Retrieving Tax Audit Criteria to Estimate Tax Audit Impact.

PRELIMINARY VERSION

Daniele Spinelli[1] Paolo Berta[1] Alessandro Santoro[2] *

[1]Department of Statistics and Quantitative Methods, University of Milano-Bicocca

[2]Department of Economics, Management and Statistics, University of Milano-Bicocca

Abstract

The criteria used by tax authorities around the world to select taxpayers to be audited are mostly unknown. In this paper, we use a machine learning approach and a random forest model to retrieve these criteria from a panel of tax audits and tax reports referring to a large sample of Italian self-employeds and sole proprietorships. It emerges that to prioritize riskier taxpayers the tax authority uses the information it obtains during the so-called 'expiration period', i.e. the period going from the target year to the year of the audit. Using these audit criteria to match audited with unaudited taxpayers within a CEM approach, we also show that the audits do have a short-run positive impact on subsequent reports and also that even such a short-run effect is likely to justify an increase in the number of audits with respect to observed levels.
Keywords: Optimal Tax Administration, Enforcement Elasticity of Tax Revenue, Machine Learning.
JEL Numbers: H26; C55

*Corresponding author `daniele.spinelli@unimib.it`, Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, Edificio U7 20126 Milan (Italy). Authors thank the Italian Revenue Agency that has provided the data in the context of a research agreement with the University of Milan-Bicocca.

1 Introduction

In their authoritative assessment of tax administration literature, Slemrod and Yitzhaki (2002) noted that there was little systematic guidance offered by the public finance literature on "the reality of evasion, the necessity of enforcement and the costs of collection". Twenty years after, it would be fair to say that much evidence has been gained on the former (the reality of evasion) but very little progress has been made on the latter two issues. The credibility revolution in the study of tax compliance (Slemrod and Weber (2010)) has brought in a huge knowledge about the magnitude of tax evasion as well as on the impact of hypothetical random audits on tax compliance. But this massive flow of results (summarized by Mazzolini et al. (2021)), though reasonably consistent with each other, has failed to have any visible impact on actual enforcement policies.

This gap may be explained also by the empirical approach and the type of data used in the literature.

Most of the papers about the impact of audits on subsequent compliance are based on field-controlled experiments, where audits are conducted randomly (see, for a summary of these papers Mazzolini et al. (2021). The availability of a natural counterfactual ensures the internal validity of these studies, that, on average, yield positive and significant estimatesof the impact of audits on subsequent tax compliance. However, as Slemrod (2016) stresses, their external validity is more problematic for two reasons.

First, taxpayers audited within these field-controlled experiments are usually (albeit not always) informed that they have been randomly selected for research purposes. Thus, these audits may not have the same impact as an operational (real-world) audit would do. In principle, one may think that a real audit prompts a stronger reaction than a research audit, however the sign of the difference depends on the prevalence of the target effect or of the bomb-crater effect. In the former case, the audited taxpayer feels to be a target of the revenue agency only if the audit is a real one, and therefore the enforcement elasticity measured within field-controlled experiments could be underestimated. In the latter case, i.e. the prevalence of the bomb-crater effect, the audit taxpayer feels to be safer after she has been (really) audited, and therefore research audits would overestimate the elasticity (or, to be more precise, would underestimate the negative impact).

Second, and more importantly for the purposes of the present study, Slemrod (2016)recalls that taxpayers audited for research purposes may not be representative of those who are typically subject to audit, and their behavior may not be representative of those who are normally targeted for operational audits.

The latter remark is of particular importance here. As shown by Keen and Slemrod (2017), the measure of the enforcement elasticity is key to understand whether additional spending on enforcement is or not justified, once private (compliance) and administrative costs are taken into account. Now, although random audits are used in collaboration with academic and researchers, the vast majority of revenue agencies uses risk-based audit criteria, because they are (rightly) believed to be more efficient. Thus, in practice, the impact of interest is that associated with an increase of audits, conditional on the application of the audit criteria. A benevolent social planner can decide whether it is socially profitable to increase the budget for audits -and whether assigning it to the Revenue Agency- only by estimating the elasticity of reported income with respect to real-world enforcement policies and by comparing it to the cost-revenue ratio, where both private ad administrative costs are factored in (for an exact formulation of the latter, see equation 10 in Keen and Slemrod (2017).

Now, the literature on the impact of operational audits so far has not been able to retrieve audit selection criteria. These are particularly important for taxpayers who enjoy greater opportunities to evade, i.e. self-employed and sole proprietorships. These criteria are of interest in themselves because audits, in the absence of third-party information, must be based on some alternative source of information, that, in turn, is used to define risk criteria.

In this paper we use a perfectly balanced panel of 662,241 Italian taxpayers registered for VAT purposes, i.e. deriving their income mostrly from self-employment and sole proprietorship, observed for 5 years.

We do three things.

First, we apply a machine learning model (a random forest) to retrieve the audit criteria that are used by the Italian tax authorities. By doing so, we obtain evidence on the importance of the information accumulated by the tax authority during the 'expiration period', i.e. the time between the period for which a tax declaration is issued and the period when the tax declaration can be

audited. We find that best predictors of the probability to be audited for the tax declaration issued in a given year (the target year) are personal income (PIT) tax bases and VAT turnover values reported not only in the target year but also during the expiration period. This is supportive evidence for the idea that tax authorities act rationally, by making use of all the information available at the time of the audit selection, including that which accrues after the target year that can be legally used. Additionally, we provide evidence that the audit probability is a decreasing function of the profitability rate, i.e. the ratio between PIT taxbase and VAT turnover, as measured in *in the target year* and also of the rate of increase of VAT turnover measured *in the years in between the target year and the audit year*.

Second, we use the audit criteria to identify taxpayers that were not audited but whose characteristics are very similar to audited ones, i.e. we use audit criteria to match audited and unaudited taxpayers. This corresponds to the idea that, by updating the information during the expiration period, the tax authority selects a pool of suspicious tax reports but it cannot audit all of them because of budget constraints. We apply a CEM (*Coarsened Exact Matching*) algorithm and we use the resulting weighting scheme to estimate the average treatment effect on the treated (ATT) both in levels and in elasticity terms. On average, we obtain positive and significant values of ATT, in line with the literature on the impact of operational audits, but almost exclusively with reference to PIT taxbase. The latter result can be interpreted as a rational reponse to the audit criteria: given that taxpyaers are audited because they report a low profitability, the audited taxpayers react by increasing this profitability in the years following the audit.

Third, we plug the estimated enforcement elasticity in the formula for the computation of optimal enforcement elasticity provided by Keen and Slemrod (2017), after adapting it to the piecewise-linear structure of the Italian personal income tax. Then, we perform some back-of-the-envelope calculations to find that, for a plausible range of values of administrative and compliance costs, the number of audits performed by the Italian tax authority is suboptimal.

2 Brief summary of the literature on tax audit criteria and tax audit impact

Although it is common knowledge that cutoffs and risk scores are widely used, the exact formulation of audit criteria followed by tax authorities is unknown (Andreoni et al., 1998) [1] and the literature is scant.

The main exception is the paper by Alm et al. (2004), who examine the process by which firms are selected for a sales tax audit and the determinants of subsequent firm compliance behavior, focusing upon the Gross Receipts Tax in New Mexico. Their purpose is, first, to identify the audit rule and, second, to examine subsequent compliance. The difference between the present paper and Alm et al. (2004) lies in the methodological approach. Alm et al. (2004) use a two-stage selection model, where the first-stage is used to retrieve the audit rule and the second to estimate compliance.

The difficulty with this approach is that the choice of variables to be inserted in the first stage is arbitrary, as it cannot be based on any economic model. As Alm et al. (2004) acknowledge, the audit rule followed by tax authorities is informal and not clearly related to the economic determinants of tax compliance. To put it alternatively, it would be wrong to impose audit criteria that are based on the *normative* theory of tax evasion to estimate the *actual* impact of the enforcement activity.

For this reason, in this paper we use a machine learning approach based on a random forest model to retrieve the audit rule. The advantage of our approach is that it is data-driven, and it exploits all the explanatory variables included in the data. This choice has at least two benefits: first, in the absence of a priori knowledge about the rules determining the audits, this approach makes it possible to use *all* the information contained in the data to summarise the criteria used by the agency. Second, this allows for the identification of additional, not clearly defined, patterns of audit selection.

A paper that uses operational audits is Løyland et al. (2019) who analyze the compliance effect

[1] The formula used by the US to compute the DIF score has traditionally been kept secret Reinganum and Wilde (1988), although taxpayers have somehow learnt what are the most important pieces of information used to compute the DIF. In Italy, the Revenue Agency compute a presumptive value of turnover and taxpayers reporting a lower-than-presumptive turnover know to have a higher probability to be audited under a method called Business Sector Studies (Santoro and Fiorio, 2011). In such a case, the formula for the presumptive value is known, but the exact difference between the probability to be audited if a report is below the presumptive one is unknown. Moreover, BSS is only one of the many audit criteria used by the Revenue Agency

of risk-based tax audits in Norway. They exclude self-employed taxpayers and focus on self-reported deductions among wage earners and transfer recipients as outcome. They find a positive effect of audits on future compliance in terms of a fall in self-reported deductions. However, the response to an audit on self-reported deductions by wage earners can hardly provide a reliable estimate of the general elasticity to changes in implementation policies.

On the contrary, Beer et al. (2019) employ a tax administrative data and operational audit information from a sample of approximately 7,500 self-employed U.S. taxpayers to investigate the effects of operational tax audits on future reporting behavior. They find that reported taxable income is estimated to be 64% higher in the first year after the audit than it would have been in the absence of the audit.

