerland and Ireland, getting a total of 23 ATPC.

Within our dataset, the GUOs are distributed geographically as in figure 1.

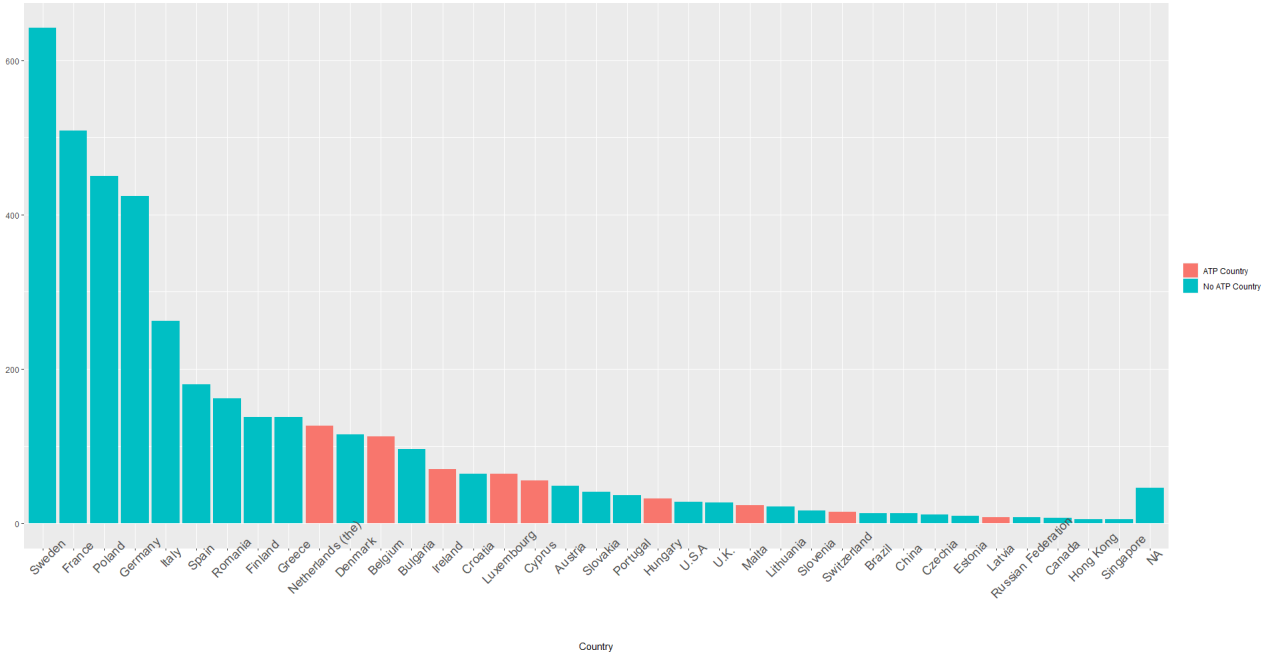


**Figure 1:** Number of GUOs within each country

The figure shows the representativeness of countries within the dataset (for visibility reasons, only countries with more than 5 GUOs are shown). Net of Sweden, the number of GUOs within each Eurozone country appears to be proportionate to the size of the country itself. The large number of GUOs resident in Sweden (642), is explained by the fact that, in the

time period considered, there has been a boom in start-ups, of a corporate nature, linked to new technologies. Countries with fewer than 6 GUOs are grouped in the NA category. The countries in red are those that we have defined as ATPC.

In addition, our dataset presents 14,781 (11.5%) subsidiaries residing in an ATP country, which in the 90% of the cases is an EU country and in the remaining cases a non-EU European country. Only a residual fraction of the subsidiaries reside in an ATP country outside the EU, and the reason for this is to be found in the opacity of the tax systems of these countries which does not allow the identification of companies located there.

We classify GUOs based on the presence of at least one company of the group (either the GUO itself, or a subsidiary) in one of the ATPC. This results in a dichotomic variable (at least one company in ATPC and No company in ATPC) and which we will consider as target variable for our first prediction exercise. In our dataset, this breakdown results in 1,574 (39.05%) GUOs with at least one company located in an ATPC and 2,457 (60.95%) with no company located in an ATPC. It is possible to make a further partition, considering only the 3,522 GUOs who are not resident in an ATPC country. Among these, 1,065 (30.24%) control a company located in an ATPC country (table 3).

**Table 3:** GUOs with at least one company located in an ATP Country

| | Freq | % |
|---|---|---|
| GUO is in ATPC | 1,574 | 39.05 |
| GUO is not in ATPC | 1,065 | 30.24 |

However, we cannot exclude that the subsidiary was part of the same group also *before* 2021. Therefore, our dependent variable will be defined as the probability to locate or to maintain a subsidiary in a ATPC (a tax-haven) in 2021.

The predictors are financial variables extracted from GUO's financial statements from 2015 to 2019.

After the selection, the dataset comprises, for every GUO three main types of variables a) asset variables (intangible and tangible assets, stocks, capital, debts, loans ecc.), b) variables from the profit and loss section (operating revenue, sales, cost of employees, taxation, profit and loss for the period ecc.) and c) other variables including sector of activity, country of residence,

legal form etc.

As for prediction methods, only Decision Tree (DT) (Hastie et al., 2001) and Random Forest (RF) (Breiman, 2001) are considered for the classification purpose. As we discuss in subsection 5.1, these two methods allow an intuitive metric of variable importance and, therefore, are easier to interpret.

DT belongs to the family of supervised learning approaches. The goal of using a DT is to create a training model that can be used to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data). DT is the graphical representation of all the possible solutions to a decision based on certain conditions. In this algorithm, the training data is splitted into two or more homogeneous sets. For more details, please see Hastie et al. (2001).

