ARTICLE IN PRESS

Economic Systems xxx (xxxx) xxx

Contents lists available at ScienceDirect

Economic Systems

journal homepage: www.elsevier.com/locate/ecosys

Using accounting information to predict aggressive tax location decisions by European groups

Matteo Borrotti[*], Michele Rabasco, Alessandro Santoro

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

ARTICLE INFO

JEL classification:
H25
H26
C53

Keywords:
Aggressive tax planning
Machine learning
Tax havens
Tax compliance by multinationals

ABSTRACT

Although locating a company in a tax haven is not illegal per se, it is likely to be part of a scheme purported to erode the tax base or to shift profits to less-taxed jurisdictions. For this reason, this type of location decision is usually targeted by anti-avoidance laws, that can take the form either of specific rules or general standards that, ex-post, sanction or limit the location decision. However, rules entail higher drafting costs and are easy to circumvent whereas standards entail more uncertainty costs. The goal of this paper is to illustrate that the risk of aggressive location decisions can be predicted ex-ante using publicly available data and that this prediction can be used by tax authorities. In the paper, we do two things. First, we use publicly available accounting data for the period 2015–2019 on 4031 group ultimate owners (GUO) of active listed companies resident in one of the 27 European Union countries to predict the probability that these companies would have at least a subsidiary in a tax haven, by spring 2021, as well as the intensity in the use of tax havens. Second, we discuss how this prediction can be used by tax authorities in the context of a new administrative preventive approach that complements the traditional legal approach. This approach can increase welfare by reducing uncertainty, thus increasing investments and economic growth.

1. Introduction

According to the definition provided by the European Commission (European Commission, 2012), aggressive tax planning (ATP) consists of “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 that uses 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 the extent to which a member state of the European Union (EU) can be seen as favoring 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), in which profits shifted from non-tax havens to tax havens and the associated

[*] Correspondence to: DEMS, University of Milano-Bicocca, Milan, Italy.
E-mail address: Matteo.borrotti@unimib.it (M. Borrotti).

https://doi.org/10.1016/j.ecosys.2023.101090
Received 29 January 2022; Received in revised form 10 January 2023; Accepted 11 January 2023

Please cite this article as: M. Borrotti, M. Rabasco and A. Santoro, Using accounting information to predict aggressive tax location decisions by European groups, Economic Systems, https://doi.org/10.1016/j.ecosys.2023.101090

loss of potential corporate tax revenue are precisely quantified. Both approaches suggest that the decision by a multinational corporation (MNC) to locate a subsidiary in a country that offers a favorable tax regime (for the sake of simplicity: tax haven) is one of the main symptoms of ATP, albeit not the only one.

Although locating a company in a tax haven is not illegal per se, it is likely to be part of a scheme purported to erode the tax base or to shift profits to less-taxed jurisdictions For this reason, the location of a company in a tax haven is already targeted in either the design or the application of some existing anti-avoidance laws.

For example, the location of a subsidiary in a tax haven is a necessary but not sufficient condition for the application of rules on controlled foreign companies (CFC rules), which are designed to allow the country of a parent company to tax profits that the company shifts to the subsidiary. Moreover, many laws include lists of countries that are considered aggressive (black lists) or not completely cooperative (grey lists). Also, specific laws are designed to deny the benefits arising from the hybrid mismatch agreements, i.e. agreements that exploit differences in the tax treatment of an entity or instrument under the laws of two or more tax jurisdictions, that make use of special purpose vehicles (SPV), which are usually located in tax havens.

The main motivation of our paper is the fact that, although the introduction of rules to limit the impact of SPVs (OECD, 2012) and CFCs (Clifford, 2019) has proven successful to some extent, they are costly to implement.

To see why, consider that, in general, laws can take the form of either rules or standards. Rules are laws that are costly for national legislators to create whereas standards are laws that are costly for taxpayers to interpret. In sum, rules entail higher drafting costs, but standards entail more uncertainty costs (Kaplow, 1992). Although uncertainty costs could be deemed higher thus suggesting the adoption of rules, rather than standards, the application of rules in the taxation field is prone to the “discontinuity” problem: small change in transactions can lead to large changes in tax liabilities (Weisbach, 1999). Therefore, taxpayers have a strong incentive for creating ways in which to circumvent tax rules by finding or creating loopholes.

In particular, rules can be avoided by appealing to the principle of freedom of establishment. This principle is a fundamental one for the EU, and it is often at odds with anti-avoidance rules, thus pushing countries to revert to standards or to the application of general principles. For example, in some national laws, CFC rules cannot be applied if the taxpayer resident in the EU can show that it carries on “substantial economic activity” abroad. Similarly, the benefits from hybrid mismatches that involve SPVs cannot be denied if the “beneficial owner” is resident in the EU. In these cases, the application of rules ultimately depends on the application of very general and uncertain principles.

To sum up, under a legal approach, the efficiency of rules is undermined by drafting costs and loopholes; at the same time, the efficiency of standards is diminished by uncertainty and compliance costs. Moreover, it is not uncommon for rules, in practice, to be converted into standards in order to avoid loopholes, thus increasing uncertainty costs.

The goal of this paper is to illustrate that the risk of aggressive location decisions can be predicted using publicly available data and that this prediction can be used by tax authorities with a preventive administrative approach that can complement that based on rules, that is, the traditional legal approach. In the paper, we do two things.

First, we use publicly available accounting data from ORBIS for the period 2015–2019 on 4031 group ultimate owners (GUOs) of active listed companies resident in one of the 27 EU countries to predict the probability that these companies would have at least a subsidiary in a tax haven in spring 2021. We also predict the intensity in the use of tax havens, that is, the share of the companies that would be located in a tax haven by spring 2021 among all the subsidiaries.

Second, we discuss how this prediction can be used by tax authorities in the context of an administrative preventive approach that complements the traditional legal approach. This approach can increase welfare by reducing uncertainty, thus increasing investment and economic growth.

Our main contributions as follows. First, we develop and test a machine learning model (i.e., random forest) to predict whether an MNC will have a subsidiary located in a tax haven by spring 2021. The proposed model is based on variables for three years (2017–2019) and achieves an accuracy rate, the ratio between correctly classified cases and total cases, of around 75%. This was possible by using only publicly available accounting data and exploring (nonlinear) relationships between explanatory variables. Given our analysis, we identify financial, profitability, and size variables as the most important predictors (Grubert, 2019; Gallemore et al., 2014). However, most important, we show that prediction performance significantly increases when at least two of the most important variables are used together with the random forest model, thus allowing for nonlinear relationships.

As our second contribution, we propose a *sequential random forest* (Sequential-RF) to predict the intensity of the use of tax havens and we show that, although the most important variables for this prediction are similar to those identified by the classification model, they are not exactly the same.

Our third and last contribution is to show that the predictions can be used in practice with an administrative preventive approach that combines some features of existing procedures, namely, cooperative compliance and tax rulings, with some additional powers granted to the tax authority.

The rest of the paper is organized as follows. Section 2 presents the related literature. Section 3 describes the data creation procedure and the final dataset. Section 4 is devoted to the definition of the explanatory variables used in this work. Section 5 introduces the machine learning model for predicting the risk of using tax havens and summarizes main results, while in Section 6 we propose an approach for predicting the proportion of subsidiaries that a GUO will locate in a tax haven country, the sequential random forest. Section 7 describes how predictions can be used by tax authorities. Some conclusions and suggestions for future work are summarized in Section 8.

2. Related literature

The importance of the decision to locate a company in countries that offer favorable tax treatment emerges clearly in the paper by Gabriel Zucman and his associates who are involved the *Missing Profits* project.

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. The creation of subsidiaries in the three tax havens (Bermuda, Ireland, and the Netherlands) is a necessary condition for the scheme to be implemented.[1]

Another stream of literature that is of interest here is that of looking at the impact of rules adopted by national legislators to respond to aggressive location decisions by MNCs.

CFC rules are studied by Clifford (2019), who exploits what she calls the "tax thresholds." Interest and royalty income by foreign subsidiaries are included in the parent company's tax base only if the subsidiary is located in a jurisdiction that applies a tax rate below the threshold. Clifford (2019) finds that over the period 2003–2013 MNCs redirect profits into subsidiaries just above the threshold and change incorporation patterns to locate fewer subsidiaries below the threshold and more above it.

The application of the CFC rules within the EU was redefined by the Anti-Tax Avoidance Directive (ATAD) adopted there in 2016. According to ATAD, member states can choose to apply CFC rules under two options. If the first option is adopted, certain predefined categories of undistributed passive income (dividends, interests, royalties, and other income from financial activities) of the CFC are attributed to the parent company. ATAD states that, in such a case, CFC income should not include the share from "substantive economic activities" conducted in a foreign EU or European Economic Area (EEA) country. If the second option is adopted, income of the CFC is attributed to the parent company if it comes from nongenuine arrangements put in place for the essential purpose of obtaining a tax advantage.

Thus, the application of CFC rules within the EU ultimately relies on proof of the (absence of) "economic substance" or the presence of "an essential purpose of a tax advantage". In this way, an apparently sharp and precise *rule* is actually converted into a more vague and uncertain *standard or principle*.