Mazzolini et al. (2021) estimate the impact of operation audits using an approach based on fixed-effects difference-in-difference comparisons with an ex-ante matched sample of non-audited taxpayers. To address concerns about the endogenous selection into audit, they provide evidence for the common trends assumption and find that, on average audited self-employed workers report a subsequent income which is approximately 8.4% higher than the variation recorded by non audited matched taxpayers. To match audited and non-audited taxpayers they use gender, industry, province, age decile and income quartile (in the beginning period, 2007).

In the present paper we use the same dataset analyzed by Mazzolini et al. (2021) and we aim to estimate the impact of operational audits as they do. However, the difference between the present paper and Mazzolini et al. (2021) is twofold.

First, we aim to identify audit criteria, which are not investigated by Mazzolini et al. (2021). To do so, we use an approach that allows us to fully exploit the richness of the panel and, in particular, the time lag between the period for which a tax declaration is issued and the period when the tax declaration is audited. This is known in legal terms as the 'expiration period' and it exists in almost every country. We show that, as it is reasonable to expect, the tax authority uses the information that it accumulates during the expiration period, and that this information consists mainly in the dynamics of reported income during that period as compared to the business cycle.

Second, and consequently, we use audit criteria as our main matching variables, and therefore our estimate of the elasticity can be intepreted as the additional tax base that would emerge by

increasing the number of audits, given the audit criteria.

Another very relevant paper is that by Advani et al. (2022) for the UK. Their very rich dataset allows them to estimate dynamic effects of tax audits, finding that audits raise reported tax liabilities for five years after audit, effects are longer lasting for more stable sources of income, and only individuals found to have made errors respond to audit. Also, and more importantly for our paper, they argue that the aggregate additional revenue after audit is at least 1.5 times the underpayment found at audit, implying substantially more resources should be dedicated to audit than a static comparison would suggest.

3 Institutional background

In Italy, individual taxpayers are required to report their incomes yearly on all personal incomes earned in each tax year. The latter is based on the calendar year. Incomes earned in a given tax year have to be reported between May and September of the following calendar year. For instance, incomes earned between January 1st and December 31st of year t-1 have to be reported between May and September of year t.

After reports are issued, they can be audited.

If a tax report is issued by the taxpayer in due time, AE can audit it until the 31st of December of the fifth year following the year of the tax report. For instance, tax reports referring to year t can be audited within the *ordinary expiration period* , i.e. until the 31st pf December of year $t+6$.

However, if no tax report is issued within due time, the expiration period is increased by two years (so called *extra-ordinary expiration period*), implying that a tax report referring to year t can be audited until the 31st of December of year $t+8$.

According to the AE definition a 'year t' audit is an audit initiated (i.e. for which the audit notice has been sent to the taxpayer) between July 1st of year t-1 and June 30th of year t. Using this notation, the latest possible audit of a tax report referring to year t and issued within due time, i.e the audit conducted during the second-mid of year t+6, is issued in audit year t+7.

In Figure 1, there are two audit years, $t-1$ and t, and two report periods, one referring to period $t-2$ and the other referring to period $t-1$.

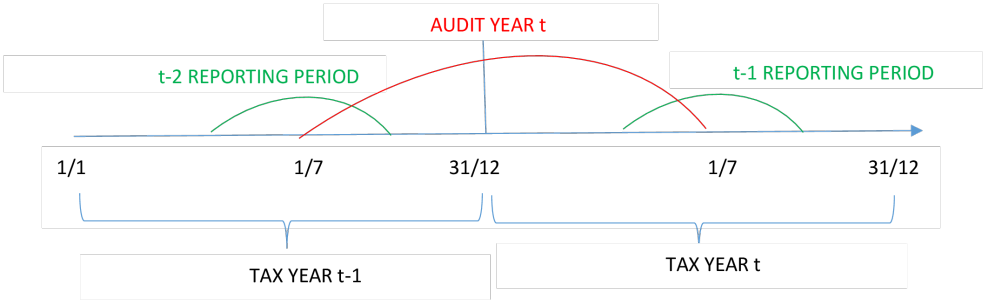


Figure 1: Time frame

By looking at Figure 1 the first year when an audit of year t can be expected to have an impact is year t. Although some accounting registrations and related tax payments (typically those associated to VAT which is paid on monthly bases or every 3 months) for tax year t may already have been made when the audit occurs, the tax report referring to year t is issued between May and September of year t+1, thus surely after a 'year t' audit. This implies that the taxpayer can adjust her accounting and reporting behaviour to react to the audit. Clearly, a neater impact can be recorded in year t+1, which is surely initiated after the audit of year t is concluded.

In general, for every 'year t' audit we distinguish between:

- the *target year*, i.e. the year to which the audited tax reports refers;
- the years between the target year and the audit year, that are part of the expiration period;
- the treatment years, where the audit is expected to have an impact, which include year t and the following ones.

4 Data description

We analyze a perfectly balanced panel of Italian taxpayers using data from two sources, both released by the Italian Tax Authority, AE (*Agenzia delle Entrate* or Revenue Agency). The first dataset contains information from the Tax Return Register "Anagrafe Tributaria", which includes the tax

reports of all Italian taxpayers. The available sample comprises the universe of VAT registered taxpayers with legal residence in three of the most populated Italian regions, namely Lombardy (located in the North), Lazio (located in the Centre) and Sicily (located in the South), which taken together account for around one third of the population of VAT registered taxpayers. These taxpayers usually obtain their income mainly from self-employment and from sole proprietorships.

The tax return dataset contains information on a set of taxpayers' demographic characteristics, like gender, age and place of residence, as well as on the main characteristics of taxpayers' economic activity, like the sector and the number of dependent workers. It includes a range of tax-related variables taken from tax returns, like income type (from self-employment or sole proprietorship), incomes from various sources, personal income tax base, i.e. taxable income equal to revenues minus costs, and VAT turnover, a proxy for the revenues.

The second source of data is the tax audit database.For each audit, the tax audit database contains information on the amount of the preliminary adjustment, the audit year, the target year and the outcome of the audit, distinguishing among null outcome, no taxpayer reaction, settlement, and legal dispute.

The tax return and the tax audit dataset are merged using an encoded taxpayer number (to ensure anonymity) and the report year.

Now, consider the distribution of audits in Table 1, where rows report the 'audit year' and columns report the 'target year' as described in section 3.

Table 1: Distribution of observations across audit years and target years

	target year					
Audit year	2007	2008	2009	2010	2011	Tot
2006	1	0	0	0	0	1
2007	0	1	0	0	0	1
2008	104	0	0	0	0	104
2009	764	54	0	0	0	818
2010	2016	669	30	0	0	2715
2011	4761	3463	387	52	0	8663
2012	10127	6547	1686	554	41	18955
2013	212	9381	4441	2215	536	16785
2014	115	148	7526	2290	1090	11169
Tot	18100	20263	14070	5111	1667	59211

We notice that the tax authority tends to concentrate audits in the second mid of the expiration period and, in particular, in the year before the end of the ordinary expiration period. This is shown by audits of years 2014 and 2013, for which the observation period is almost complete, where the most targeted years are, respectively, 2009 and 2008, i.e $t-5$, for which the tax report has been issued within September $t-4$.

Unfortunately, audits conducted after audit year 2012 cannot be used to estimate the impact of tax audits on subsequent reports because reports from year 2012 onwards are not available to us, and 2012 is the first year for which the impact could be observed.

On the contrary, 2010 audits should have an impact on two observable years, 2010 and 2011, and 2011 should have an impact on 2011 year.

For these reasons, we shall focus the attention on 2010 audits, for which 2007 and 2008 are the target years, 2009 is the post-target and pre-treatment (and also pre audit) year and 2010 and 2011 are the treament years. For robustness checks, we shall use 2011 audits.

The dataset originally includes 460 variables related to 662,241 taxpayers observed for 5 years (total number of observations 3,311,205). In our analysis we use a subset of 42 variables, which are summarized by type in Table 2, and selected removing those variables with 100% of missing values. After that the remaining variables were selected by excluding multicollinear or highly correlated ones (Pearson's coefficient of correlation greater than 0.90).

Table 2: Types of variables used for the statistical analysis

Label	Description
	Time invariant
Sector	Sector of operation (21 dummies)
Region of operation	Lombardy (north) Lazio (centre) Sicily (south)
age	Year of birth of the taxpayer in 2007
female	=1 if the taxpayer is female, 0 otherwise
NW	Number of dependent workers
	Time variant
PIT	Personal income tax variables: revenues, incomes, witholding taxes
VAT	VAT variables: number of positions, turnover
IRAP	IRAP variables: value of production, tax due

Among the time invariant variables the region of residence is relevant considering that including Lombardy (49.4% of observations), Lazio (26% of observations), and Sicily (24.6% of observations),

it allows us to cover North, center, and South of Italy, which are typically different socio-economic contexts. Similarly the large amount of sectors considered means that our results are not strictly conditioned by specific economic sectors but cover a wide range of business activities[2]. PIT is the personal income tax (IRPEF) whose taxbase is personal income which, in turn, is the sum of various incomes, namely income from labour (including self-employment) and income from capital (including that from partnerships). We observe each of these incomes and their single components, along with revenues. We observe also witholding taxes (*ritenute*) applied by counterparts (employer, clients, banks).

VAT is the value added tax (IVA) whose taxbase is the difference between VAT-turnover that we observe, and VAT-costs, that we do not observe. However, we observe the number of VAT-positions associated to the same taxpayer across time. IRAP is a regional taxbase whose base is the value of production, that we observe along with the tax due.