RF models are non parametric models that can be used both for regression and classification, as DT. RFs are ensemble methods belonging to the class of Bagging methods. In ensemble methods, a series of learners (*i.e.* models) is used in order to enhance the final prediction performance. Starting from the training set, a series of smaller training sets is sampled and used to train a set of weak learners. Then, all weak learners are used to predict the class of a new observation. In classification task, the predicted class will be selected in accordance with the majority voting rule. In RF, the weak learners are DT. For more details, please see Breiman (2001).

### 4.1 Pre-processing phase

The considered dataset is composed by unique observations, however the amount of null values is quite significant. Typically, two strategies of handling missing values are recommended: (a) passive and (b) active strategy.

Considering option (a), observations having a percentage of missing data greater than 50% are removed for the dataset. The same strategy is applied to remaining observations having variables with a percentage of missing data greater than 50%. Two different approaches are applied as active strategy while considering quantitative variables. In first instance, value of year $t+1$ is imputed on the same variable on year $t$, if the value of year $t$ is missing. In second instance, the median of each variables is used to fill the remaining missing values.

After dealing with quantitative missing data, the one-hot encoding technique is applied to qualitative variables to convert them into binary variables, hence increasing the predictive performance of the machine learning models. Basically, for each level of the qualitative feature a dummy variable is built. At the end, all qualitative variables are standardized in accordance with the common formula, $\frac{x-\mu}{\sigma}$ where $\mu$ is the mean of a considered variable and $\sigma$ the standard deviation of the same variable.

## 5 Predicting the risk of using tax havens

After the pre-processing phase, the final dataset is composed by 3575 observations (*GUOs*). Each observation is described by 232 variables, among them 29 are categorical variables (after one-hot encoding transformation), 202 are quantitative variables and one is the target variable. The target variable is a binary feature that leads to a classification problem. An observation is classified with label 1 if at least one company of the group is in one of the ATP countries, 0 otherwise. Almost 61% of observations have label 0 and the remaining observations have label 1. Input variables are related to the time interval from 2015 to 2019. The target variable is computed on year 2021.

As in a classical machine learning procedure, the dataset is splitted in training (80%) and test (20%). At the beginning of the procedure, DT and RF are compared in order to select the most powerful approach. For this purpose, a 10-folds cross-validation technique is used to identify the best model. Table 4 summarizes the main performance metrics, in average.

The social costs of type I errors, i.e. false positives, and type II errors, i.e. false negatives, vary across problems that are under examination, so that, in principle, it is appropriate to consider both of these errors. They are jointly measured by *accuracy* which is just the ratio between correctly classified cases and the total of cases. RF is slightly better with respect to DT in terms of accuracy leading to an higher ability to correctly predict both classes and to reduce both types of errors.

Now consider that *Recall* (also known as sensitivity or true positive rate) measures the ratio between true positive (observations correctly predicted on positive class - 1 -) and the number of actual observations on positive class. *Precision* measures the ratio between true positive (observations correctly predicted on positive class - 1 -) and the number of predicted

observations on positive class. A good performance metric that takes in consideration both precision and recall of machine learning approach is the the F1 score that is the harmonic mean of precision and recall. RF reaches a value of F1 equal to 0.678 while DT a value equal to 0.643, and again, RF records both a higher precision and a higher recall, therefore RF is selected for the next research phase.

**Table 4:** Performance metrics on validation set

| | Accuracy | Recall (Sensitivity) | Precision | F1 score |
|---|---|---|---|---|
| Decision Tree | 0.745 | 0.584 | 0.721 | 0.643 |
| Random Forest | **0.774** | **0.604** | **0.779** | **0.678** |

RF, as many other machine learning approaches, is characterized by several hyper-parameters which dramatically influence the performance of models. Therefore, careful tuning of these hyper-parameters, i.e., hyper-parameter optimization, is important. For this reason, a simple *grid-search* technique is applied in order to find the best set of hyper-parameters while considering F1 score as the objective function that should be maximise. As initial study, only the number of trees to grow (*ntree*) and the number of variables randomly sampled as candidates at each split (*mtry*) are involved in the optimization. Table 5 reports the hyper-parameters involved in the optimization.

**Table 5:** Hyper-parameters involved in the optimization.

| Hyper-parameter | Range | Step |
|---|---|---|
| ntree | $\{250, 500, 1000, 1500\}$ | – |
| mtry | $\{5 - 18\}$ | 1 |

In order to selected the best setting, a 10-fold cross-validation approach is used on the initial training set. In Figure 2, the F1 score for each combination of values is reported. The best setting is $ntree = 500$ and $mtry = 12$, that leads to an accuracy of 0.778, recall of 0.612, precision of 0.781 and F1 score of 0.684.

In our context, recall is important because every MNE which is going to use a tax haven, without being predicted as such, represents a serious risk for the efficiency and the reputation