The impact of special rules against the hybrid mismatch arrangements were examined by OECD (2012).

Hybrid mismatch arrangements exploit differences in the tax treatment of instruments, entities, or transfers between two or more countries, which often leads to "double nontaxation" that may not be intended by either country or to a tax deferral that, if maintained over several years, is economically similar to double nontaxation.[2]

A number of countries have introduced rules that specifically deny benefits that arise from certain hybrid mismatch arrangements. The thrust of all these rules is to link the domestic tax treatment of an entity, instrument, or transfer involving a foreign country to the tax treatment in that foreign country.

Although the experience of countries that have introduced rules expressly denying the benefits from hybrid mismatch arrangements has been positive overall, tax authorities noticed that arrangements can become more elaborate after the introduction of specific rules denying benefits in the case of hybrid mismatch arrangements. For example, taxpayers tried to circumvent the rules by interposing companies in an EU/EEA or treaty state, so that countries had to amend the rules by specifying that the deduction was legitimate if the company in an EU/EEA or treaty state was the beneficial owner. However, the implementation of this principle rests on the ability to show who the beneficial owner is (OECD, 2012).

Finally, another stream of literature related to our paper is on the determinants of ATP by MNCs.

Newberry and Dhaliwal (2001) look at the determinants of the location decision. They study the decision by a US MNC 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 several tax and nontax variables.

More recently, the accounting and economics literature has focused on the determinants of income shifting by US MNCs. Papers in this literature examine the impact of specific financial incentives or constraints and income shifting between specific countries or between affiliates of a given profitability level, but they all suffer from having limited information about the dependent variable: actual income shifting can be measured only by using Internal Revenue Service data on intercompany payments (De Simone et al., 2019).

Our approach and that followed in the previous literature on ATP has three differences.

First, we focus on European groups (i.e., groups consisting of companies listed in a stock market and resident in the EU-27), and we look mainly at tax competition within Europe. Both aspects of our research design are a novelty in the literature, which mainly focuses on US MNCs and 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

[1] Clearly, the exact amount of different transactions should be known to estimate the loss of tax revenue suffered by the US. As Tørsløv et al. (2018) show, the information about the magnitude of capital flows available in ORBIS is highly unreliable.

[2] 2A classical example is one in which parent A establishes an SPV in a tax haven to control company B. Then SPV borrows from another entity and uses the loan to inject capital into B. If the SPV is treated as a part of B's group by B's country and as a transparent by A's country the interest paid by the SPV will be deducted from B's consolidated tax base as well as from A's tax base. This is known as a "double deduction" scheme. In this case, a special rule may disallow the deduction of the interest expense from B's tax base if the same deduction is allowed from A's tax base, or vice versa.

associated with a specific type of tax planning, as in the paper 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).

Third, we use a machine learning (ML) approach and, rather than preselecting explanatory variables that should be associated with ATP according to a given theory or approach and then testing the significance of their correlation, we let the data reveal which predictors are the best in the decision to locate or maintain a company in a tax haven.

The application of ML methods, especially for prediction, detection, and targeting purposes, is beneficial in several fields. In Barboza et al. (2017), ML models are tested to predict bankruptcy one year before the event. They find that, in general, ML models have better, more accurate predictive performance than traditional statistical models. Another field in which ML is extensively employed is the detection of financial frauds, as shown by Sadgali et al. (2019). Finally, Badal-Valero et al. (2018) discuss how ML algorithms can be applied to find patterns by money launderers, whereas Andini et al. (2022) (2022) use ML to propose a targeting rule, with the goal of increasing the effectiveness of a public guarantee program.

3. Dataset

We use data from the ORBIS database, which provides firm-level data on over 400 million companies and entities worldwide, along with detailed information on company structure.[3] The starting point of the analysis is the identification of active listed companies that are resident in one of the 27 EU countries.[4] Among them, we consider for analysis only those identified by ORBIS as corporations. These companies are labeled as LISTED, because they are listed on a stock market, not necessarily one in the EU.

The second step is the identification of the group to which every LISTED company belongs. In turn, this requires the identification of the global ultimate owner and of all companies that are part of the same UO (subsidiaries, in short). Both steps are completed using data available in spring 2021.

ORBIS identifies UOs by analyzing 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 company to its UO must be 50.01%.[5] A company is considered the UO if it has no identified shareholders or if the proportion of control by its shareholders is not known.

We refer to a GUO to distinguish it from a domestic UO (DUO), that is, the highest company in the path between a particular company and its GUO located in the same country as that company. Beginning with the LISTED companies, we identify 4031 GUOs subdivided as follows: 2654 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 and obtain a set of 127,827 subsidiaries.[6] The composition of the final dataset of 131,858 companies, updated to at least 2015, is shown in Table 1.

Approximately two-thirds of the dataset are companies in groups whose GUO is a LISTED company. Groups with a noncompany GUO are larger but less populated than groups with nonlisted GUO. Table 2 gives more details on the size of the groups and also illustrates considerable variability in group size. 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 size = 1, the GUO itself) and groups of considerable size (up to 2114 companies).[7]

4. Definition of the variables and choice of prediction methods

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

For each company, we observe the country of fiscal residence in spring 2021, and consequently we determine whether it is a subsidiary located in a tax haven or nontax haven. More precisely, the list of countries with a high risk of aggressive tax planning (ATPC) is as follows. The starting point is the list of 12 countries identified by the European Council (EC) in its official journal as the EU list of noncooperative jurisdictions for tax purposes (European Council, 2021):: American Samoa, Anguilla, Barbados, Fiji, Guam, Palau, Panama, Samoa, Seychelles, Trinidad and Tobago, US Virgin Islands, and Vanuatu. To these, we add the Cayman Islands and Oman, which were on the list until the previous revision. To these 14 countries, we add seven countries for which at least 13 of the 33

[3] ORBIS is a commercial database provided to the Organisation for Economic Cooperation and Development (OECD) by the electronic publishing firm Bureau Van Dijk.

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

[5] The ORBIS guide specifies that the procedure for the UO 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."

[6] With reference to the subsets of OWN_GUOs and OTHER_COMPANY_GUOs, enables us to extract the entire 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 Appendix A

[7] However, it should be noted that the lower average number of groups with 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.

Table 1
Dataset composition: GUOs and subsidiaries.

Category	Freq.	%
OWN_GUO	2654	65.8
OTHER_COMPANY_GUO	562	14.0
OTHER_INSTITUTIONS_GUO	815	20.2
All GUOs	4031	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

Source: Authors' calculation based on the ORBIS database.

Table 2
Group size by GUOs.

category	min	q1	median	mean	q3	max
OWN_GUO	1	2	6	32.68	20	2114
OTHER_COMPANY_GUO	1	5	15	53.33	50	1542
OTHER_INSTITUTIONS_GUO	1	1	5	18.62	15	689

Source: Authors' calculation based on ORBIS database.

indicators of ATP used by the EC (Meldgaard et al., 2015) are present. To these 14 countries, we add seven countries for which at least 13 of the 33 indicators of ATP used by the EC Tørsløv et al. (2018) but not yet in our list, Switzerland and Ireland, for a total of 23 ATPC countries.

The GUOs in our dataset are distributed geographically as shown in Fig. 1, which illustrates the frequency of countries in the dataset (for visibility reasons, only countries with more than five GUOs are shown). The number of GUOs in each European country (other than Sweden) appears to be proportionate to the size of the country. The number of GUOs in each European country (other than Sweden) appears to be proportionate to the size of the country. The large number of GUOs in Sweden (642) is explained by the fact that, during the period considered, a boom occurred in corporate startups linked to new technologies. Countries with fewer than six GUOs are grouped in the NA category. The countries that we define as ATPC appear in red in the figure.

In addition, our dataset has 14,781 (11.5%) subsidiaries in an ATP country, which is an EU country in 90% of the cases and a non-EU European country otherwise. Only a residual share of the subsidiaries are in an ATP country outside the EU, because of the opacity of the tax systems in these countries, which prevents the identification of companies located there.

We classify GUOs based on the presence of at least one company in the group (either the GUO or a subsidiary) in one of the ATPC. This results in a dichotomic variable (at least one company in an ATPC or no company in ATPC) that we consider the target variable in our first prediction exercise. In our dataset, this breakdown results in 1574 (39.05%) GUOs with at least one company located in an ATPC and 2457 (60.95%) with no company located in an ATPC. Among the former, 1065 (67.7%) are groups whose GUO is not located in an ATPC country. (Table 3).

However, we cannot exclude that the subsidiary was part of the same group *before* 2021. Therefore, our dependent variable is defined as the probability of locating or maintaining a subsidiary in an ATPC (tax haven) in 2021. This clearly is an important limitation in our data that could be eliminated by using more refined information.