In Tables 3 we report the descriptive statistics for some of the most important variables

Table 3: Summary Statistics

	Mean	SD	Median	1st percentile	99th percentile
Female	0.249	0.432			
Age	47.45	11.84	46	25	78
Number of workers	.823	2.623	0	0	11
VAT Turnover	105776	273561	46749	0	1138767
PIT income	29906	76651	16340	0	224551
IRAP production value	27200	1958998	8053	0	274691

5 Methods

5.1 Random forest approach to identify the audit rule determinants

Tax authorities use risk-based rules to select taxpayers to audit, but these rules are not disclosed. Therefore, our first aim is to identify audit rule determinants and to discuss them. We expect that audit rules are multidimensional and non-linear. For this purpose, we use the random forests

[2]The sectors involved in our analysis are distributed as follow: Trade (retail and wholesale) 26.9%, Professional services 21.8%, Building and construction 11.4%, Agriculture 10.4%, Industry 6.1%, Other services 4.3%, Restaurants and hotels 4.2%, Health services 3.7%, Storing and transport services 2.6%, Services for firms 2.5%, and a 6.1% of other residual sectors.

(RF) classifier , a machine learning method based on ensembles of decorrelated classification trees (Breiman, 1999, 2001; Hastie et al., 2009). The RF is capable to detect interactions and non-linearities in data without superimposing functional forms to the relationship between the covariates and the response variable.

The response variable of the RF is a binary indicator ($Audit_i$) equal to 1 if the taxpayer's income in target years 2007, 2008 or 2009 has been target of a tax audit in audit year 2010. The set of the potential predictors is given by the 42 variables whose classification is reported in Table 2. Compared to other classification methods (i.e. logstic regression) the RF classifier have the advantage to be able to detect complex patterns and non-linearity. A feature of the RF classifier is that returns the variable importance; this measure is based on node purity as measured by the Gini coefficient and reflects the contribution of each variable to improve the model's capability to correctly identify audits. In our framework, we aim to classify the audits by identifying its determinants through ranking the importance of each predictor. The variables at the top of this ranking are said to be the audit determinants. The set of the audit determinants (vector $\boldsymbol{X}$) is the main output of interest for our RF classifier as it gives information on the variables that the are used by the revenue agency to define the (unobserved) audit rules. The RF is characterized by a set of hyperparameters, which must be set appropriately by to maximize the predictive capability of the RF (Claesen and De Moor, 2015). The combination of the hyperparameters values gives a total of 70 random forests, the details of hyperparameter tuning are given in Section 9.1.

The flexibility of the RF has the drawback that the results are not directly interpretable. Therefore, for the sake of ease of interpretation, we provide partial dependence plots. These plots are similar in vein to marginal effects for regression as they convey information about the relationship between the probability of being audited $\Pr(Audit_i = 1)$ audit determinants $\boldsymbol{X}$. Partial dependence plots are the level curves of $\Pr(Audit_i = 1)$ as a function of $\boldsymbol{X}$. In our framework, we consider a pair of audit determinants, $\boldsymbol{x_1} \in \boldsymbol{X}$ and $\boldsymbol{x_2} \in \boldsymbol{X}$. The partial dependence plots are obtained as follows:

1. assign $\boldsymbol{x_1} = x_{1i}$ and $\boldsymbol{x_2} = x_{2i}$ to all the taxpayers in the sample;

2. For each taxpayer j, compute the audit probability conditioned on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. This is $p_j =$

$\Pr(Audit_j = 1 | \boldsymbol{x}_1 = x_{1i}, \boldsymbol{x}_2 = x_{2i})$;

3. Calculated the average conditional probability $\bar{p}$;

4. plot the level curves of $\bar{p}$ with respect to (x_{1i}, x_{2i}).

The process above must be repeated for all the levels of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. We use the random forest that maximizes the predictive accuracy to obtain the audit probabilities.

5.2 Using audit determinants to estimate the impact of audits

The impact of audits on taxpayers' behavior can be measured by looking at subsequent changes either in tax bases or in tax liabilities. As the latter are not observable to us, we look at variations in *PIT taxbase*, the personal income tax base, and in *VAT turnover*, the value of turnover relevant for the value added tax, as our outcomes of interest. For this purpose we apply a difference-in-differences (DID) model (Cunningham, 2021) combined with coarsened exact matching (Iacus et al., 2011, 2012).

Our main DID model aims to estimate the impact of audits conducted in audit year 2010. Thus, the treatment variable is $Audit_i$ (see Section 5.1). Our analysis measures the impact of the audit on tax reporting behaviour in two periods (2010 and 2011). Our model specification is reported in Equation 1.

$$y_{it} = \alpha + \sum_{t=2007}^{2011} \beta_t Year_t + \gamma Treated_i + \sum_{t=2007}^{2011} \delta_t Treated_i Year_t + \boldsymbol{\psi x_i} + \epsilon_{it} \tag{1}$$

where $Year_t$ is a set of period-specific dummies and $\boldsymbol{x}$ is a vector of time-invariant control variables (sector-specific and region- specific fixed effects). All the parameters indicated by Greek letters are regression coefficients to be estimated. We clusterize standard errors at a taxpayer level.

The main parameters of interest are δ_{2010} and δ_{2011}, as they represent the average treatment effect on the treated (ATT). Further, coefficients δ_{2007}, δ_{2008} and δ_{2009} are crucial for supporting the parallel trend assumption. This assumption requires that the treated and untreated cohorts had a stable difference in outcomes over time before the treatment occurred (Cunningham, 2021; Angrist and Pischke, 2014).

Therefore, in our model specification, we want to make inference on δ_{2010} and δ_{2011} provided that the parallel trend assumption is satisfied and, hence, δ_{2007}, δ_{2008} and δ_{2009} are not significant.

A reason for which the parallel trend assumption my not hold is related to the non-experimental nature of our data: even though audit rule are not observable, audits are assigned on the basis of observed tax report characteristics rather than being randomly assigned. Thus, to create an estimation sample in which, conditional on observables, treatment assignment is independent of potential outcomes (Angrist and Pischke, 2014), a conditioning method is required. For this purpose, we use coarsened exact matching (CEM) and use the determinants of audit assignment as matching variables; these variables are $\boldsymbol{X}$, the most imporant variables in classifying audited taxpayers from the RF classifier. CEM achieves covariate balance implies that taxpayers have the same levels of audit determinants prior the audit and, thus, have the same probability of being audited. Hence, differences in the post-audit outcomes can be attributed to the effect of the audit itself.

The advantage of CEM compared to propensity score matching (PSM)[3] is that it does not require any functional form as it creates strata by dividing numerical variables into discrete intervals (i.e. coarsening), whereas each class of categorical variables represents a stratum. Further, CEM is highly effective in removing imbalance between treatment and control groups (Iacus et al., 2011, 2012; Berta et al., 2017), which is a recurrent issue with stochastic matching methods. As a consequence, matched units are in the area of common support by construction.

Each observation is classified according to the combinations of the strata. For each stratum, the CEM calculates taxpayer-specific weights w_i. The weight of observation i in stratum s is defined as follows:

$$w_{is} = \begin{cases} 1 & \text{if } i \in \tau \\ \frac{N^U N_s^T}{N^T N_s^U} & \text{otherwise} \end{cases} \tag{2}$$

In the bottom case of Equation (2), N^U and N_s^U refer to the number of untreated taxpayers in the whole sample and in stratum s respectively. Similarly, N_s^T and N^T are the number of audited taxpayers in stratum s and in the whole sample. Unmatched units are outside of the area of common support and receive zero weight. These weights are used to estimate the regression in Equation (1).

[3]For an extended description of the risks in using PSM in real data applications we refer the reader to the seminal paper by King and Nielsen (2019).

In specifying the CEM model, our goal is to maximize the external validity by matching as many treated taxpayers as possible while ensuring balance in terms of treatment determinants. This practically translates in maximizing the share of audited taxpayers subject to the constraint that CEM produces insignificant differences in the treatment determinants.

6 Results

6.1 The audit rule

Our RF classifiers (70 in total) are characterized by a correct classification rate ranging from 54% to 73%, depending on the hyperparmeters . Sensitivity (the share of correctly classified audited taxpayer, N=2,324) ranges from 62% to 73%, while specificity varies from 54% to 72% (the share of correctly classified non-audited taxpayer). Even though accuracy measure vary depending on the hyperparameters, the RF classifiers consistently signal that the most important variables for audit assignment are PIT taxbase and VAT turnover related to target years 2007,2008 and 2009. These are the pre-audit values of the outcomes considered in our causal inference and are considered as the audit determinants $\boldsymbol{X}$ (see Section rf.methods). Table 4 reports the most import variables and their median ranking obtained from the 70 RF classifiers.

Table 4: Audit rule determinants

Median Ranking	Variable
1	2007 VAT turnover
2	2008 VAT turnover
3	2009 VAT turnover
4	2007 PIT tax base
5	2008 PIT tax base
6	2008 PIT tax base
7	Age
8	2007 IRAP value of production
9	2007 IRAP tax
10	2007 Immovable property income

From the audit assignment determinants identified above we use the partial dependence plots (Figure 2) to provide a graphical understanding of the audit rule.Consistent with the notation of

Section 5.1 we assign $\boldsymbol{x_1}$ = *PIT taxbase 2009* and $\boldsymbol{x_2}$ = *VAT turnover 2009*. The relationship between $\boldsymbol{x_1}$,$\boldsymbol{x_2}$ and audit probability is plotted at the subsample identified by the medians of *PIT taxbase 2007*, *PIT taxbase 2008*, *VAT turnover 2007* and *VAT turnover 2008*. The audit probabilities are obtained from a RF classifier characterized by correct classification rate equal to 72% , sensitivity equal to 66% and specificity equal to 72% . The hyperparameters are $m = 16$ sampled covariates for each tree, $B = 500$ decision trees; the oversampling of the audited class corresponds to a proportion of non audited to audited equal to 2/3. Alternative specifications of the RF classifers are robust to different hyperparameters and are discussed in the Appendix (Section 9.1). Therefore, each panel in Figure 2 can be interpreted as a cell of a 4 x 4 matrix. The first and the second column identify taxpayers with 2007 PIT taxbase below and above the median, respectively, and the third and fourth columns have the same interpretation with respect to the values of 2008 PIT taxbase. On the other hand, the first and the second row identify taxpayers with 2007 VAT turnover values above and below the median, respectively, and the third and fourth raw have the same interpretation with respect to values of 2008 VAT turnover. Therefore, the location of a taxpayer in a given cell describes her tax reports in the target years.