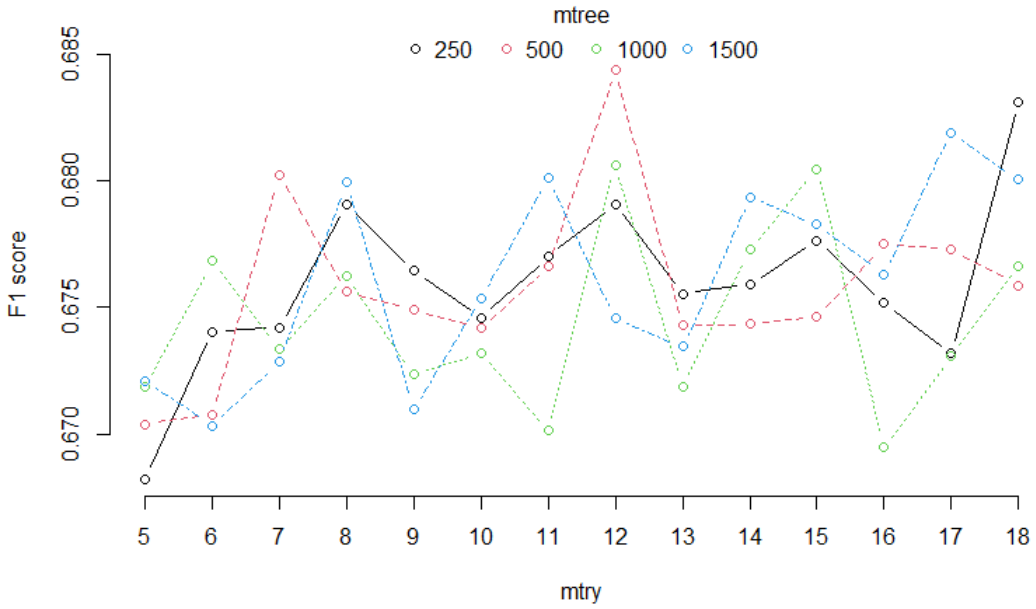


**Figure 2:** Hyper-parameters tuning performance with respect to F1 score.

of the tax administration. In the previous model, recall is below 70% and this needs to be increased.

For this reason, the probability threshold used to assign a new observation to label 1, $P(Y = 1|X = x)$, is optimized. Also in this case, a 10-fold cross-validation is implemented. In order to find the best compromise between precision and recall, F1 score is considered as objective function. In Figure 3 is reported the behaviour of F1 score over the interval $\{0.05, 0.95\}$ of probability threshold. The highest value of F1 score is reached when probability threshold is equal to 0.40. It means that a new observation will be assigned to label one when $P(Y = 1|X = x) > 0.40$. In the 10-folds cross validation setting, this algorithmic configuration got accuracy = 0.757, recall = 0.727, precision = 0.680 and F1 score = 0.702 (in average).

RF and its best configuration of parameters ($ntree = 500$ and $mtry = 12$) is than re-trained on the whole training set for the final test. The final performance on the test set, considering the tuned probability threshold ($P(Y = 1|X = x) > 0.40$), are: accuracy = 0.768, recall = 0.747, precision = 0.667 and F1 score = 0.705. Figure 4 shows the Receiving Operating Characteristic (ROC) curve. Considering the final performance, RF seems to be a stable model for predicting the target variable.

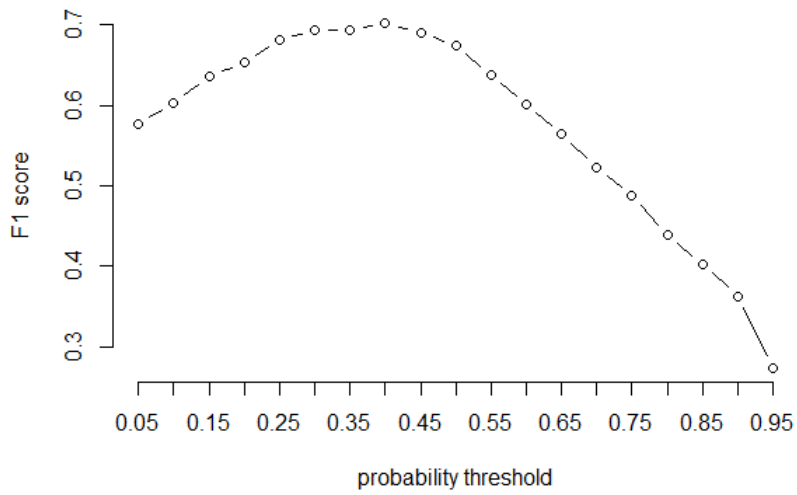


**Figure 3:** F1 score behaviour over probability threshold.

### 5.1 Variable importance

Variable importance can easily be measured when decision trees and random forests are adopted. Decision trees iteratively splits a dataset with the aim of decreasing as much as possible the heterogeneity of the target variable within each subset. Each split is based on a variable and a threshold. Therefore, the importance of any variable can me measured by its ability to decrease the weighted heterogeneity in each tree, or the *impurity* of the tree. The importance of a variable within a random forest model is obtained averaging such decrease in impurity over trees (Hastie et al., 2001; Breiman, 2001).

We partition the variables in six subsets:

- positive profitability: variables entering with a positive sign in the profit-and-loss statement of the GUO
- negative profitability: variables entering with a negative sign in the profit-and-loss statement of the GUO
- profitability: variables describing the profitability of the GUO
- size: variables describing the size of the GUO
- financial: variables describing the financial structure of the GUO

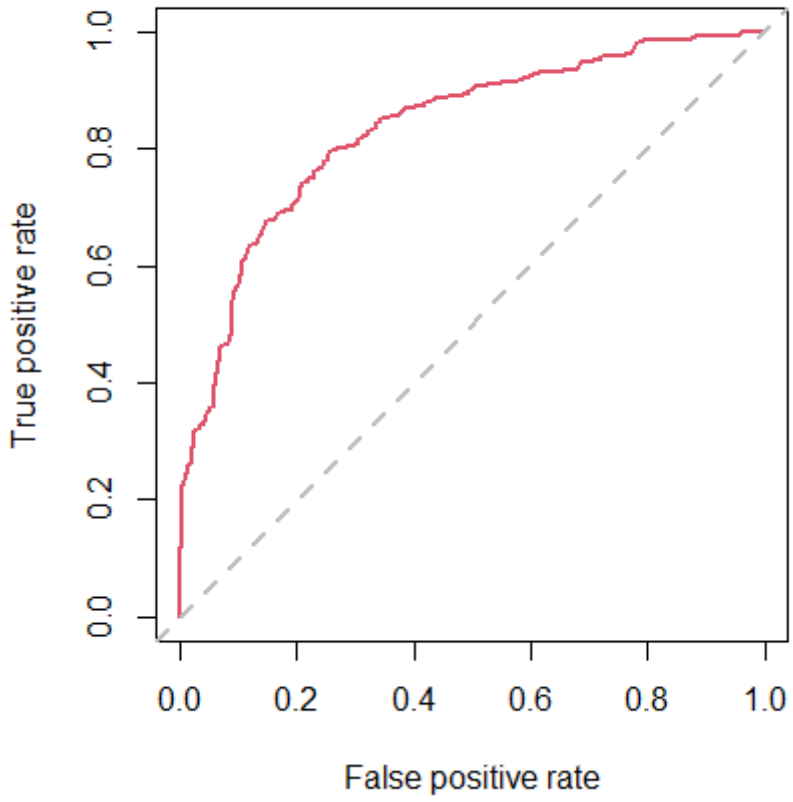


**Figure 4:** ROC curve.

- categorical: variables describing some time-invariant features of the GUO (country, sector, year of incorporation, etc).