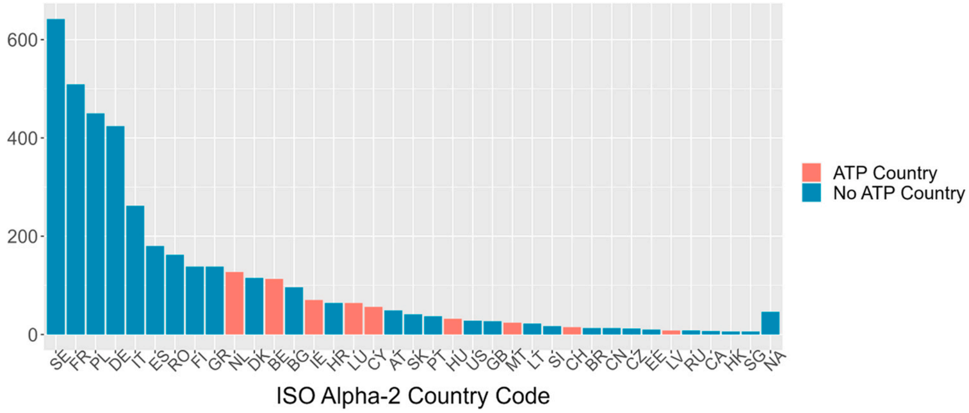

Fig. 1. Number of GUOs in each country.
Source: authors' own elaboration on ORBIS database.

Table 3
GUOs with at least one company located in an ATP country.

	Freq.	%
GUO is in an ATPC	509	32,3
GUO is not in an ATPC	1065	67,7

Source: authors' own elaboration on ORBIS database.

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

For every GUO, the final dataset comprises three main types of variables: (1) balance-sheet variables (intangible and tangible assets, stocks, capital, liabilities), (2) income statement variables (operating revenue, sales, cost of employees, taxes), and (3) firm characteristics, including sector of activity, country of residence, and legal form.

For the prediction methods, we considered a set of models suitable for the purpose of classification. More precisely, we compared logistic regression, decision tree (DT), and random forest (RF). Logistic regression is a well-known discriminant statistical model suitable for binary classification (Cox, 1958; Hastie et al., 2001). Among other ML approaches, we selected decision tree (Hastie et al., 2001) and random forest (Breiman, 2001) because they enable an intuitive measurement of variable importance and, therefore, are easier to interpret. More details about DT and RF are given as follows.

DT is a member of 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 (a tree composed of levels and leaves) of all the possible solutions to a decision based on certain conditions. In this algorithm, the training dataset is recursively split into two or more homogeneous sets, in order to obtain a locally optimal choice based on a given explanatory variable and a given threshold over it (top-down greedy approach), until a stopping rule is reached.[8] As a result, we obtain a tree in which the final nodes (leaves) provide a classification for the variable that we want to predict. Heterogeneity in subsamples is measured with the Gini index. When the classification results in only two classes, as in our case, the Gini index is calculated as $2h(1\text{-}h)$, where h is the proportion in the second class. Among the main advantages of DT are the ability to consider nonlinear effects while keeping the results interpretable and the robustness to heterogeneous data (Hastie et al., 2001).

RFs are nonparametric models in the category of ensemble methods. In ensemble methods, a series of learners (i.e., models) is used in order to enhance the final prediction performance. Starting with 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 RF, the weak learners are DT. The procedure can be summarized as follows: first, each DT is trained by employing only a random subset of m features for each tree and only a subsample of the initial sample, then trees are ensembled and averaged. Following the literature, m usually is selected as $m = \sqrt{p}$ where p, is the total number of explanatory (or input) variables. RF usually improves the accuracy of the prediction with respect to decision trees. Additionally a small value of m in building a RF helps when we consider a problem with a large number of correlated input variables (for more details, see Appendix C, and, for a more technical explanation, Hastie et al. (2001) and Breiman (2001)).

Before the data are analyzed, a preprocessing phase is needed. The dataset considered is composed of unique observations; however, the amount of null values is quite significant.[9]

Typically, two strategies are recommended for handling missing values: passive and active. Here, we use a combination of them. First, following the passive strategy, observations that have more than 50% of data missing are subtracted from the dataset. The same strategy is applied to remaining observations with variables whose percentage of missing data is greater than 50%. Then, both approaches are applied as active strategies in considering quantitative variables. First, if the value of year t is missing, the value of year $t + 1$ is imputed from the same variable on year t. Second, a stochastic imputation is implemented. Given the nonmissing values for a given variable, a value is sampled from the empirical distribution and a small noise factor is added to that value to increase variability. The noise is sampled from a normal distribution with a mean of 0 and standard deviation of 1.

After we deal with the quantitative missing data, all the quantitative variables are standardized in accordance with the common formula, $\frac{x-\mu}{\sigma}$, where μ is the mean of a variable considered and σ the standard deviation of that variable.

Furthermore, the one-hot encoding technique is applied to qualitative variables to convert them into binary variables, hence increasing the predictive performance of the ML models. Basically, a dummy variable is built for each level of the qualitative characteristic.

5. Predicting risk from the use of tax havens

After the preprocessing phase, the final dataset is composed of 3575 observations (*GUOs*), and each observation is described by 207 variables. In Appendix B, we report the complete list of input variables described at the macro level (e.g., a single variable

[8] When dealing with DT, we need to handle the usual ML trade-off issue: higher complexity (number of levels/leaves) typically leads to higher in-sample prediction power at the cost of a decrease in power out-sample (overfitting problem). Imposing one stopping rule is a way to manage that issue.

[9] Considering the quantitative variables, we have 17% of missing values per variables on average. One variable has more than 50% missing values. In addition, 10% of observations have more than 50% missing values.

Table 4
Performance metrics for the validation set.

	Accuracy	Recall (Sensitivity)	Precision	F1 score
Logistic regression	0.594	0.519	0.576	0.479
Decision Tree	0.754	0.547	0.754	0.637
Random Forest	**0.776**	**0.603**	**0.781**	**0.679**

Source: Authors' calculation based on the ORBIS database.

“activity” is reported, instead of different sectors of operation). The target variable is a binary feature that leads to a classification problem. An observation takes a value of 1 if it has one or more subsidiaries in ATP countries, and 0 otherwise. Almost 61% of the observations take a value of 0, and the remainder take a value of 1. The input variables are related to the period 2015–2019. The target variable is calculated using information from 2021.

As in a classical ML procedure, the dataset is split between training (80%) and testing (20%). At the beginning of the procedure, logistic regression, DT and RF are compared in order to select the more powerful approach (for more detail about default parameters and implementation, see Therneau et al. (2022), Breiman et al. (2022)). For this purpose, we use a 10-fold cross-validation technique to identify the best model. Table 4 summarizes the main average performance metrics.

The social costs of type I errors (false positives) and type II errors (false negatives) vary across the problems under examination, so, in principle, it is appropriate to consider both types of errors. They are both measured by *accuracy*, which is the ratio between correctly classified cases and the total number of cases. Logistic regression achieves accuracy of 0.594 and DT 0.754. RF is slightly better with respect to DT in terms of accuracy, leading to a greater ability to correctly predict both classes and to reduce both types of error.

Recall (also known as sensitivity or the true positive rate) measures the ratio between the true positive (observations correctly predicted on positive class) and the number of actual observations of positive class. *Precision* measures the ratio between true positive (observations correctly predicted on positive class) and the number of predicted observations of positive class. A good performance metric that considers both precision and recall of ML approach is the F1 score, which is the harmonic mean of precision and recall. RF has a value of F1 that equals 0.679, whereas DT has a value of 0.637 and logistic regression a value of 0.479; again, RF achieves both higher precision and higher recall, therefore it is selected for the next research phase.

ML models have superior prediction performance compared to a traditional econometric model. In this comparison, logistic regression considers only the variables as principal effects and does not include interactions or quadratic terms. Although it would be possible to include them, and therefore to increase the predictive performance of the logistic regression, ML models are the most sensible way to deal with linear and nonlinear interactions among predictors.

5.1. Selection of tax years

RF models do not directly handle time, so to analyze the time impact, we decompose the initial dataset in different subsets. More precisely, six subsets of increasing time coverage are considered. The first one includes only values measured in 2019, the second subset adds 2018, the third adds 2017, and the fourth includes 2016; the fifth is the entire dataset, which includes information from 2015 until 2019. All subsets contain the categorical variables because they are time invariant. Additionally, to empirically understand whether the inclusion of additional years is relevant to improvement in the performance of our approach, a subset containing all the categorical variables and the average of all quantitative variables over the full period is considered.We apply a 10-fold cross-validation technique to each subset and estimate RF with default parameters.

The results are presented in Table 5. Two things emerge. Adding information from 2018 and 2017 to the first subset, which includes only 2019, steadily increases the predictive performance according to all measures. On the contrary, adding 2016 decreases the predictive performance, and adding 2015 does not increase it. Second, using average values, rather than values from single years decreases, albeit slightly, the predictive performance, probably because it flattens the relationships between the variables.

The pre-processing phase may have had an impact on these results because the missing values imputation approach implicitly adds redundant information. However, the finding that the information content of the data used for prediction decreases over time is not new in the literature (Bajgar et al., 2020).

On the basis of this evidence, only data from the period 2017–2019 are considered for the quantitative variables used for prediction.

Table 5
Predictive performance of RF using different time subsets. Column *Avg* reports the performance considering the average over time.