For instance, Panel 11 (forth row, first column) depicts the levels of audit probability as a function of 2009 VAT turnover and PIT taxbase for a taxpayer having 2008 VAT turnover above the median and 2007 PIT taxbase below the median.

Inside cells, the color palette transition from blue to red indicates the probability of audit as a function of the information available to the tax authority in the years following the audited one, i.e. the value of PIT taxbase in 2009 (on the horizontal axis) and the value of VAT turnover in 2009 (on the vertical axis). Thus, by looking at horizontal and vertical changes of colour one can appreciate how the audit probability, for a given subgroup, changes as a function of value reported in 2009.

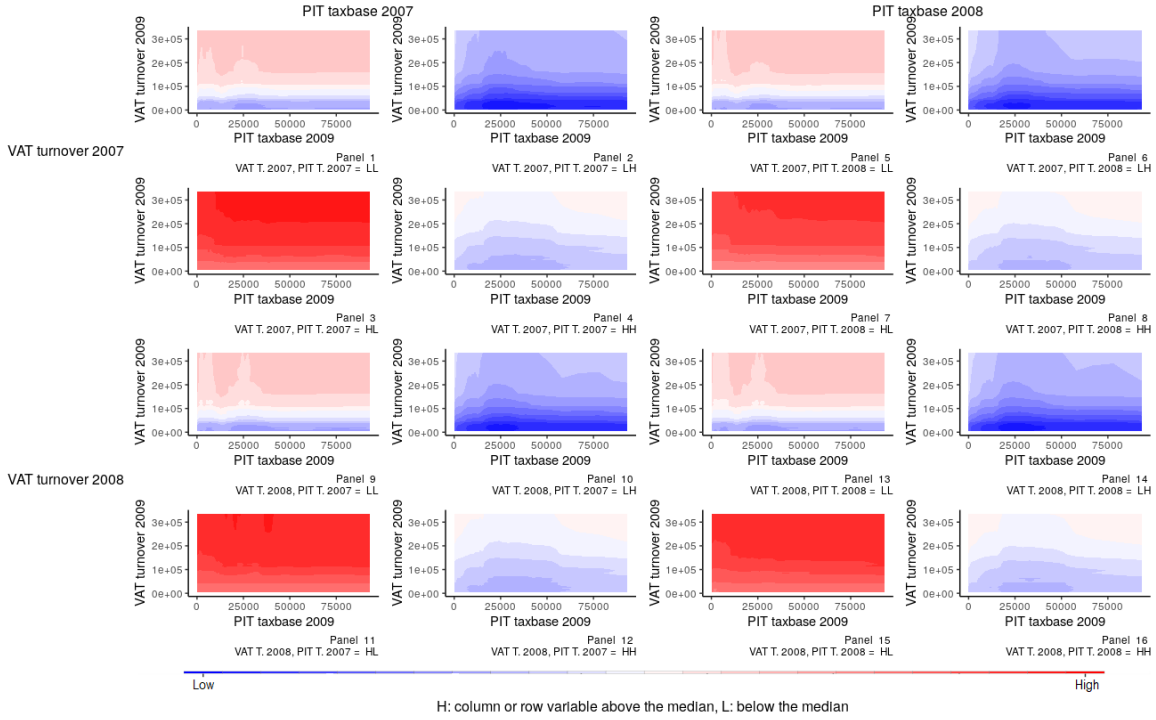


Figure 2: Partial dependence plot for the RF

Visual inspection of Figure 2 reveals some key ingredients of the audit rule followed by the tax authority.

First, if one further divides the 4x4 matrix in 4 submatrices - e.g the upper left submatrix would be the one including cells 1,1, 1,2, 2,1 and 2,2 - , it is immediate to note that they display a very similar pattern. This suggests a stability of behaviour of taxpayers: if a taxpayer reports a value higher (lower) than the median of the relevant variable in 2007 she tends to do it also in 2008 and, consequently, the probability to be audited in 2010 is the same because the taxpayers are the same.

This allows us to focus on any submatrix, for example the first submatrix, where the targeted year is 2007 for both PIT taxbase and VAT turnover.

We observe that:

- the upper left cell - 1,1 in the first quadrant - i.e the *low VAT and low PIT values in the audited year* is characterized by a vertical change from blue to pink;

- the upper right cell - 1,2 in the first quadrant- i.e. the *low VAT and high PIT values in the audited year* is dominated by the blue colour, with a slight change from deep to light blue;
- the lower left cell -2,1 in the first quadrant- i.e the *high VAT and low PIT values in the audited year* is dominated by the red colour;
- the lower right cell - 2,2 in the first quadrant - i.e. the *high VAT and high VAT values in the audited year* is characterized by a vertical change from light blue to almost pink colour.

These patterns can be interpreted as follows.

The first risk criterion that the tax authority seems to have in mind is the presence of a low PIT taxbase, i.e of a low income. As a matter of fact, cells corresponding to lower than median values of the PIT taxbase, i.e cells belonging to the first or third columns, are entirely or partly red coloured. More precisely, a lower than median PIT taxbase is seen as suspect either i) when it is coupled with a high VAT turnover in the target year as shown by the lower-left cells or ii) when the VAT turnover is low in the target year but is increased afterwards as shown by the upper-left cells in each submatrix.

The type of behaviour that the tax authority is seemingly targeting is that of a taxpayer who reports, either in the target years or afterwards, a high VAT turnover and a low PIT income. Given that the former corresponds, approximately, to the revenues declared by the taxpayer while income is the difference between revenues and costs, this behaviour could be summarized as that of a taxpayer inflating costs while reporting revenues correctly.

This interpretation is strengthened when cells of even columns, second and fourth, are examined. In these cells, taxpayers reporting high PIT taxbase are included and the average probability to be audited is lower than that for low PIT reports. Moreover, reporting a low VAT turnover and a high PIT taxbase in the target year is seen as less suspect than reporting both high in the target years; and increasing VAT turnover in years following the target year is seen as suspect.

Why should the tax authority focus so much on evasion accomplished through inflating costs rather than that accomplished through underreporting revenues? The answer may be related to the fact that, during the observed period, taxpayers were subject to *business sector studies, BSS* that provided the taxpayer with a level of presumptive turnover calculated by the tax authority (Santoro

and Fiorio, 2011). On the other hand, no presumption on costs was available for the taxpayer during the period we observe in the data.

This interpretation may explain the fact that reporting a higher VAT turnover in the year following the target year tends to increase the probability to be audited even when the PIT is not low in the target year. The tax authority is seemingly using the information accruing after the target year to audit taxpayers that may have been formally compliant with the presumptive turnover suggested by the BSS but who, at the same time, have over-reported costs. This is an example of a dynamic process of information updating which is broadly in line with that outlined by Advani et al. (2022) as we shall discuss in the concluding section.

6.2 Application of CEM

We use the most important variables from the RF classifier for matching. The set of variables (PIT base and VAT turnover in the pre-treatment period) includes also the pre-treatment outcomes. To this matter, Chabé-Ferret (2015) outlines that the bias of matching is lower when the number of pre-treatment outcomes is higher. Moreover, Lindner and McConnell (2019) outline that matching on pre-treatment outcomes may reduce bias. Our matching strategy resulted in a matched sample of 279,422 out of 658,744 unaudited taxpayers and of 2,085 out of 2,324 treated taxpayers. This means that 90% of the audited and 42% of the unaudited have been restricted to the area of common support. The unmatched untreated units may be non-auditable, while the unmatched audited taxpayers may have been audited because they have extreme values of the audit determinants and, thus, CEM is not able to detect their counterfactual. Matched units are in the area of common support; covariate balance in terms of audit determinants (the most important variables in our RFs) is depicted in Table 5. Prior to applying CEM, the sample was unbalanced in terms of all the variables, in particular the PIT taxbase was significantly lower in the audit population in all of the years prior the audit. Conversely, the same stratum had a significantly larger VAT turnover (more than double). The application to CEM led to two balanced populations with non-significant differences in terms of pre-audit PIT taxbase and VAT turnover (bottom panel of Table 5).

	PIT 2009	PIT 2008	PIT 2007	VAT 2009	VAT 2008	VAT 2007
Pre-Matching	*	***	***	***	***	***
Unaudited	29,308	30,195	30,100	102,962	109,785	105,990
Audited	26,270	25,926	25,299	216,062	241,304	234,918
Post-Matching						
Unaudited	23,453	23,236	23,211	114,636	121,506	117,769
Audited	22,792	23,080	22,258	114,351	123,025	117,746

*** p<.01, ** p<.05, * p<.1

Table 5: CEM results for 2010 audits

6.3 Parallel trend checks and ATT estimates

We now estimate Equation 1 using as dependent variables the amount of PIT taxbases and VAT turnover observed for treated and matched untreated taxpayers in years 2010 and 2011. Recall that the treatment variable here is the 2010 audit, so that 2010 is the first tax report after the audit and 2011 is the second one.