Figures 5 and 6 reports the 10 most important variables for the first 3 (profitability-related) subsets and for the latter 3 subsets.

It appears that variables describing positive profitability, size and financial structure of the GUO are the most important ones. Other variables are of a lesser importance and, in particular, time-invariant characteristics (such as the country of incorporation or the sector) are of a negligible importance as compared to others. There are some interesting cross-correlations between variables that we classify in different subsets. In particular, note that the four most important variables are the amount of assets, the amount of other current liabilities and of shareholder funds and the level of sales. Overall, this analysis seems to suggest that the decision to locate (or to maintain) a subsidiary in a tax haven is more associated with the size and the profitability of the GUO rather than on the specific sector or country of the GUO.

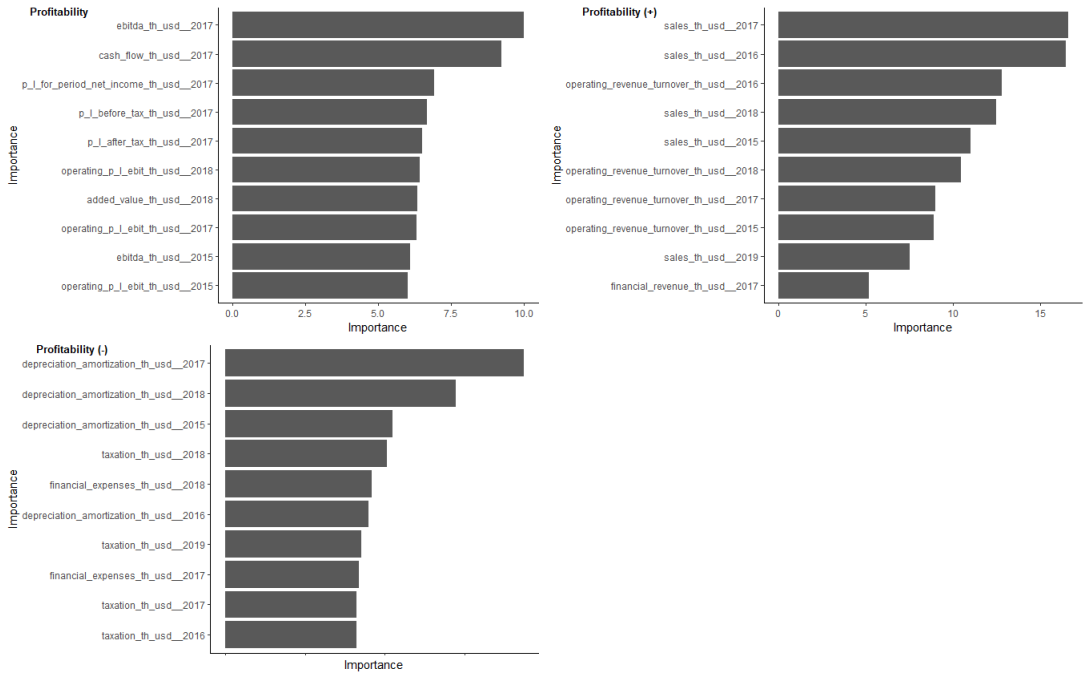


**Figure 5:** Variables ordered for importance and divided in three groups related to the type of information: profitability, profitability positive and negative.

### 5.2 Analysis of time impact

Previous analysis suggests that many variables are selected as predictor repeatedly in different years, from 2015 to 2019. Random forest models do not directly handle time so to analyze the time impact we decompose the initial dataset in a series of subset with decreasing years. More precisely, five subsets are considered. The first one includes only variables related to 2019, the second subset describes the period 2019 - 2018, the third is related to 2019 - 2018 - 2017 and the fourth 2019 - 2018 - 2017 - 2016. The last subset is the entire dataset. All subsets contain the categorical variables since they are time invariant. On each subset, a 10-folds cross validation technique is applied and a random forest with default parameters is estimated. Results are presented in Table 6. The predicted performance of the random forest model increases while adding years until 2017 than performance starts to deteriorate. The pre-processing phase may impact the results since the missing values imputation approach implicitly adds redundant information, however variables related to the period 2019-2017 provide sufficient information for the prediction. This result confirm that the information content of the data decreases in time (Bajgar et al., 2020).