Metrics	2019	2019–2018	2019–2017	2019–2016	2019–2015	Avg
Accuracy	0.771	0.772	0.775	0.773	0.776	0.765
Recall (sensitivity)	0.596	0.600	0.606	0.604	0.603	0.575
Precision	0.775	0.774	0.778	0.775	0.781	0.775
F1 score	0.673	0.675	0.680	0.677	0.679	0.659

Source: Authors' calculation based on the ORBIS database.

Table 6
Hyperparameters involved in the optimization.

Hyperparameter	Range	Step
ntree	{250, 500, 1000, 1500}	–
mtry	{5 − 18}	1

Source: Authors' calculation based on the ORBIS database.

5.2. Fine tuning of the parameter optimization

RF, like many other ML approaches, is characterized by several hyperparameters, which dramatically influence the performance of models. Therefore, fine tuning of these hyper-parameters, that is, hyper-parameter optimization, is important. For this reason, a simple *grid-search* technique is applied in order to find the best set of hyperparameters, considering the F1 score as the objective function to be maximized. Initially, 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 6 reports the hyperparameters involved in the optimization.

In order to select the best setting, a 10-fold cross-validation approach is used on the initial training set. In Fig. 2, the F1 score for each combination of values is reported. The best setting is *ntree* = 250 and *mtry* = 13, which has accuracy of 0.780, recall of 0.613, precision of 0.785, and F1 score of 0.687.

In our context, recall is important because every MNC that uses a tax haven, without being predicted as such, represents a serious risk to the efficiency and reputation of the tax administration.

To further increase recall, the probability threshold used to assign a new observation to label 1, $P(Y = 1|X = x)$, is optimized. Here as well, 10-fold cross-validation is implemented. In order to find the best compromise between precision and recall, the F1 score is considered the objective function. Fig. 3 shows F1 scores over the interval {0.05, 0.95} of the probability threshold. The highest F1 score is reached when the probability threshold is 0.40. This means that a new observation will be take a value of one when $P(Y = 1|X = x) > 0.40$. In the 10-fold cross-validation setting, this algorithmic configuration achieves accuracy of 0.757, recall of 0.722, precision of 0.682, and F1 score of 0.701 (on average).

RF and its best configuration of parameters (*ntree* = 250 and *mtry* = 13) is then retrained on the entire 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 = 0.749, recall = 0.672, precision = 0.662 and F1 score = 0.667. Fig. 4 shows the Receiving Operating Characteristic (ROC) curve. Considering the final performance, RF seems to be a stable model for predicting the target variable.

A fairly parsimonious RF model that only uses publicly available accounting data for GUOs for the period 2017–2019 is used to predict the presence of a subsidiary of that GUO in a tax haven by spring 2021. In particular, the final model has an accuracy rate of 75% and recall of 67.2% on the training set. Thus, the probability of identifying a GUO that will locate a subsidiary in a tax haven is (almost) twice as high as that obtained by randomly selecting from the set of all GUOs, which, in turn, equals the frequency of GUOs that will actually have a subsidiary in a tax haven (39.

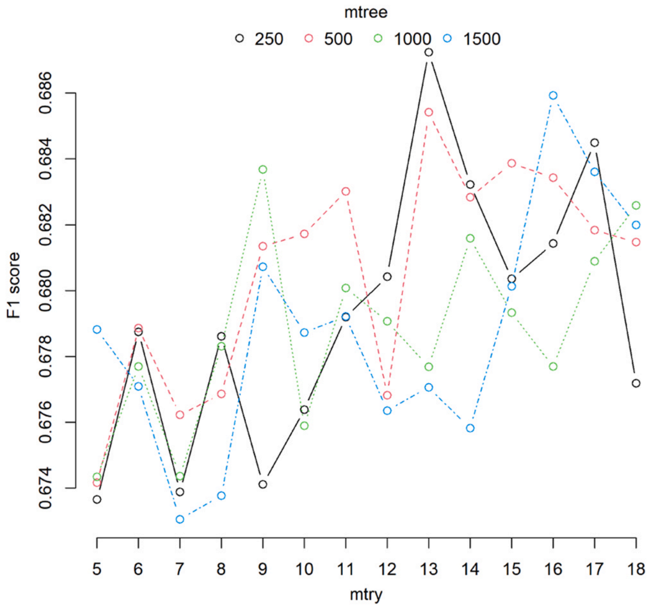

Fig. 2. Hyperparameters fine-tuning performance with respect to the F1 score.
Source: Authors' calculation based on the ORBIS database.

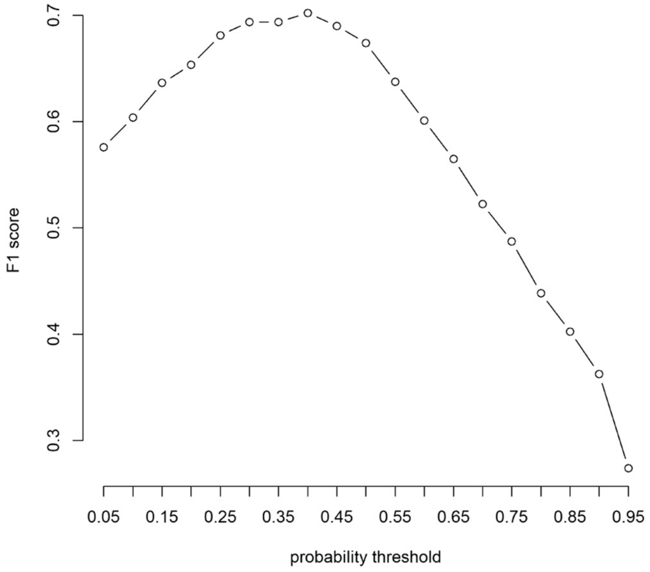

Fig. 3. F1 score behavior over the probability threshold.
Source: Authors' calculation based on the ORBIS database.

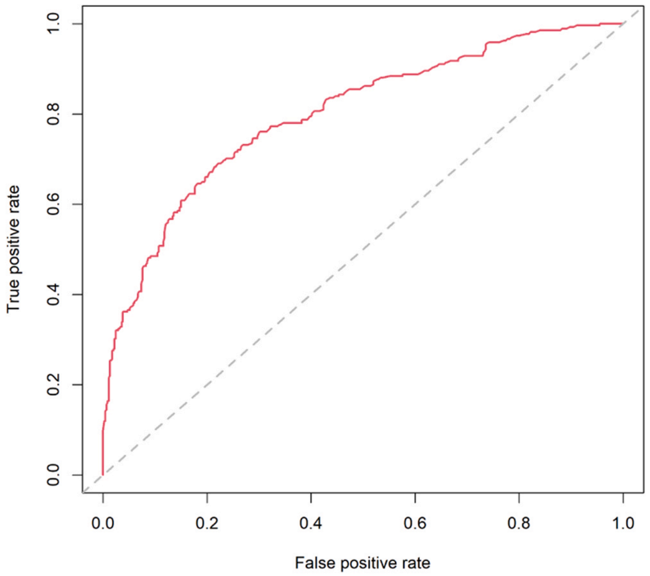

Fig. 4. ROC curve.
Source: Authors' calculation based on the ORBIS database.

5.3. *Variable importance*

Variable importance can easily be measured with decision trees and random forests. Decision trees iteratively split a dataset with the goal of decreasing the heterogeneity of the target variable within each subset as much as possible. Each split is based on a variable and a threshold. Therefore, the importance of any variable can be measured by its ability to decrease the weighted heterogeneity in each tree or the *impurity* of the tree. The importance of avariable within an RF model is obtained to average this decrease in impurity across trees (Hastie et al., 2001; Breiman, 2001).

We divide, somewhat arbitrarily, the variables into six subsets: .

- profitability: variables describing the profitability of GUOs;
- size: variables describing the size of GUOs;
- financial: variables describing the financial structure of GUOs;
- categorical: variables describing some time-invariant features of GUO (country, sector, year of incorporation, etc.);