Along with ATT estimates of *absolute differences*, we also provide ATT estimates of *semi-elasticities*, i.e. of the differences between logs of PIT taxbases and of VAT turnover reported by treated taxpayers with respect to those reported by matched untreated taxpayers in each of the two years[4].

Table 6 reports the coefficients of interest from Equation 1 ($\delta_t, t \in \{2007, ..., 2010\}$), while Figure 3 gives a graphical display of the same estimates.

[4]Semi-elasticities are obtained from a log-linear model estimated using a Poisson Pseudo Maximum Likelihood (PPML) estimator. Contrarily to log transforming the dependent variable and estimating an OLS regression, this approach is robust to the presence of zeros in the dependent variable.

	(PIT)	(VAT)	(PIT semiel.)	(VAT semiel.)
Pre-Audit				
$Audit \times Year_{2007}$	-953.103	-22.260	-0.042	-0.000
	(1097.367)	(4263.237)	(0.049)	(0.036)
$Audit \times Year_{2008}$	796.616	1540.592	0.035	0.013
	(561.492)	(1583.326)	(0.024)	(0.013)
$Audit \times Year_{2009}$	291.799	-262.583	0.013	-0.002
	(966.980)	(2074.515)	(0.042)	(0.018)
Post-Audit				
$Audit \times Year_{2010}$	2951.688	5032.081	0.125	0.043
	(1020.044)	(2941.297)	(0.042)	(0.025)
	***	*	***	*
$Audit \times Year_{2011}$	3805.904	8393.066	0.158	0.072
	(1147.011)	(3619.088)	(0.045)	(0.030)
	***	***	***	***

*** p<.01, ** p<.05, * p<.1
Clustered std. err in parentheses
Estimates adjusted for region and sector

Table 6: ATT estimates

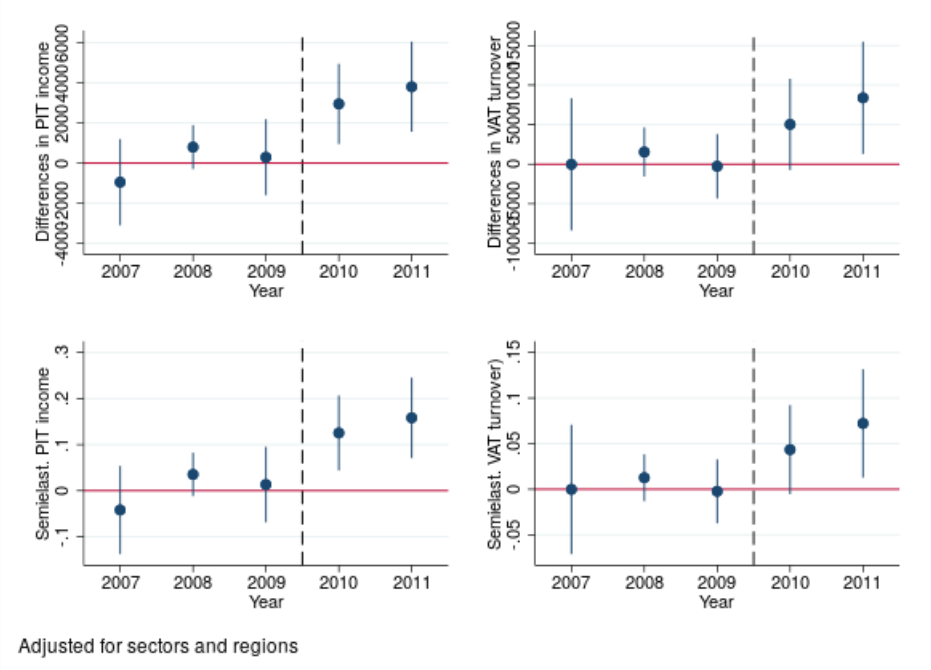


Figure 3: Event study DID for 2010 audits

In the years prior the treatment (top panel of Table 6), our matched DID model did not present significant differences on the pre-treatment outcomes, which are also the variables used by the tax authority to select the audited. Thus, the pre-audit time window is from 2007 to 2009, while 2010 and 2011 are the post-audit time periods, and the parallel trend condition is satisfied.

When we consider *asbolute differences* and *semi-elasticities* the qualitative results are very similar. The audits have significant impact on PIT taxbases in both years, but stronger in 2011 (the target year following the audit year), while they have a weaker, albeit significant impact on VAT turnover only in the 2011. Thus, taxpayers react more by increasing PIT taxbases than by increasing VAT turnover and more in the year following the audit.

The first result suggests a rational response by taxpayers to the audit criteria. Recall from 2 that a high probability to be audited is associated to a low reported value of the PIT taxbase, either associated to a low VAT turnover or associated to a high VAT turnover which is also high in the years following the audited one. If one imagines that, during the audit, taxpayers understand that the PIT taxbase is particularly important as an audit criterion, it is reasonable they increase PIT taxbases without increasing VAT turnover. Note also that this is supportive of the idea of a target rather than a bomb-crater effect so that estimates of the audit impact based on random audits would underestimate the impact of the audit if taxpayers are aware that they have been chosen randomly and not on the basis of a risk assessment.

The second result also fits with the accounting procedures. Although we noted before that an audit conducted in 2010 is surely initiated before a tax report for year 2010 is issued, it should be recalled that a 2010 audit could be conducted during the first six months of 2010, thus when some accounting operations have already been recorded. Now, if the audited taxpayer was prone to evade again before receiving the audit (s)he probably underestimated profitability also during the months of 2010 from January to that when the audit actually occurred. Thus, the rational response (the increase in profitability) could only be effective in the remaining months of 2010, and this is a possible explanation of the lower magnitude of such a response with respect to that recorded in 2011.

6.4 Robustness check

As a robustness check, we repeat the same approach to 2011 audits. In this framework, pre-audit years are the three years before the audit period (2008 to 2010), while the only post-audit year is 2011. As expected, the CEM improved balance of the audit determinants as the audited and non-audited matched population do not statistically differ in terms of pre-audit PIT taxbase and VAT turnover (Table 7). However, our CEM for 2011 audits is capable to match 3'220 out of 7'604 treated taxpayers (42%) e 60'591 out of 653'464 non-audited taxpayers, with a matching performance inferior to that recorded in our 2010 estimate.

	PIT 2010	PIT 2009	PIT 2008	VAT 2010	VAT 2009	VAT 2008
Pre-Matching	***	***		***	***	***
Unaudited	29,878	29,279	30,169	103,715	102,564	109,330
Audited	33,303	30,907	31,116	176,949	171,797	189,030
Post-Matching						
Unaudited	13,012	12,969	12,358	54,967	57,101	59,305
Audited	12,991	12,824	12,288	54,931	57,133	59,396

*** p<.001, ** p<.01, * p<.05, # p<.1

Table 7: CEM results for 2011 audits

Under this caveat, it is worth noticing that the results, displayed in Table 8, are qualitatively similar to thos obtained previously. Again the parallel trend fro 2011 is satisfied and 2011 audits have a clear impact on PIT taxbase, both in terms of absolute values and semielasticity, while there is no significant impact on VAT turnover.

	(PIT)	(VAT)	(PIT semiel.)	(VAT semiel.)
Pre-Audit				
$Audit \times Year_{2008}$	-69.062	90.959	-0.006	0.002
	(262.568)	(1139.834)	(0.021)	(0.019)
$Audit \times Year_{2009}$	-75.902	-58.725	-0.006	-0.001
	(192.025)	(486.939)	(0.015)	(0.008)
$Audit \times Year_{2010}$	47.460	-127.248	0.004	-0.002
	(168.434)	(488.090)	(0.013)	(0.009)
Post-Audit				
$Audit \times Year_{2011}$	992.097	-69.542	0.071	-0.001
	(245.412)	(767.457)	(0.018)	(0.014)
	***		***	

*** p<.001, ** p<.05, * p<.1
std. error in parentheses

Table 8: Robustness check: ATT for 2011 audits

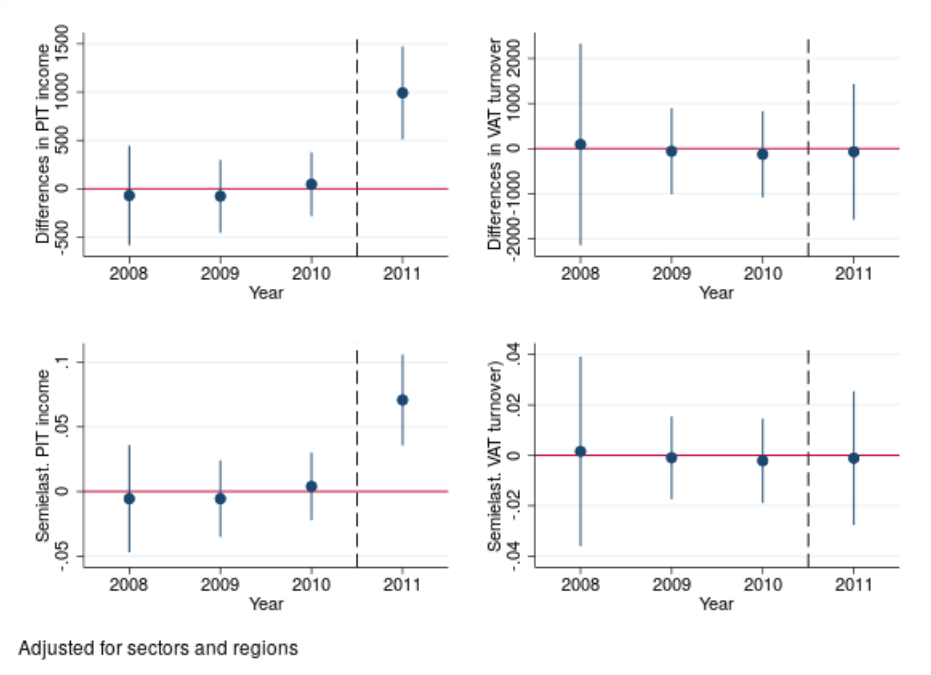


Figure 4: Event study DID for 2011 audits

In sum, results indicate that operational audits, similarly to experimental audits, do have a positive impact on taxpayer's compliance in the years immediately following the audit. This suggests some further considerations on the optimal level of auditing.