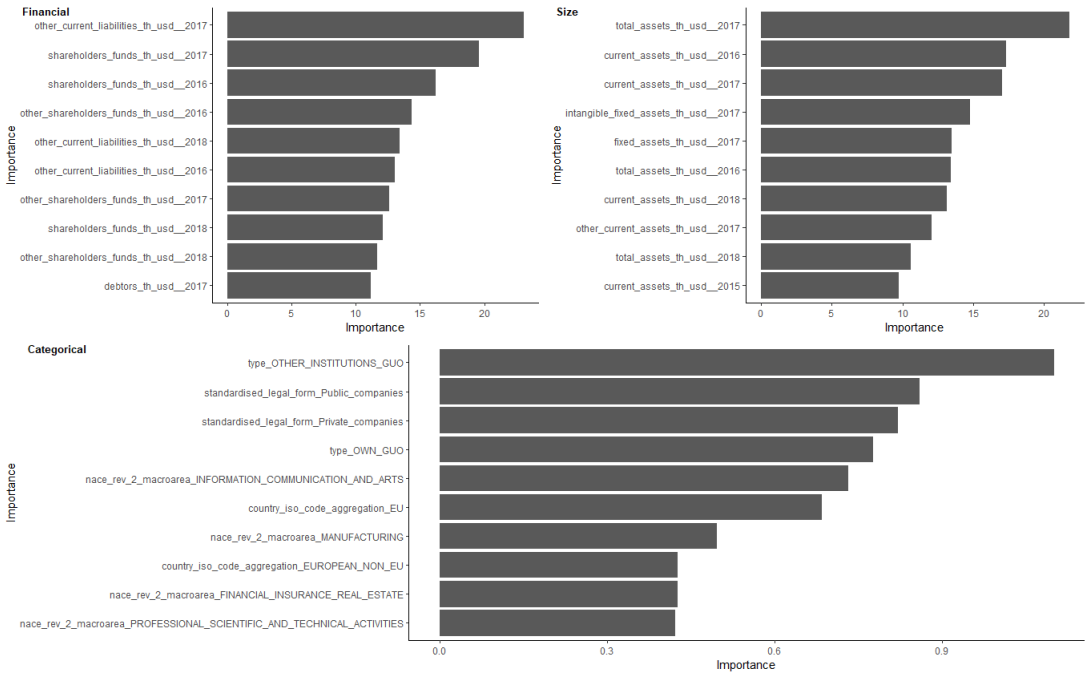


**Figure 6:** Variable importance for financial, size and categorical variables.

**Table 6:** Time impact on RF performance.

| Metrics | 2019 | 2019 - 2018 | 2019 .. 2017 | 2019 .. 2016 | 2019 .. 2015 |
|---|---|---|---|---|---|
| Accuracy | 0.767 | 0.771 | 0.777 | 0.771 | 0.774 |
| Recall (sensitivity) | 0.583 | 0.601 | 0.611 | 0.594 | 0.604 |
| Precision | 0.772 | 0.769 | 0.779 | 0.774 | 0.779 |
| F1 score | 0.663 | 0.673 | 0.683 | 0.671 | 0.678 |

## 6 Predicting the intensity in the use of tax havens

As described in Section 4, for each company is observed if the fiscal residence in 2021 is located in a tax haven or in a non-tax haven country. In the previous analysis (Section 5), we constructed a target variable following a binary classification setting. More precisely, a Global Ultimate Owner (GUO) is classified with label 1 if at least one company of the group is in one of the ATP countries, 0 otherwise.

In this section, we propose an approach for predicting the proportion of subsidiaries that a GUO will located in a tax haven country. For this reason, a new quantitative target variable is introduced with range between 0 and 100. If the value is equal to 0 no subsidiary is located in a tax haven country. If the value is equal to 100, all subsidiaries are located in a tax haven

country.

In order to solve this problem, we operate in a sequential manner. Ideally, the problem can be solved in two phases. As first phase, given a GUO as a new observation we can predict if it has at least one subsidiary in one of the ATPC. In the second phase, if the GUO has at least one subsidiary in one of the ATPC then we predict the proportion of subsidiaries in the ATPC. Figure 7 shows the phases of the sequential approach. Model 1 is a classification model and we exploit the Random Forest (RF) with $ntree = 500$, $mtry = 12$ and $P(Y = 1|X = x) > 0.40$ presented in Section 5. Similarly to Model 1, Model 2 is a 10-folds cross-validation optimized regression RF with $ntree = 1500$ and $mtry = 13$. The tuning optimization procedure is the same applied on Section 5. In order to train Model 2, same data presented in Section 5 is used. Differently from Model 1, the training set contains the new quantitative target variable which is computed as the ratio between the number of subsidiaries in the GUO in a tax haven country and the total number of subsidiaries in the GUO. The training set is then used for training Model 2. The whole approach is called *Sequential-Random Forest* (Sequential-RF).

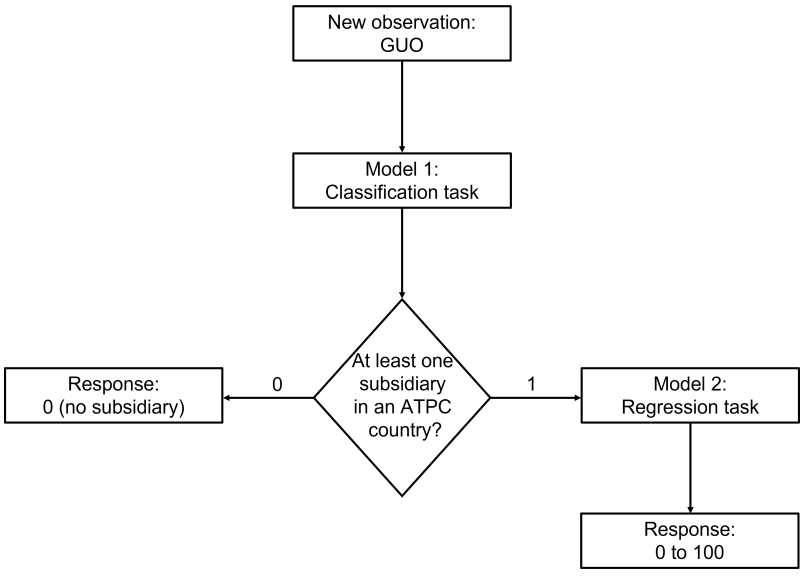


**Figure 7:** Scheme of the sequential approach.

As performance metric of the whole sequential approach, we consider the Root Mean Square Error (RMSE) defined as in Eq. 1.

$$RMSE = \sqrt{\frac{1}{N_t}\sum_{i=1}^{N_t}(y_i - \hat{y}_i)^2} \tag{1}$$

In Eq. 1, $N_t$ is the number of observations in the test set, $y_i$ is the actual proportion for

the $i-th$ observation and $\hat{y}_i$ is the prediction for the $i-th$ observation. The Sequential-RF reached a RMSE equal to 35.86 and was able to correctly assign 0 to 78% of GUOs. This result is compared with the prediction obtained by Model 2 on all the test set without applying, as first step, Model 1. In this case, we got a slightly lower RMSE (34.32) but no GUOs have been predicted with 0 subsidiaries in non-tax haven countries. From a practical point of view, this suggest that a Sequential-RF is a better model for solving the general problem of identifying GUOs with subsidiaries in a non-tax haven or tax haven country and the corresponding proportion when necessary.