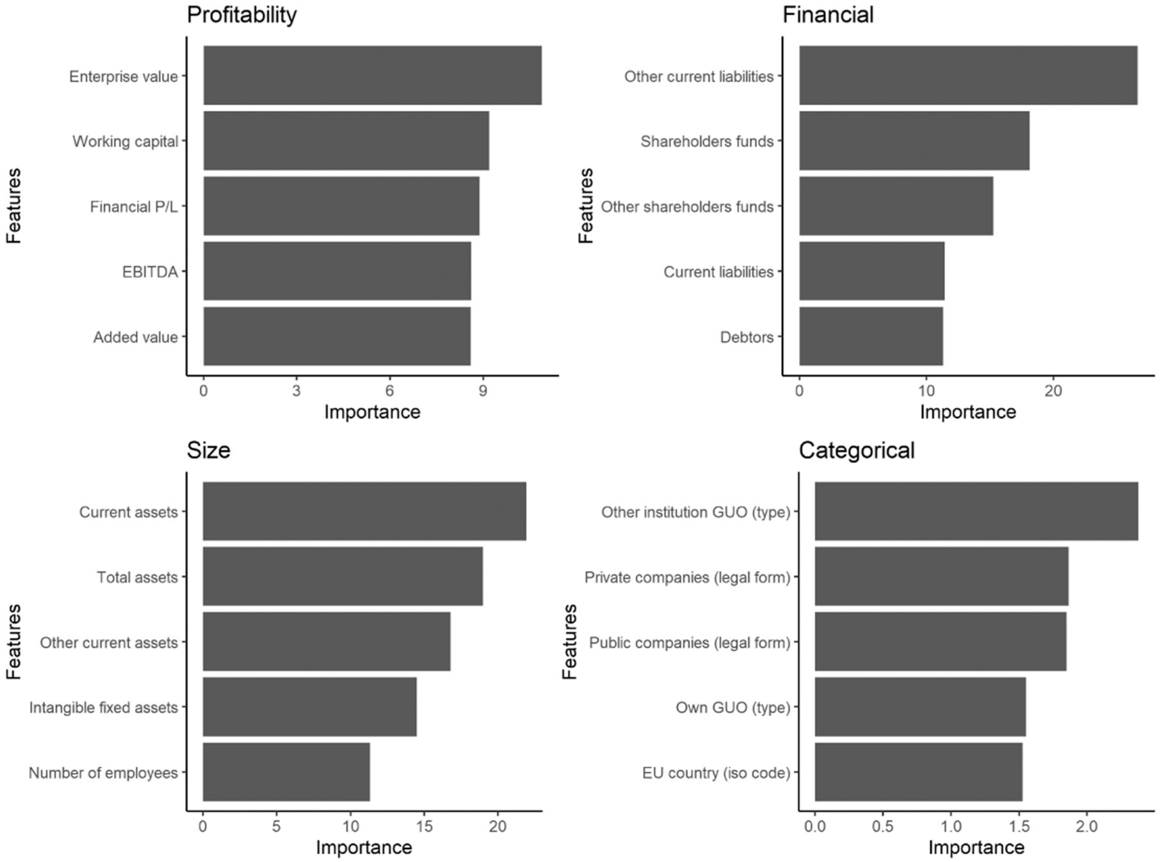

Fig. 5. Variables in order of importance and divided into four groups based on the type of infor mation: profitability, financial, size, and category.
Source: authors' own elaboration on ORBIS database.

- positive profitability: variables with a positive sign in the profit-and-loss state- ment of GUOs;
- negative profitability: variables with a negative sign in the profit-and-loss statement of GUOs;

Admittedly, some categories and variables are interrelated. In particular, some variables that we classify as "financial"—such as the amount of current liabilities or the value of capital—are closely related to other variables—such as current or total assets—that we classify as "size."

Fig. 5 reports the five most important variables for the first four groups, and Fig. 6 reports the three most important variables for the last two (profitability-related) subsets. To increase the interpretability of the output, the reported importance is the average of the variable importance values of each variable over the relevant period (e.g., the importance of the variable "other current liabilities" is the average of the variable importance values of "other current liabilities" in the period 2017–2019).

It appears that variables describing the financial structure, the profitability, and the size of a GUO, are the most important predictors. In particular, the "best" predictor of the probability of using tax havens is the value of "other current liabilities"—a category that lumps together all kinds of short-term debt—the second- and third-most important are sales and operating revenues, respectively, while the fourth- and fifth-most important are current and total assets, respectively.

It could be argued that these results are not surprising, considering the existing literature. Intercompany debt is frequently used to shift profits, and thus the financial structure is clearly relevant for predicting tax planning (Grubert, 2019). Size has already been found to be associated with the use of tax havens (Gallemore et al., 2014), and, there is no reason to use tax havens in absence of profits. However, the fact that these variables are more important than others does not imply that they can be used instead. In an RF model, predictive performance increases when at least two of these variables are used together, allowing all the possible nonlinear interactions among them.

To illustrate this last point, we now consider the most important five variables—other current liabilities, sales, operating revenue turnover, current assets, and total assets—and we construct five corresponding RF models, from the simplest one, which uses only other current liabilities, to one that employs all the five variables as predictors. Table 7 shows the results in terms of recall and F1 scores and shows that predictive performance increases when at least two of the top five variables are included in the RF model. In addition, we highlight that a logistic regression with only the first variable (other current liabilities) achieves recall of 0. If we consider a logistic regression with all five variables and interactions, we obtain recall of 0.250 and an F1 score of 0.375.

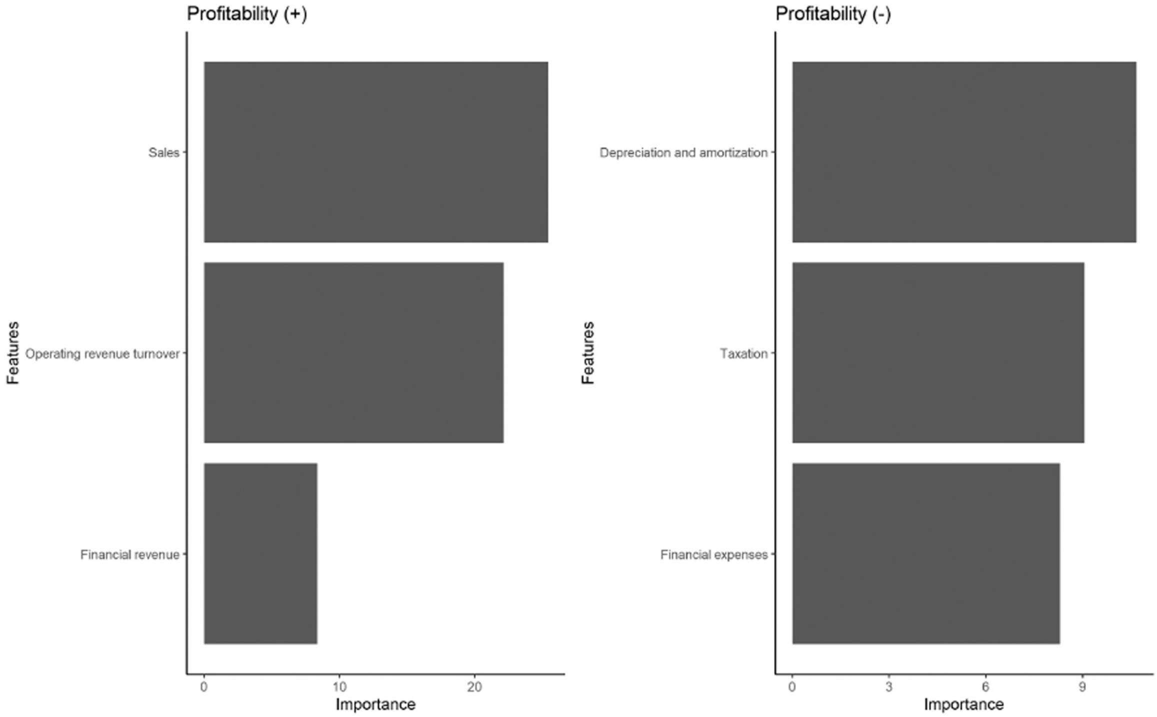

Fig. 6. Variable importance of positive and negative variables for profitability.
Source: Authors' calculation based on the ORBIS database.

Table 7
Performance metrics on test sets for five models with an increasing number of variables.

	Variables	Recall (Sensitivity)	F1 score
Model 1	other current liabilities	0.579	0.534
Model 2	Model 1 + sales	0.645	0.586
Model 3	Model 2 + operating revenues	0.654	0.590
Model 4	Model 3 + current assets	0.629	0.575
Model 5	Model 4 + total assets	0.646	0.593

Source: Authors' calculation based on the ORBIS database.

One last technical remark is needed. The use of highly correlated variables as predictors can be an issue in analyzing variable importance.[10] For this reason, the literature proposes some techniques for variable selection such as the one proposed by Genuer et al. (2010). In this paper, the number of variables randomly sampled as candidates at each split (*mtry*) is optimized by the use of cross-validation. From the perspective of prediction, the impact of the inclusion of the correlated variables is thus mitigated.

6. Predicting the intensity in the use of tax havens

In this section, we propose an approach for predicting the proportion of subsidiaries that a GUO will locate 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 0, no subsidiary is located in a tax haven country. If the value is 100, all subsidiaries are located in a tax haven country.

We operate in a sequential manner. First, we predict whether a GUO will at least have one subsidiary in an ATPC, an exercise equivalent to the one conducted earlier. Second, we predict the proportion of subsidiaries in an ATPC conditional on the GUO having at least one subsidiary in an ATPC. Fig. 7 shows the steps in the sequential approach. Model 1 is a classification model, and we exploit the RF with $ntree = 250$, $mtry = 13$, and $P(Y = 1|X = x) > 0.40$, presented in Section 5. Like Model 1, Model 2 is a 10-fold cross-validation optimized regression RF with $ntree = 1500$ and $mtry = 15$. The fine-tuning optimization procedure and the data are discussed in Section 5. Unlike Model 1, the training set contains the new quantitative target variable, which is calculated as the ratio

[10] With two highly correlated variables (e.g., t and t + 1 depreciation), once one is used in the model, the importance of the other is going to be reduced because the ability of the second variable to decrease the weighted heterogeneity in each tree is greatly lowered as the first variable decreased the impurity of the tree already.