7 Some back-of-the-envelope calculations

The literature on optimal tax administration was started by Masyhar (1991) where an implicit condition for the optimal size of tax administration is that the additional revenue gained from stricter enforcement is equated to the sum associated additional compliance and administrative costs, with the latter weighted more heavily than the former because they need to be paid for from distorting taxation.

A major merit of Keen and Slemrod (2017) was to recognise that this condition could be expressed in terms of elasticity of tax revenues, analogously to what happens in the optimal taxation literature. In particular, Keen and Slemrod (2017) show that when a piecewise-linear tax schedule is applied and when both the compliance and administrative costs are homogeneous functions of the level of the administrative activity, denoted by α, then the optimal level of α is given by

$$E(T, \alpha) = \frac{A}{T} + \frac{C}{v'T} \tag{3}$$

where $E(T, \alpha)$ is the elasticity of the tax revenue, T, with respect to α, C are private compliance cost, A are public administrative costs and v' is the marginal social utility of an additional Euro of public spending. Basically, equation (3) is simply saying that the elasticity of revenues should be equal to the the weighted sum of average administrative and compliance costs.

On the left-hand side, $E(T, \alpha)$ is the sum of a *direct effect semi-elasticity*, i.e. the percentage increase in additional taxes collected with the audit, and of an *indirect effect semi-elasticity*, i.e. the percentage increase in additional taxes reported by the audited taxpayer after the audit. Denoting the former with E_{de} and the latter with E_{inde} we can write

$$E(T, \alpha) = E_{de} + E_{inde} \tag{4}$$

The direct effect elasticity can be computed using data on adjusted taxbases and adjusted taxes, including sanctions. The indirect effect is more easily estimated looking at the taxbase semi-elasticity, as we did in Section 6.3 and then computing the associated elasticity of the tax revenue. These two elasticities are linked as follows

$$E_{inde} = E(z, \alpha)\frac{T_z}{\overline{T}} \tag{5}$$

where T_z and $\overline{T}$ are the marginal and average tax rate, respectively. Note that, with a flat tax schedule, $T_z = \overline{T}$ while, with a (weakly) progressive tax schedule, $T_z \geq \overline{T}$, so that the ratio $\frac{T_z}{\overline{T}}$ can be interpreted as a progressivity multiplier.

We now turn to some back-of-the envelope calculations of these semi-elasticities.

As for the direct effect, on average, every Euro of preliminary adjustment generates 10.2 cents of additional taxes in 2010 audits and 11.01 cents of additional taxes in 2011; therefore in our data we can assume $E_{de} = 10,6\%$, very close to the estimate of 10% provided for the US by Alm (2012) [5].

According to the Italian personal income tax schedule (IRPEF), the (legal) marginal tax rates applicable in 2010 and in 2011 were the following: 23% for incomes between 0 and 15,000 Euro, 27% for incomes between 15,000 and 28,000 euros, 38% for incomes between 28,000 and 55,000 Euro, 41% for incomes between 55,000 and 75,000 Euro and 43% for incomes above 75,000. Average rates vary accordingly.

The distribution of PIT taxbases reported by the 1,372 taxpayers audited in 2010 and matched using the CEM algorithm across vingtiles is reported in the Appendix (see Tables 10 and 10). Using these data, a pair of plausible values is $T_z = 26.7\%$ and $\overline{T} = 25,1\%$ so that $\frac{T_z}{\overline{T}} = 1,06$. In other words, despite the fact that the Italian PIT schedule has 5 brackets, the distribution of reported incomes is so skewed on the left that the actual tax system is very mildly progressive. A more progressive schedule at the bottom, or a more even distribution of reported taxbase would yield a much higher value of the progressivity multiplier.

Recalling that the minimum estimate of the indirect semielasticity that we obtain is 7%, the minimum value of the LHS of (3) consistent with our estimates is $7\%X1,06 + 10.6\% = 18,02\%$.

Turning to the RHS, we know that the administrative cost of the Italian revenue agency in Italy amounts to approximately 3 billions of Euro per year, while the total amount of (State) taxes amounts to approximately 450 billions, so that $\frac{a}{T} = 0,67\%$ a value very close to that reported by

[5] Advani et al. (2022), using random audits, obtain an indirect effect which is 1.5 larger than the direct one. The difference between their results and ours may be due not only to the nature of audits (random in their paper, operational in ours) but also to the nature of incomes audited, as they find heterogeneous responses across income types.

Keen and Slemrod (2017) for the US.

We do not have information about the private compliance costs in Italy, so that we can take the US case as a benchmark. In US, private compliance costs are estimated at 11% so that $\frac{c}{T} = 0.11$, thus compliance costs are about 16 times larger than administrative costs. In Italy the tax system is fairly complex, so that it might be reasonable to assume a higher bound of $\frac{c}{T} = 0.15$. If, $v' = 1, 2$, we can assume $\frac{c}{T} = 0.125$, so that the maximum value of the RHS of (3) consistent with our estimates is $0,7\% + 12.5\% = 13,25\%$.

This line of reasoning suggests the observed level of audits is not optimal. If we assume that both the direct and the indirect effects and elasticities are decreasing in the number of audits, because increasing audits imply to audit also less risky taxpayers, then we can conclude that our back-of-the-envelope calculations suggest that the observed level of audits is suboptimal.

8 Concluding remarks

The contribution of this paper to the literature is twofold.

First, we show that machine learning techniques are useful for the analysis and the design of tax administration policies, at least to the same extent they have proven to be useful for the design of public policies aimed at preventing crime (see Chandler et al. (2011), Kleinberg et al. (2018) and Meijer and Wessels (2019)) or at improving the design of welfare services (see Rockoff et al. (2011) and Andini et al. (2018); Finkelstein and Notowidigdo (2019)). Clearly, the idea of applying machine learning to tax administration is not completely new, as tax authorities around the world are now well aware of the advantages that can potentially be associated to a better targeting of their audit policies(OECD, 2017). However, in the current debate it is not clear how this can actually be done.

The present paper shows that machine learning can be used to analyze the current audit rule adopted within a tax administration, to evaluate its impact and, in turn, to calculate the optimal level of audits. A limitation of our dataset is that we observe only a short period after the audit. As shown by (Advani et al., 2022), considering a longer period will likely yield even higher estimates of the indirect effect and, consequently, of the optimal number of audits. Clearly, one could also use machine learning to compare different audit rules or to identify the best ones. To do the latter,

however, machine learning should be used to retrieve risk criteria from observed behaviour in random audits that measure true propensity to evade. Then, these criteria should be used to design a new audit rule and its direct and indirect effects should be measured, for example by randomly selecting risky taxpayers to be audited. In the latter case, the evaluation of impact could be done by a standard difference-in-difference models where unaudited taxpayers provide the counterfactual, and the matching approach we adopted here would not be necessary.

Second, we contribute to the literature which analyzes the impact of operational tax audits by looking at the information used by the tax authority and also at how this information becomes available to the audited taxpayer. In particular, (Advani et al., 2022) explain their results by making appeal (also) to the updates of the information held by the tax authority. They suggest that performing an audit provides the tax authority with more accurate information on a taxpayer's income at a point in time. In subsequent years, information from the audit will make evasion of more stable income sources easier to detect, but for less stable income sources the effect will rapidly wear off. Thus, taxpayers who have stable incomes will learn from the audit that their incomes are easier to detect and will consequently evade less.

We complement this finding by suggesting that the tax authorities also consistently use, when they select taxpayers to audit, the information they have gathered during the 'expiration period' i.e. the period in between the audited year and the year of the audit. In particular, we provide suggestive evidence for the Italian tax authority looking at how much and in what directions two basic indicators - PIT taxbase and VAT turnover - have changed during the expiration period. Also, we observe that the taxpayer reacts to the audit by increasing the reported PIT taxbase, consistently with the idea that, during the audit, the taxpayer has perceived this to be the main indicator driving the audit selection.

9 Appendix

9.1 Random forest hyperparameter tuning

The RF is a classifier defined by a collection of tree-structured classifiers, where each tree splits nodes using a randomly selected set of taxpayers' characteristics. Assuming that B is the number

of trees we want to estimate in each RF, let m the subset of variables to be selected at every split in each tree, and n the minimum node size. The detailed procedure is the following:

- For b=1 to B

    1. Draw a bootstrap sample of observations with replacement from the training dataset
    2. Fit a classification tree on the bootstrapped data. Nodes are split iteratively by:
        (a) from the p variable randomly select a m -dimensional subset of taxpayers' characteristics;
        (b) select the variables and the cut-off points that maximize the node purity (measured by the Gini index);
        (c) split each node into two daughter nodes.

- Obtain the RF class prediction by aggregating the prediction of each tree. An observation is assigned to a class (audited or non audited) based on the majority of "votes" defined by each tree.