Also for Model 2, the variable importance is analyzed considering the same six subsets presented in Section 5.1. Figures 8 and 9 reports the 10 most important variables, respectively, for profitability, positive and negative profitability subsets and for size, financial and categorical subsets. The variables are very similar to those displayed in Figures 5 and 6.

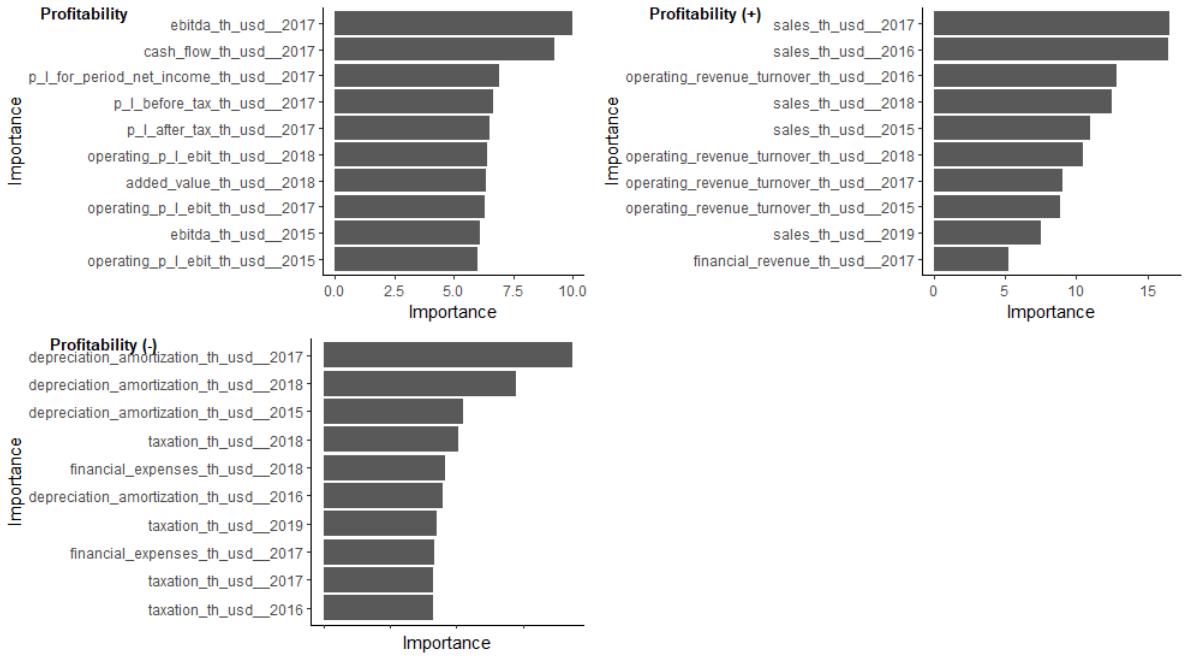


**Figure 8:** Variables ordered for importance and divided in three groups related to the type of information: profitability, profitability positive and negative.

## 7 Policy implications

Consider a tax authority which is evaluating the risk that a 'national' MNE locates (or maintain) a subsidiary in a tax haven. Suppose that the tax authority has in mind an intervention (a 'treatment') that can be effective in preventing this risk, but that this treatment is costly (both