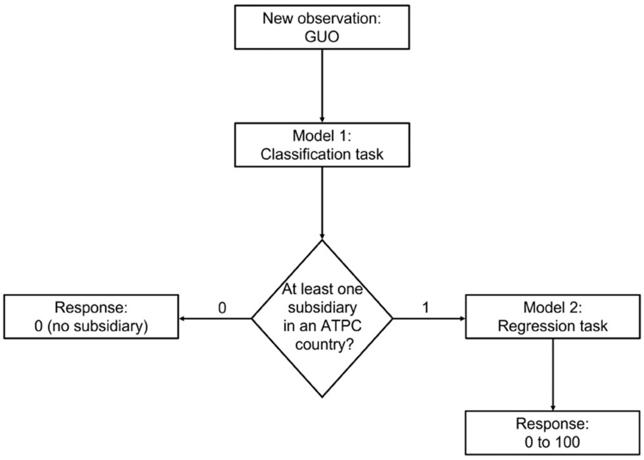

Fig. 7. Scheme of the sequential approach.

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 entire approach is called *sequential random forest* (sequential RF).

As performance metric of the sequential approach, we consider the RMSE defined as in Equation (1).

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

In Equation (1), N_t is the number of observations in the test set, y_i is the actual proportion for the ith observation and $\hat{y}_i$ is the prediction for the ith observation. The sequential RF achieved an RMSE of 41.24 and correctly assigned 0–79% of GUOs. This result is compared with the prediction obtained by Model 2 on the entire test set without, as first step, applying Model 1. In this case, we obtained a higher RMSE (61.75) and no GUOs were predicted with 0 subsidiaries in non-tax haven countries. From a practical perspective, this suggests that a sequential RF is a better model for solving the general problem of identifying GUOs with subsidiaries in a tax haven or non-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 5.3. Here, the variable importance is calculated as the mean decrease in Mean Square Error (MSE). MSE is defined as $MSE = \frac{1}{N_t}\sum_{i=1}^{N_t}(y_i - \hat{y}_i)^2$. Fig. 8 reports the most five important variables for profitability, financial, size, and category. Fig. 9 reports the three most important variables for positive and negative profitability subsets. The variables are very similar to those displayed in Figs. 5 and 6.

Although the most important variables for this prediction are related to the GUO's financial structure, profitability, and size—as in the prediction of the risk to locate a subsidiary in a tax haven—they are not exactly the same. In particular, the value of capital, rather than that of current liabilities, is the most important financial variable and the cost of employees is the most important size variable.

This is an important result: to predict the intensity in the use of tax havens the tax authority should look at variables that are similar and related, but not exactly equal, to those that it should use to predict the mere presence of a subsidiary in a tax haven. This difference may be important when the tax authority implements the results of the prediction exercise, as we illustrate in the next section.

7. The use of predictions by tax authorities

The prediction exercises that we conducted in previous sections provide national tax authorities with a list of GUOs that are more likely to locate or *maintain* a subsidiary in a tax haven, or that are predicted to have a higher share of subsidiaries in a tax haven.

The same methodological approach could be used when the tax authority also has information about the existence of a subsidiary in a tax haven at the time of the prediction. This would enables the tax authority to obtain separate lists of GUOs that are more likely to locate the first or additional subsidiaries in a tax haven and additional lists of GUOs with higher predicted shares of first (or additional) subsidiaries in tax havens. The selection of GUOs to be treated clearly depends on the budget available. For example, if the available budget is small, the tax authority should focus on GUOs with highest probability of reaching a very high share of subsidiaries in a tax haven.

The selected GUOs should be subject to a special *verification procedure*. This procedure, typically to be conducted partly from the desk and partly on site, are intrinsically different from traditional audits. Audits are conducted ex-post, with the goal of confirming whether a given behavior has occurred. The verification procedure we are discussing here is preventive, as it aims to preempt tax evasion. The necessity of strengthening this kind of ex-ante activity is often stressed by the tax administration, as seen by looking at reports by the OECD's Forum on Tax Administration (FTA). In particular, large companies have increased their compliance activity with nontraditional approaches. The trend is to move from post-filing of tax return examination to real-time evaluation of risk and compliance issue resolution. Many countries have implemented various programs to provide certainty to large taxpayers and early identification and resolution of compliance issues (OECD, 2009).

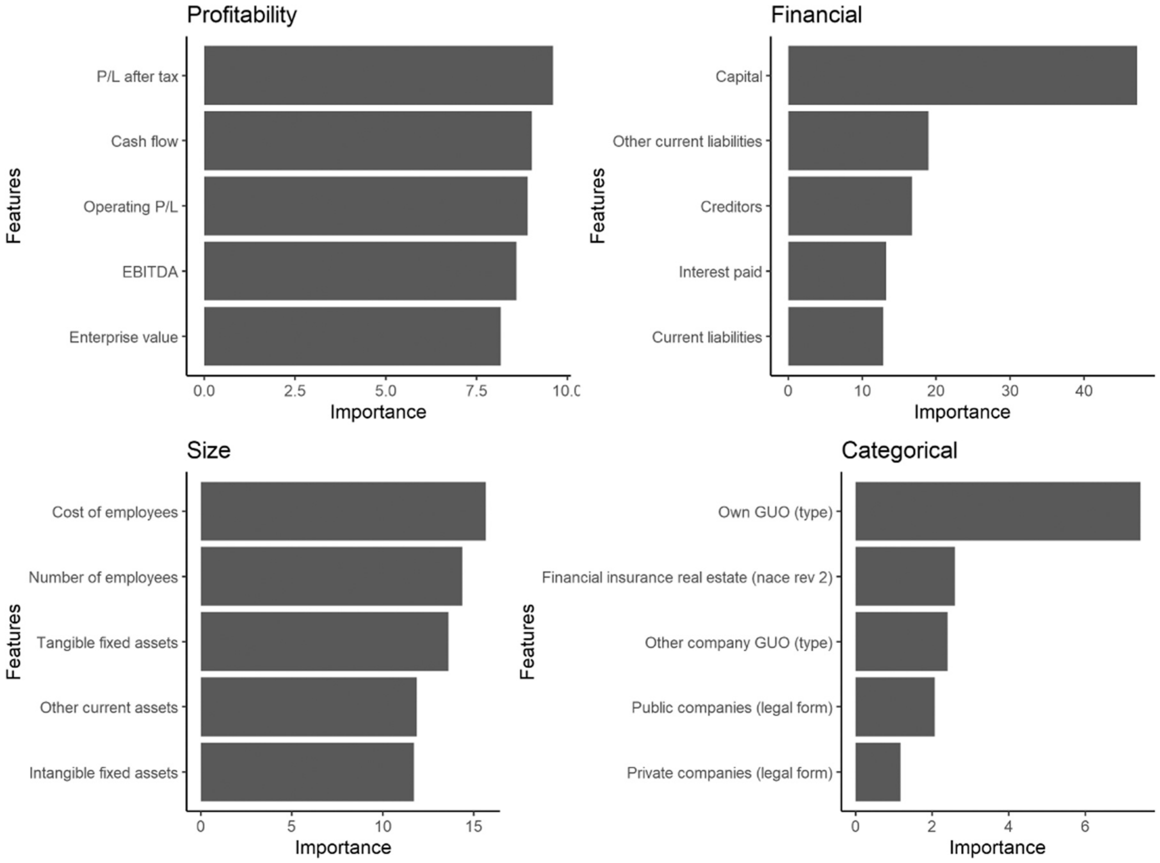

Fig. 8. Variables ordered for importance and divided in four groups related to the type of information: profitability, financial, size, and category.
Source: authors' own elaboration on the ORBIS database.

The ex-ante verification procedure in this case should work as follows. GUOs involved should be asked to disclose the operational details of any subsidiary that it will create in tax havens in the foreseeable future (e.g., in the following year). The information should be such that the tax authority can issue a conditional tax ruling, that is a statement about the legitimacy of the scheme to which the tax authority commits. This commitment is the first incentive for the GUO, which knows in advance whether a certain program is legitimate (lower uncertainty costs). Clearly, if the ruling is positive, such that the location decision is not deemed to lead to tax evasion or tax avoidance, the commitment is conditional on the group's operating consistently with the information disclosed during the verification procedure. Additional incentives may be provided to high-risk GUOs, such as a reduction in penalties if a dispute arises between the parties or as the admission to simplified procedures for the reimbursement of tax credits. However, a GUO may refuse to cooperate or fraudulently deny any intention to locate a subsidiary in a tax haven. In that case, as well as if the GUO deviates from the information previously disclosed, the GUO should know that both the probability of being audited and of paying sanctions in the event of a successful audit are increased by a given magnitude.

The verification procedure described uses elements that are typical of policies already in place, namely, cooperative compliance and tax rulings, but it grants the tax authority additional powers.