The advantage in the random selection of splitting candidate covariates is related to variance reduction through the introduction of node splits based on variables and criteria that otherwise would be overlooked (Hastie et al., 2009; Breiman, 1999, 2001). As node splitting is based on node purity, that depends on the prevalence of the class to be predicted, the rarity of the audits would result in very high specificity (i.e. capability to correctly predict the non-audited class) but low sensitivity (i.e. capability to correctly predict the audited class).

However, our main purpose is the prediction of audited taxpayers. To tackle this issue we under-sample the most frequent class (the non-audited) (Chen et al., 2004; Weiss et al., 2007). We use the non audited-to-audited sampling ratio as an additional RF hyperparameter. This is done by fixing the number of sampled audits equal to the number of audits in our dataset and adjusting the number of sampled non-audited observations; the selected non audited-to-audited ratios are 1:2 and 2:3. By forcing the algorithm to sample a smaller number of non-audited, we train our RF to predict the audits more accurately (i.e. increasing sensitivity) at the cost of reducing the specificity.

Besides the sampling scheme, the RF requires the specification of B, the number of decision trees to be fit and m, the number of variables to be sampled as candidates to each split. We decided to

perform RF including $B = (50, 100, 200, 500, 1000)$ different trees. Concerning hyperparameters m, Hastie et al. (2009) suggest using $m = \log_2 p+1$ as reference value, which means, in our case, $p = 134$ and $m = 8$. Starting from this reference, we decided to test a vector of $m = (4, 6, 8, 10, 12, 14, 16)$. The combination of all the above mentioned hyperparameters yields a total of 70 random forests from which we extract the out-of-bag (OOB) accuracy measures and the variable importance in terms of reduction in the Gini impurity index. We use the set of the most important variables as inputs in a matching approach explained in Section 5.2.

9.2 Audit rule determinants

The results of the OOB accuracy for our RF classifiers can be observed in Figure 5 and summarized in Table 9. Figure 5 reports that the majority of the fitted RF is clustered in the False Positive rate range 0.25-0.40, while the sensitivity (True Positive rate) is clustered around 0.70. Two opposite classes of RF are also evident from Figure 5: the bottom-left class (lower sensitivity and higher specificity) and top-right class. The former is given by the 2:3 sampling scheme, the latter is given by the 1:2 sampling scheme and yields to a sensitivity ranging from 75% to 80%. This means that the lower is the non-audited to audited sampling proportion the better is the ability of the RF to correctly predict the audits (in Figure 5 the empty markers have higher sensitivity). Correct classification rate (CCR) of our RFs ranges from 54% and 71% and it is largely determined by sensitivity. In these terms the best performances are given by hyperparameters $m = 16$, $B = 500$ and 2:3 sampling. Such RF is characterized by 66% sensitivity and 72% specificity. In contrast, the 1:2 sampling RF with the largest CCR (66%) is characterized by $m = 14$, $B = 500$, specificity equal to 72% and sensitivity equal to 66% respectively.

Figure 5: Accuracy measures for the RF

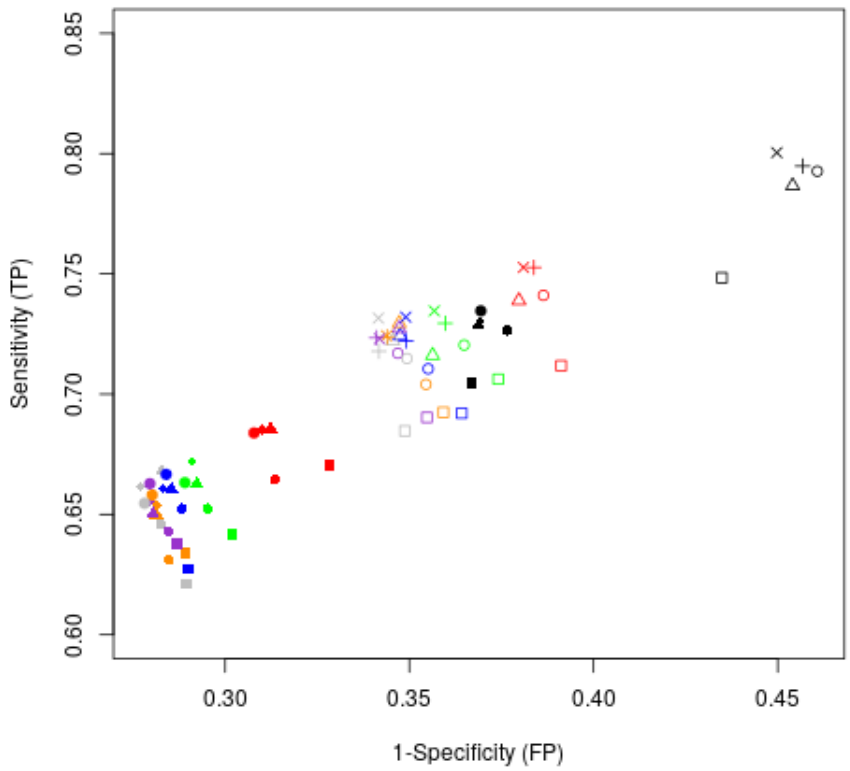
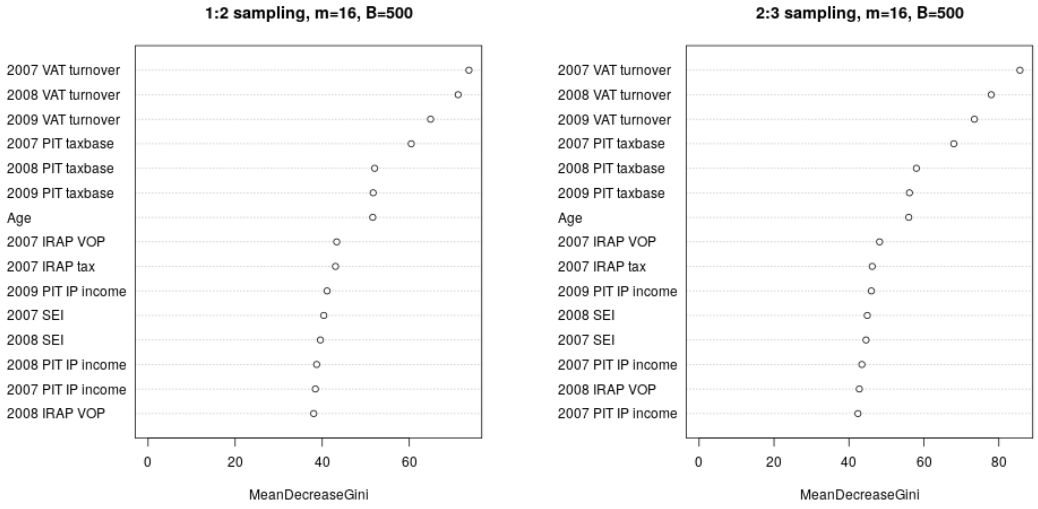


Table 9: Accuracy measures for the RF

Sampling	Measure	Sensitivity	Specificity	CCR
1:2	Min	0.6847	0.5392	0.5401
	Median	0.7239	0.6449	0.6451
	Mean	0.7296	0.6296	0.6300
	Max.	0.8004	. 0.6589	0.6591
2:3	Min.	0.6211	0.6234	0.6238
	Median	0.6606	0.7109	0.7108
	Mean	0.6658	0.6983	0.6981
	Max	0.7346	0.7230	0.7228
Overall	Min	0.6211	0.5392	0.5401
	Median	0.7043	0.6561	0.6564
	Mean	0.6977	0.6639	0.6641
	Max	0.8004	0.7230	0.7228
Sampling: non-audited to audited proportion CCR: correct classification rate				

In Figure 6 we can appreciate the variable importance in our RFs . The left panel represent the most accurate 1:2 sampling RF, whereas the right panel is related to the most accurate 2:3 sampling RF. The first six rows of top and central panels show that the greatest contribution in terms of Gini impurity index are provided by the *PIT taxbase* and *VAT turnover* referring to the target years before the audit (2007, 2008 and 2009). This variables are the audit determinants. This is additional evidence for the idea that the tax authority uses all the information available at the time of the audit. To gain more insights on how this information is used we turn to partial dependence plots.