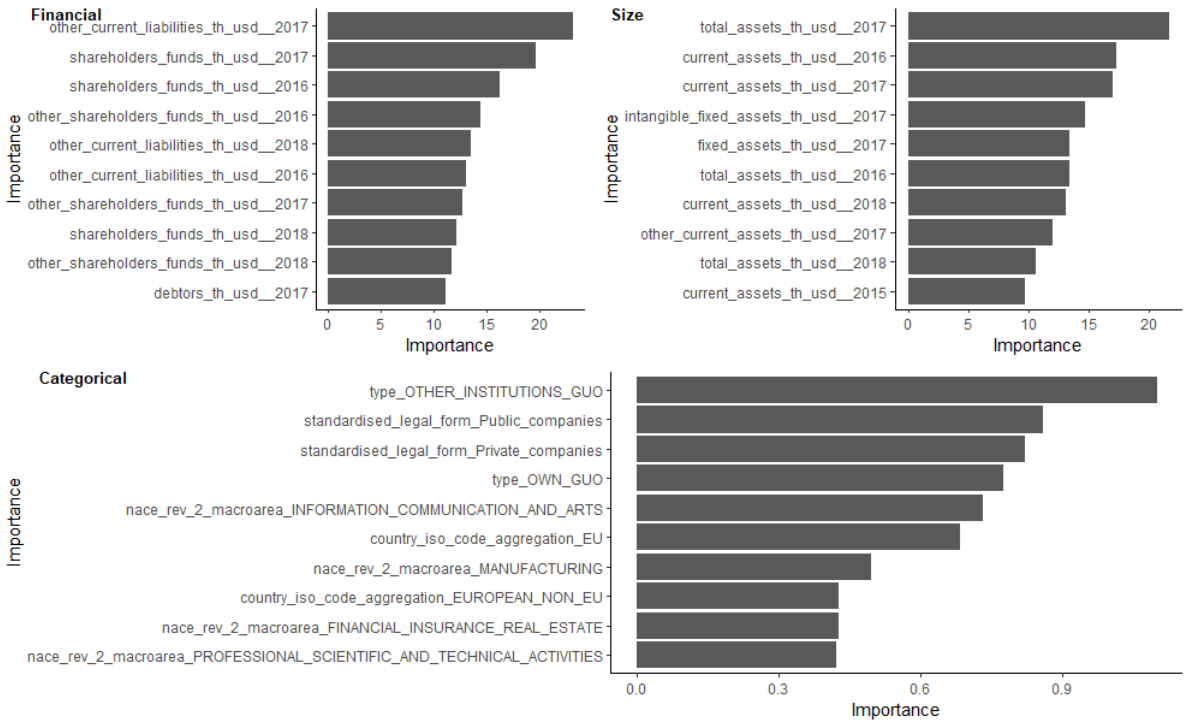


**Figure 9:** Variable importance for financial, size and categorical variables.

for the tax authority and for the company). This treatment can consist in an audit having the purpose to collect more information on the tax planning activities of the company, i.e. an ATP-audit. Each ATP-audit is costly for the tax authority, that has to devote to it specific human resources, and also for the company, that has to provide the type of information requested in the audit, thus diverting resources from the production activity (ultimaltely, compliance costs are opportunity costs). Previous results show that the tax authority can select the companies to be ATP-audited using exclusively publicly available financial information and that, by doing so, it can first, prevent some companies from reducing their tax payments, and, second, save public and private money reducing useless audits.

More precisely, suppose that the tax authority defines a set of $N$ GUOs that could be subject to an ATP-audit at time $t$ and that the agency has a budget constraint such that only $n < N$ of these GUOs should be audited. Initially, the tax authority only knows that $p$ is the probability that any GUO will locate or maintain a subsidiary in a tax haven at time $t + 2$ so that there are $pN$ true positives and $(1 - p)N$ true negatives. This implies that, if the tax authority audits randomly $n$ taxpayers, $np$ would be true positives, whilst $n(1 - p)$ would be false positives, so that the false positive rate is equal to $(n/N) \leq p$ and the false negative rate equals $1 - (n/N) \geq (1 - p)$.

Now, by using accounting information available at time $t$, the tax authority can assign a probability $q$ of the use of tax havens by time $t+2$ to *each* company and consequently can select for ATP-audits only companies that have a $q$ higher than this given minimum probability. If the share of true positives among the $n$ audited using the threshold $q$ is $s$, then with this policy there will be $sn$ true positives and $(1-s)n$ false positives. Consequently, the false positive rate would be equal to $(n/N)(1-s)/(1-p)$ and the false negative rate would equal $1-(sn/pN)$. Clearly, if $s > p$, i.e if the share of true positives among the audited is higher than the portion of true positives in the population, both errors are decreased with respect to random auditing.

In our first application of the RF algorithm, the share of MNEs having a company located in a tax haven, $p$, is slightly less than 40%, while $s$, which is the precision rate, varies, across different models, around 70%. If $n/N = p = 40\%$ the false positive rate (Type I error) and the false negative rate (Type II error) would both decrease by 50%, from 40% to approximately 20% and from 60% to approximately 30%, respectively.

Note also that the similarity between most important predictors of the risk of using tax havens, i.e. our first application, and of the intensity of this use, our second application, suggests that the prediction problem is quite simple for the tax authority. A knowledge of a limited number of variables allows the tax authority to make a fairly reliable prediction of the individual risk of tax planning through strategic location decisions.

### 7.1 Using the predictive approach in the implementation of the CoCo scheme: evidence from Italy.

A more specific use of the approach suggested here can be hypothesized in the context of CoCo schemes, that we already described, in their general features, in the Introduction. Here we look more closely at the CoCo scheme implemented in Italy, which share many features with others adopted internationally.

The Italian CoCo (known as *Regime di adempimento collaborativo*) scheme was established with the legislative decree of 5 August 2015, n. 128. Companies equipped with a system (known as tax risk control framework, TCF) for detecting, measuring, managing and controlling tax risk, understood as the risk of operating in violation of tax regulations or in contrast with the principles or purposes of the tax system, can, in principle, participate. However, within

companies equipped with a TCF, eligibility is restricted to resident and non-resident companies (with a permanent establishment in Italy) that achieve a volume of revenues of no less than 5 billion euros. According to the Italian tax authority's website, the purpose of the scheme is to establish a relationship of trust between the administration and the taxpayer that aims to increase the level of certainty on relevant tax issues. This objective is pursued through constant and preventive dialogue with the taxpayer on facts, including the anticipation of control, aimed at a common assessment of situations likely to generate tax risks. Eligible companies that want to adhere to scheme issue a request, which is evaluated by the tax authority. If the evaluation is positive, the companies and the tax authority enter into the agreement. For companies, the benefits are the following:

1. shortened preliminary ruling procedure [6];
2. application of penalties reduced by half, for the risks communicated in a timely and exhaustive manner, where the tax authority does not agree with the position of the company;
3. exemption from presenting guarantees for the reimbursement of direct and indirect tax credits [7].

The Italian law allows single companies to enter the scheme even if they are not GUOs. However, in practice the tax authority aims at involving in the scheme the GUO and, possibly, all or at least some of its subsidiaries. As of end of December 2021, 62 single companies had been admitted in the scheme, corresponding to 29 non-financial groups (29 GUOs and 22 subsidiaries) and to 7 financial and banking groups (7 GUOs and 4 subsidiaries).

In our dataset, we can retrieve 9 GUOs of non-financial groups that have been admitted in the scheme and also 5 GUOs that do satisfy eligibility requirements but have not entered the scheme as of December 2021. Note that all of these are fiscally resident in Italy. As expected, and given the importance of the size of the company as predictor of the probability to locate a subsidiary in a tax haven, the average probability of locating or maintaining a subsidiary in a

[6] A ruling is a written interpretation of tax laws and of their implementation; it is normally issued by the tax authority upon request by the taxpayer

[7] In Italy, requests of reimbursement of tax credits exceeding given thresholds are subject to further scrutiny unless the taxpayer presents guarantees, such as a compliance visa issued by a certified tax pratictioner

tax haven in 2021, as predicted by our best model using accounting from period 2017-2019, is very high among the 14 GUOs, and it amounts at 83.7%. Only 3 out of the 14 GUOs considered here have a predicted probability lower than 40%, i.e. the average observed in the sample and 1 of these 3 has not entered in the scheme as of December 2021.