The term "cooperative compliance" (CoCo) was proposed by the OECD (2013) as an evolution in the "enhanced relationship" between tax authorities and taxpayers advocated by the FTA in 2008. CoCo programs cover a wide range of policies adopted by 21 FTA member countries to increase voluntary tax compliance by large businesses, especially MNCs. In particular, CoCo programs aim to reduce base erosion and profit-shifting practices that happen within MNCs using a variety of programs (OECD, 2013). Also, CoCo programs can be interpreted as ways in which inequality is addressed in the distribution of the tax burden between large MNCs and small domestic firms. When a tax authority and a company decide to enter into a CoCo agreement, they commit to mutual information disclosure. On the one hand, businesses disclose information about the schemes that 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 privileged access to a ruling and a reduction in the likelihood of being audited. The CoCo agreement can thus be summarized as "transparency in exchange of certainty."

The main difference between CoCo programs and the policy we advocate here is that the former are based on voluntary agreements, such that companies self-select for the treatment, whereas we propose compulsory verification, requiring a special additional powers granted to the tax authority. The difference in these two treatments is justified by the fact that the risk analysis shows

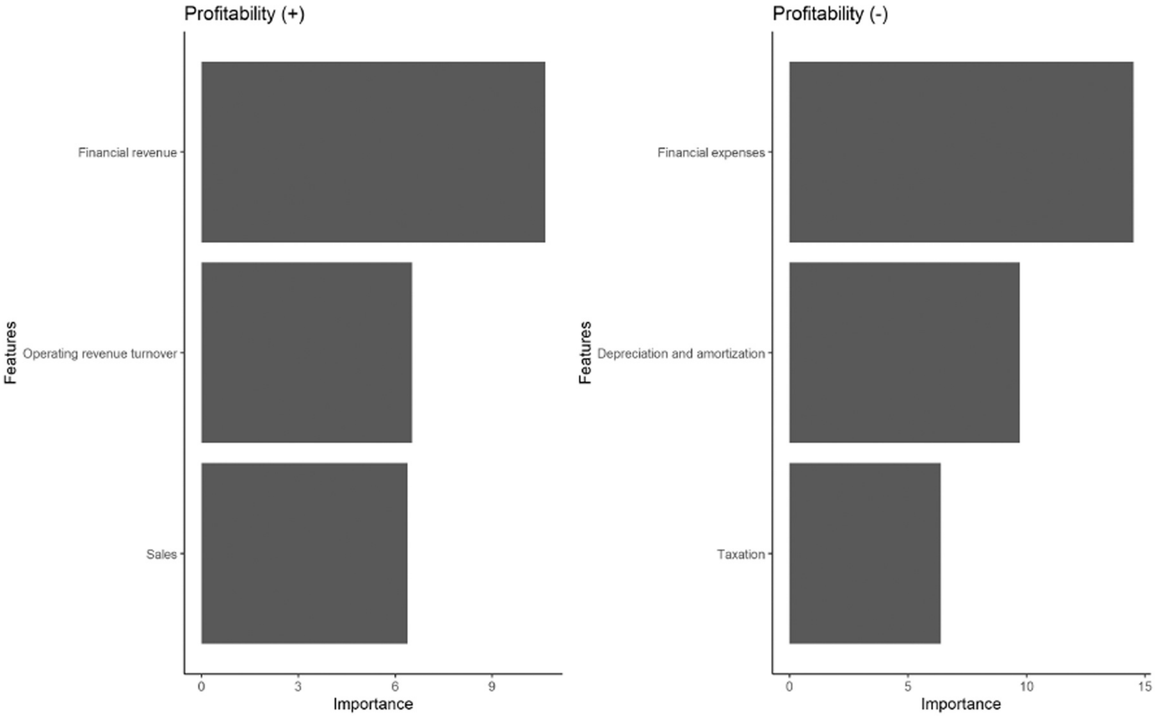

Fig. 9. Variable importance for profitability, positive and negative.
Source: Authors' calculation based on the ORBIS database.

evidence of a high probability of a specific risk, that is, the use of subsidiaries located in tax havens. Therefore, a subset of MNCs usually eligible for CoCo programs would be subject to these verification activities.

Tax rulings are common among all tax administrations. However, tax rulings are normally issued by the tax authority at the request of the taxpayer. Therefore, it is the company that decides which information to disclose and when to disclose it. Our proposal envisages a tax ruling concerning the reason and nature of location decisions, as defined by the tax administration. Clearly, incentives are needed to elicit information from the company, which is why, in our proposal, the tax ruling is accompanied by additional benefits.

8. Concluding remarks

In this paper, we show that the presence of a subsidiary of a European GUO in a tax haven at year *t* can be predicted reasonably well using publicly available accounting data about the financial structure, size and profitability of the GUO for the period between *t-4* and *t-2* with an RF model, which clearly outperforms more traditional models (e.g., logistic regression). Also, we show that the same applies if the tax authority wants to predict the intensity in the use of tax havens, that is, the share of subsidiaries located in tax havens, although the specific variables to be used for prediction slightly change.

We discuss the use of these predictions in the context of a new ex-ante administrative approach that, in our view, should complement the traditional ex-post legal approach, thus providing a concrete example of the new compliance risk management advocated, among others, by the OECD. To do this, we assume that also the information on the number and location of subsidiaries at the time of prediction is available to the tax authority. So, the tax authority could draw up a list of high-risk GUOs and set up a special verification procedure, with the goal of collecting information from the GUO on how the subsidiaries to be located in tax havens in the near future will operate. Commitment to the tax ruling on the application of anti-avoidance rules, as well as additional incentives for GUOs disclosing the relevant information should be included in the program. Also, an increased probability of auditing and increased sanctions for high-risk GUOs that do not disclose this information or deviate from it should be part of the policy.

This kind of policy can be enhanced by further improving available data. For example, under the OECD's BEPS (Base Erosion and Profit Shifting) Action 13, all large MNCs are required to prepare a country-by-country (CbC) report with aggregate data on the global allocation of income, profit, taxes paid and economic activity among tax jurisdictions in which it operates. This CbC report is shared with tax administrations in these jurisdictions, for use in high-level transfer pricing and BEPS risk assessments. In our context, this information can be used by the tax authorities to differentiate the risk of using intercompany debt or of setting up SPV or other specific schemes for base erosion and profit shifting. Accordingly, specific profile risks, that is, probabilities that a given tax scheme will be used in a given country, could be obtained using the methodological approach illustrated here.

Finally, the credibility of the policy advocated here does not rely exclusively on the incentives and disincentives of the preventive administrative scheme.

On October 8, 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 digitization of the economy. According to Pillar 1, tax rights over 25% of the residual profit of the largest and most profitable MNCs would be reallocated to the jurisdictions where their customers and users are located. According to Pillar 2, GloBE (Global Anti-Base Erosion) rules provide a global minimum tax of 15% on all MNCs with annual revenue over 750 million euros. The two-pillar strategy, if implemented, in theory reduces the incentives for relocating activities and subsidiaries in tax havens to reduce taxes. Clearly, the scope of the agreement and the number of countries that eventually adhere to it are crucial parts of the implementation of the strategy. Given the large number of countries that have agreed to the Inclusive Framework (141 as of December 2021), it seems fair to conjecture that, in the near future, MNCs will face higher transaction costs for locating or maintaining subsidiaries in tax havens. The type of data mining and predictive administrative action explored in this paper could be a more credible tool in this context. Any MNCs that is considering the use of 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 of being closely monitored by the tax authority, which uses the approach we explore in this paper.

Appendix A. Appendix

In the case of OTHER_INSTITUTION_GUO, to reconstruct the corporate group, we proceed as follows. We identify the set of all the subsidiaries for each LISTED company and INTER- MEDIATE company, where a subsidiary is a company at least 50.01% of whose shares are owned by the GUO. Fig. A.1 illustrates the ownership structure for the set of OTHER_INSTITUTIONS_GUO.

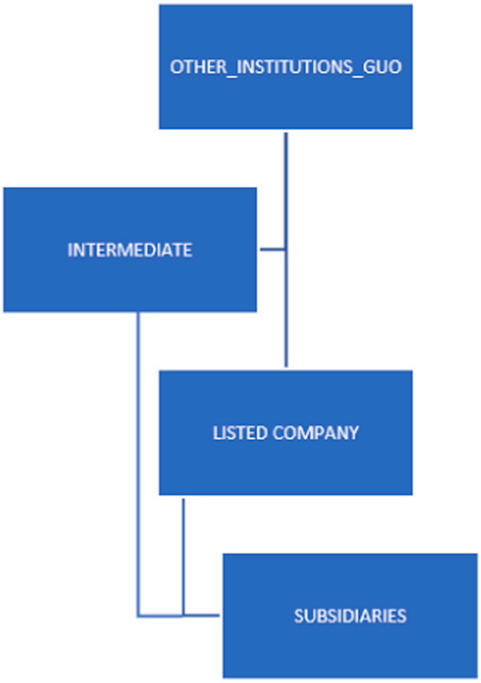

Fig. A.1. Diagram of the reconstruction of a group headed by an OTHER_INSTITUTION_GUO.

B. Appendix

The following table reports the list of all the variables (at the macro level) used as input in machine learning approaches. Table B.1.

Table B.1
List of the variables used in each year.