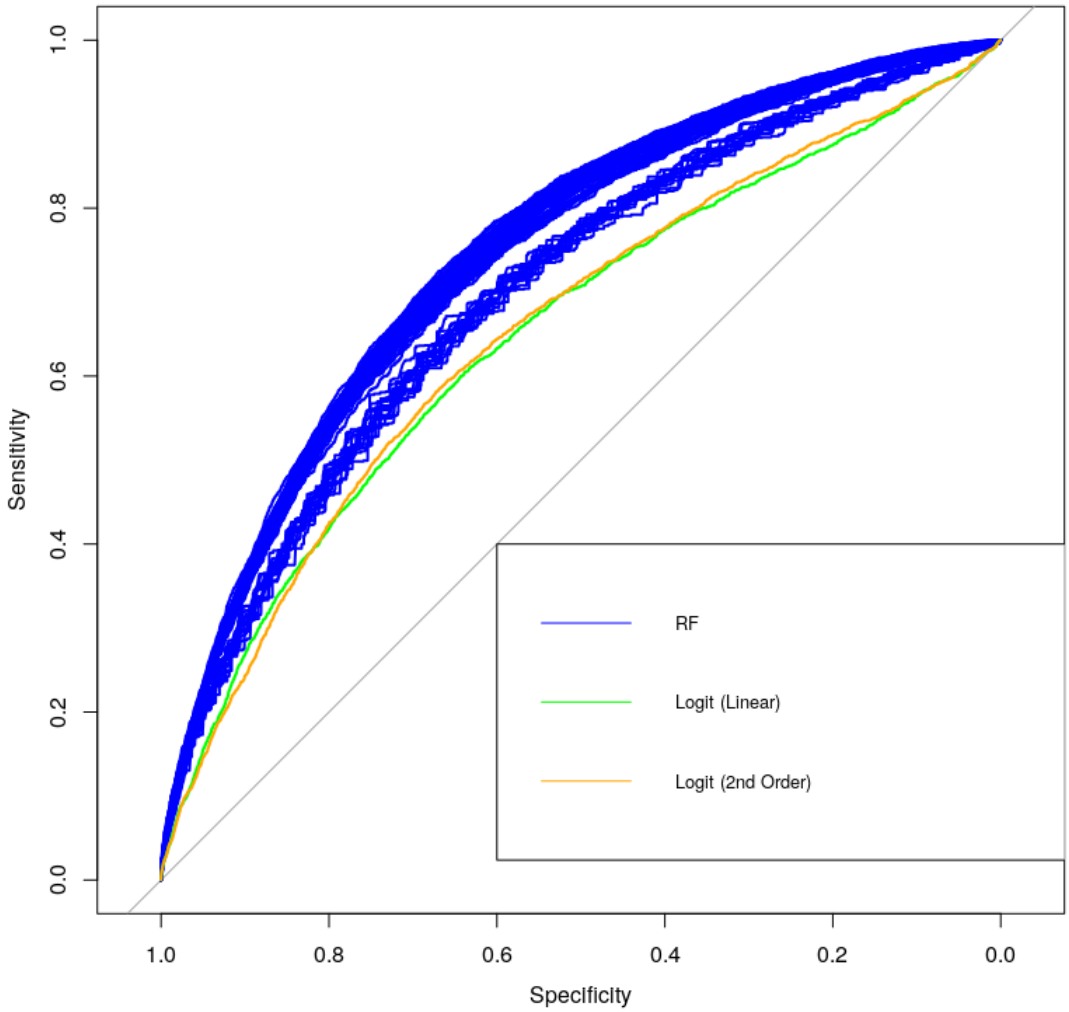


(a) (b)

Figure 6: Summary of variable importance: RF with maximum CCR.

We also compare the predictive accuracy of our RF classifiers with logistic regression using receiver operative characteristic (ROC) curves. Our logistic models include as covariates the most important variables obtained from the RF in two specifications: 1) linear form; 2) a second order polynomial form including all the interactions. All of the RF classifiers outperform the regressions in terms of area under the curve (AUC) (Figure 7). For reference, the polynomial logistic model's AUC is equal to 0.653 , while the same figure is equal to 0.755 the best performing RF. The ROC analysis in Figure 7 also highlights using any threshold (in terms of votes) from the RF classifier to discriminate between audited and non-audited leads to more accurate prediction than any threshold (in terms of predicted probabilities) from the logistic models. This means that, in our framework, the RF is globally more accurate than logistic regression.

Figure 7: Accuracy measures for the RF



9.3 Data for BOE calculation

	Post Audit						*Pre Audit*								
	2011			2010			2009			2008			2007		
vingtile	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1670	730	1	1372	610	39	294	192	6	154	101	1	1	1
4	1677	3870	2864	1426	3678	2597	295	2649	1349	189	2667	1427	2	1535	660
5	3877	5751	4746	3685	5674	4778	2717	4563	3733	2681	4673	3683	1558	3765	2797
6	5807	7475	6750	5679	7232	6464	4576	6296	5502	4685	6632	5729	3774	5748	4788
7	7481	8921	8184	7234	8772	7987	6317	7785	6994	6634	8001	7317	5761	7494	6671
8	8927	10088	9475	8779	10177	9504	7788	9207	8451	8033	9315	8731	7499	8984	8258
9	10110	11457	10797	10178	11257	10668	9212	10532	9904	9328	10467	9925	9006	10219	9589
10	11470	12767	12113	11292	12760	12132	10551	11871	11144	10477	11737	11182	10221	11250	10713
11	12774	14327	13584	12788	14185	13481	11875	13164	12511	11775	13266	12554	11261	12617	12001
12	14336	16224	15226	14193	15741	14918	13187	14698	13873	13283	14846	14003	12622	14253	13356
13	16237	18296	17233	15742	17453	16641	14700	16191	15415	14855	16721	15801	14257	15833	15051
14	18331	21021	19661	17476	20066	18658	16228	18481	17409	16752	18928	17816	15842	17950	16815
15	21040	24064	22713	20071	23875	21853	18513	21716	20179	18938	21958	20451	17954	21162	19525
16	24076	29659	26626	23882	28124	26107	21808	26002	23861	21997	25838	23757	21204	25357	23091
17	29709	37811	33375	28166	36124	31906	26067	33605	29315	26184	32996	29261	25422	32695	28684
18	37910	51245	43536	36141	47979	41014	33853	46145	38884	33114	46886	39201	32854	45883	38813
19	51250	86351	64722	48108	89196	62486	46439	83726	59710	46902	77180	57303	45925	78325	58989
20	86890	2215674	208520	90490	1995983	198785	84223	996150	178093	79250	1270914	184009	80280	993152	175887

Table 10: PIT taxbase

References

Advani, A., W. Elming, and J. Show (2022). The dynamic effects of tax auditsthe dynamic effects of tax audits. *Review of Economics and Statistics*.

Alm, J. (2012). Measuring, explaining, and controlling tax evasion: lessons from theory, experiments, and field studies. *International tax and public finance 19*(1), 54–77.

Alm, J., C. Blackwell, and M. McKee (2004). Audit selection and firm compliance with a broad-based sales tax. *National Tax Journal 57*(2), 209–227.

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

Andreoni, J., B. Erard, and J. Feinstein (1998). Tax compliance. *Journal of Economic Literature 36*(2), 818–60.

Angrist, J. D. and J.-S. Pischke (2014). *Mastering'metrics: The path from cause to effect*. Princeton university press.

Beer, S., M. Kasper, E. Kirchler, and B. Erard (2019). Do Audits Deter or Provoke Future Tax Noncompliance? Evidence on Self-employed Taxpayers? IMF Working Papers, Fiscal Affairs Department 223, International Monetary Fund.

Berta, P., M. Bossi, and S. Verzillo (2017). % cem: a sas macro to perform coarsened exact matching. *Journal of Statistical Computation and Simulation 87*(2), 227–238.

Breiman, L. (1999). Random forests. *UC Berkeley TR567*.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Chabé-Ferret, S. (2015). Analysis of the bias of matching and difference-in-difference under alternative earnings and selection processes. *Journal of Econometrics 185*(1), 110–123.

Chandler, D., S. D. Levitt, and J. A. List (2011). Predicting and preventing shootings among at-risk youth. *American Economic Review 101*(3), 288–92.

Chen, C., A. Liaw, L. Breiman, et al. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley 110*(1-12), 24.

Claesen, M. and B. De Moor (2015). Hyperparameter search in machine learning. *arXiv preprint arXiv:1502.02127*.

Cunningham, S. (2021). *Causal Inference: The Mixtape* (1 ed.). Yale University Press.

Finkelstein, A. and M. J. Notowidigdo (2019). Take-up and targeting: Experimental evidence from snap. *Quarterly Journal of Economics*, 1505 – 1556.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference and prediction* (2 ed.). Springer.

Iacus, S. M., G. King, and G. Porro (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association 106*(493), 345–361.

Iacus, S. M., G. King, and G. Porro (2012). Causal inference without balance checking: Coarsened exact matching. *Political analysis 20*(1), 1–24.

Keen, M. and J. Slemrod (2017). Optimal tax administration. *Journal of Public Economics 152*, 133–142.

King, G. and R. Nielsen (2019). Why propensity scores should not be used for matching. *Political Analysis 27*(4), 435–454.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics 133*(1), 237–293.

Lindner, S. and K. J. McConnell (2019). Difference-in-differences and matching on outcomes: a tale of two unobservables. *Health Services and Outcomes Research Methodology 19*, 127–144.

Løyland, K., O. Raaum, G. Torsvik, and A. Øvrum (2019). Compliance effects of risk-based tax audits.

Masyhar, J. (1991). Taxation with costly administration. *The Scandinavian Journal of Economics 93*(1), 75–88.

Mazzolini, G., L. Pagani, and A. Santoro (2021). The deterrence effect of real-world operational tax audits on self-employed taxpayers: evidence from italy. *International Tax and Public Finance*.

Meijer, A. and M. Wessels (2019). Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration 42*(12), 1031–1039.

OECD (2017). Tax Administration 2017: Comparative Information on OECD and Other Advanced and Emerging Economies. Technical report, OECD Publishing.

Reinganum, J. F. and L. L. Wilde (1988). A note on enforcement uncertainty and taxpayer compliance. *The Quarterly Journal of Economics 103*(4), 793–798.

Rockoff, J. E., B. A. Jacob, T. J. Kane, and D. O. Staiger (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy 6*(1), 43–74.

Santoro, A. and C. V. Fiorio (2011). Taxpayer behavior when audit rules are known: Evidence from italy. *Public Finance Review 39*(1), 103–123.

Slemrod, J. (2016). Tax compliance and enforcement: New research and its policy implications.

Slemrod, J. and C. Weber (2010). Evidence of the invisible: Toward a credibility revolution in the empirical analysis of tax evasion and the informal economy. *International Tax and Public Finance 19*(1), 25–53.

Slemrod, J. and S. Yitzhaki (2002). Tax avoidance, evasion and administration. *The Handbook of Public Economics 3*(22), 1425–65.

Weiss, G. M., K. McCarthy, and B. Zabar (2007). Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? *Dmin 7*(35-41), 24.