These results are thus indicating that, on the one hand, the vast majority (7 out of 9) of the GUOs that have already entered in the CoCo scheme or that may enter in it in the future (4 out of 5) had or will have a high risk to locate or to maintain a subsidiary in a tax haven. Our analysis suggests that the tax authority should adapt the scheme to this risk. To illustrate how this can happen, let us imagine that the 5 GUOs that are eligible do request to enter the scheme in the nearest future. For the 4 that present a high risk, some adjustments could be designed. For example, before accepting to conclude the agreement with these companies, the tax authority could specifically scrutinize the location decisions, their motivations and scope and ask the company to precommit to disclose specific information on these decisions in the future. Also, the benefit of the shortened period for any ruling associated to the use of schemes that may involve strategic placement decisions could be disallowed to these companies and a same reasoning could be applied to the reduction in sanctions for the use of elusive schemes based on location decisions.

## 8 Concluding remarks

Achieving effective taxation of multinational enterprises has been a high priority on the international tax policy agenda for a long time, as governments attempt to find an effective solution to the tax challenges posed by the increasing digitalisation of the economy and the resulting changes in value creation processes.

In particular, in Europe the proposal of a minimum effective tax rate has been supported by all large countries (US, Germany, France, Italy) whose multinationals invest, through the creation of subsidiaries or special entities, and pay taxes (if any) in other countries where favourable tax rules are available. Many of these 'other' countries are, as a matter of fact, European countries (Ireland, Luxembourg, Netherlands). Taking the viewpoint of the latter countries, that we can define 'tax havens' in a broad sense, any international project that reduces the scope of tax competition is a net loss of capitals and/or tax revenues. Until today

these countries have been steadily and successfully opposing all these efforts.

There are many reasons for this. The full integration of capital markets, the absence of mandatory rules on tax competition at the international level and the need for 'unanimity vote' on tax matters at the EU level are all factors that have undermined the attempts to reduce the impact of tax havens. In this context, preventing any attempt to exploit international tax competition, including through strategic decisions, was a non-credible threat.

However, a crucial signal in favour of multilateralism, cross-border cooperation and fair taxation was sent out on 8 October 2021: The OECD/G20 Inclusive Framework on Base Erosion and Profit Shifting agreed on a two-pillar solution to address the tax challenges arising from the digitalisation of the economy. According to Pillar One, taxing rights over 25% of the residual profit of the largest and most profitable MNEs would be re-allocated to the jurisdictions where the customers and users of those MNEs are located. According to Pillar Two, GloBE (Global Anti-Base Erosion) rules provide a global minimum tax of 15% on all multinationals with annual revenue over 750 million euros.

The two-pillar strategy, if implemented, should in theory reduce the incentives to relocate activities and subsidiary in tax havens to reduce taxes. Clearly, the scope of the agreement and the number of countries that will eventually adhere is a crucial feature in the implementation of the strategy. Given the large number of countries that have subscribed to the Inclusive Framework (141 as of December 2021), it seems fair to conjecture that, in the nearest future, multinationals will face higher transaction costs to locate or to maintain subsidiaries in tax havens.

The type of data mining and predictive administrative action that we explore in this paper could be a more credible tool in this context, A MNE that is considering to use tax havens will face, on the one hand, lower expected gains, because of the application of a minimum effective rate, and, on the other hand, a higher probability to be closely monitored by the tax authority which uses the approach we explore in this paper. In this perspective, the increase in the efficiency of national tax administration policies should be seen as a*complement* rather than *substitute* for enhanced international corporate tax coordination efforts. This complementarity is particularly relevant for Europe. At the end of December 2021, the European Commission has issued a proposal for the implementation of minimum effective taxation (Pillar Two) within

Europe. This implementation, in principle, should reduce the gain that a European MNE obtains from relocating part of its activities in European tax havens. In turn, this should make national policies, including administrative ones of the type envisaged in this paper, more effective.

## References


Bajgar, M., G. Berlingieri, S. Calligaris, C. Criscuolo, and J. Timmis, 2020: Coverage and representativeness of orbis data.

Breiman, L., 2001: Random forests. *Machine Learning*, **45**, 5–32.

De Simone, L., L. F. Mills, and B. Stomberg, 2019: Using irs data to identify income shifting to foreign affiliates. *Review of Accounting Studies*, **24**, 694–730.

European Council, 2021 [Online]: Council conclusions on the revised eu list of non-cooperative jurisdictions for tax purposes. URL https://eur-lex.europa.eu/legal-content/en/TXT/PDF/?uri=uriserv:OJ.C_.2021.066.01.0040.01.ENG.

Hastie, T., R. Tibshirani, and J. Friedman, 2001: *The element of statistical learning*. Springer Series in Statistics.

Meldgaard, H., J. Bundgaard, K. Dyppel Weber, and A. Floristean, 2015: Study on structures of aggressive tax planning and indicators. Final Report Taxation Papers, Working Paper n.61, TAXUD, European Commission.

Newberry, K. J., and D. S. Dhaliwal, 2001: Cross-jurisdictional income shifting by u.s. multinationals: Evidence from international bond offerings. *Journal of Accounting Research*, **39 (3)**, 643–662.

OECD, 2013: Addressing base erosion and profit shifting. Report https://doi.org/10.1787/9789264192744-en, OECD publishing.

Tørsløv, T. R., L. S. Wier, and G. Zucman, 2018: The missing profits of nations. Working paper 24701, National Bureau of Economic Research.

Zucman, G., 2014: Taxing across borders: Tracking personal wealth and corporate profits. *Journal of Economic Perspectives*, **28 (4)**, 121–148.