Variable	Description
Added Value	Profit for period + Depreciation + Taxation + Interests paid + Cost of employees
Capital	Issued share capital (authorized capital)
Cash and Cash Equivalent	The amount of cash at bank and in hand of the Company
Cash Flow	Profit for period + Depreciation
Cost of Employees	Detail of all the employees costs of the Company (including pension costs)
Creditors	Debts to suppliers and contractors (trade creditors)
Current Assets	Total amount of current assets (Stocks + Debtors + Other current assets)
Current Liabilities	Current liabilities of the company (Loans + Creditors + Other current liabilities)
Debtors	Trade receivables (from clients and customers only)
Depreciation and Amortization	Total amount of depreciation and amortization of the assets
EBITDA	Operating profit + Depreciation
Enterprise Value	Estimate of the total value on the market of the company operations by the sum of its market capitalization, long-term debts and loans (to financial institutions) minus cash and cash equivalent
Financial Expenses	All financial expenses such as interest charges, write-off financial assets
Financial P/L	Result from financial activities of the company (financial revenue - financial expenses)
Financial Revenue	All financial revenue, such as interest and income from shares
Fixed Assets	Total amount (after depreciation) of non current assets (Intangible assets + Tangible assets + Other fixed assets)
Intangible Fixed Assets	All intangible assets, such as formation expenses, research expenses, goodwill, development expenses and all other expenses with a long-term effect
Interest Paid	Total amount of interest charges paid for shares or loans
Long Term Debt	Long-term financial debt e.g., to credit institutions, bonds
Material Costs	Detail of the purchases of goods (raw materials + finished goods). No services
NACE Rev. 2	Classification of economic activities
Total number of employees	Total number of employees included in the company’s payroll
Operating P/L [=EBIT]	EBIT. All operating revenues - all operating expenses
Operating Revenue (Turnover)	Total operating revenues (Net sales + Other operating revenues+ Stock variations)
Other Current Assets	All other current assets such as receivables from other sources (taxes, group companies), short-term investment of money, and cash at bank and in hand
Other Current Liabilities	Other current liabilities, such as pensions, personnel costs, taxes, intragroup debt, and accounts received in advance
Other Fixed Assets	All other fixed assets, such as long-term investments, shares and participation, and pension funds
Other Non-Current Liabilities	Other long-term liabilities (trade debt, group companies, pension loans, etc.) + provisions + deferred taxes
Other Shareholders Funds	All shareholder funds not linked to the issued capital, such as reserve capital, undistributed profit, including also minor- ity interests, if any
P/L after Tax	Profit before taxation - Taxation
P/L before Tax	Operating profit + Financial profit
P/L for Period [=Net income]	Net income for the year, before deduction of minority interests
Sales	Net sales
Shareholders Funds	Total equity (Capital + Other shareholders funds)
Standardized Legal Form	Legal form of companies
Stocks	Total inventory
Tangible Fixed Assets	All tangible assets, such as buildings and machinery
Taxation	All taxes related to the accounting period (paid, accrued, or deferred)
Total Assets	Total assets (Fixed assets + Current assets)
Working Capital	Indicates how much capital is used for day-to-day activities = Stocks + Debtors - Creditors

Source: ORBIS User Guide.

C. Random Forest

First, we briefly introduce how a Decision Tree (DT) is built and then describe how a Random Forest (RF) works in practice.

If we consider a classification problem, we have observations in two or more known classes, and our ultimate goal is to develop a rule or a set of rules that assign each observation to a class. Each observation is characterized by numerical or categorical input variables. A possible solution from classical statistics is the use of logistic regression. In brief, logistic regression searches for linear combination of the input variables in order to assign observations to a specific class.

DT, more specifically a classification tree, uses recursive binary splits to identify regions that are increasingly homogeneous with respect to the class variable. These regions are called nodes. Basically, at each step of the approach, an optimization is performed to identify the most homogeneous subgroups (i.e., input variable in the node and decision rule of cut-off to be used to split the data). Homogeneity is measured by the Gini index (Breiman, 2001). The tree continues to grow (splitting process) until further subdivision no longer reduces the Gini index. At this point, we obtain a fully-grown tree where the final nodes are called terminal nodes. The performance of the tree can be improved by applying specific techniques (i.e., *pruning* (Breiman, 2001)) that are based on evaluating classification error and selecting only the most informative part of the tree.

DT has some limitations: (1) trees generally do not have the same level of predictive accuracy as some other approaches, and (2) trees are not robust to change in the data. Even a small change in the training data can lead to a large change in the final estimated tree.

RF tries to overcome both limitations of DT. RF builds a number of decision trees and assigns an observation to a class based on the prediction from all the trees. First, RF selects a set of bootstrap samples from the training data. Each bootstrap training sample is made up of random selected (with replacement) observations from the initial training set. Observations that do not appear in the bootstrap training sample are called *out-of-bag* observations. Now, a classification tree is built on each bootstrap training sample. In RF, when building these DTs, every time a split in a tree is considered, a random sample of m input variables is chosen as split candidates from the full set of p input variables. The split is allowed to use only one of those m input variables. A fresh sample of m input variables is taken at each split (James et al., 2009).

When all the trees are fullgrown, every tree is used to predict the out-of-bag observations. Given all the predictions for an out-of-bag observation, a majority vote is used to assign the final class, with ties split randomly. At this point, the error rate is calculated for every observation using out-of-bag predictions and averaged over all observations. Out-of-bag estimations can be seen as cross-validated accuracy estimates.

As described, an important feature of RF is that it forces each split in the tree construction process to consider only a subset of input variables. As pointed out in James et al. (2009), on average $(p - m)/p$ of the splits do not consider the most impactful input variable, so other input variables will have more of a chance—therefore, making the average of the resulting trees less variable and more reliable.

Appendix C. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ecosys.2023.101090.

References

Andini, Boldrini, D.B. Ciani, D'Ignazio, Paladini, 2022. Machine learning in the service of policy targeting: the case of public credit guarantees. J. Econ. Behav. Organ. 198, 434–475.

Badal-Valero, Alvarez-Jareño, Pavía, 2018. Combining benford's law and machine learning to detect money laundering. an actual spanish court case. Forensic Sci. Int. 282, 24–34.

Bajgar, M., Berlingieri, G., Calligaris, S., Criscuolo, C., Timmis, J., 2020. Coverage and representativeness of orbis data. OECD Sci., Technol. Ind. Work. Pap. 1 (1), 1–63.

Barboza, Kimura, Altman, 2017. Machine learning models and bankruptcy prediction. Expert Syst. Appl. 83, 405–417.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Breiman, L., A. Cutler, A. Liaw, M. Wiener , 2022. randomForest: Breiman and Cutler's Random Forests for Classification and Regression. R Foundation. R package version 4.7–1.1.

Clifford, S., 2019. Taxing multinationals beyond borders: financial and locational responses to cfc rules. J. Public Econ. 173 (May), 44–71.

Cox, D.R., 1958. The regression analysis of binary sequences. J. R. Stat. Soc. Ser. B (Methodol.) 20 (2), 215–242.

De Simone, L., Mills, L.F., Stomberg, B., 2019. Using irs data to identify income shifting to foreign affiliates. Rev. Account. Stud. 24, 694–730.

European Commission, 2012. Commission Recommendation of 6 December 2012 on aggressive tax planning.European Commission.

European Council , 2021. Council conclusions on the revised eu list of non-cooperative jurisdictions for tax purposes. Report, Council of the EU.

Gallemore, J., Maydew, E.L., Thornock, J.R., 2014. The reputational costs of tax avoidance. Contemp. Account. Res. 31 (4), 1103–1133.

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recognit. Lett. 31 (14), 2225–2236.

Grubert, H., 2019. Intangible income, intercompany transactions, income shifting, and the choice of location. Natl. Tax. J. 56 (1), 221–242.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Element of Statistical Learning. Springer Series in Statistics.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2009. An Introduction to Statistical Learning. Springer Series in Statistics.

Kaplow, L., 1992. Rules vs standards: an economic analysis. Duke Law J. 42 (1), 557–629.

Meldgaard, H., J. Bundgaard, K. DyppelWeber, 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., Dhaliwal, D.S., 2001. Cross-jurisdictional income shifting by u.s. multinationals: evidence from international bond offerings. J. Account. Res. 39 (3), 643–662.

OECD, 2009. Forum on tax administration: Compliance management of large business task group. Report, OECD publishing.

OECD, 2012. Hybrid mismatch arrangements: Tax policy and compliance issues. Report https://www.OECD.org/tax, OECD publishing.

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

Sadgali, I., Nawal, S., Faouzia, B., 2019. Performance of machine learning techniques in the detection of financial frauds. Procedia Comput. Sci. 148, 45–54.

Therneau, T., B. Atkinson, B. Ripley, 2022. rpart: Recursive Partitioning and Regression Trees. R Foundation. R package version 4.1.16.

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

Weisbach, D., 1999. Formalism in the tax law. Univ. Chic. Law Rev. 66 (1), 860–886.

Zucman, G., 2014. Taxing across borders: tracking personal wealth and corporate profits. J. Econ. Perspect. 28 (4), 121–148.