No 497

MAY 2022

CefES
Center for European Studies
PAPER SERIES

The Optimal Number of Tax Audits: Evidence from Italy

Daniele Spinelli, Paolo Berta, Alessandro Santoro

DIPARTIMENTO DI ECONOMIA, METODI QUANTITATIVI E STRATEGIA DI IMPRESA

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

The Optimal Number of Tax Audits: Evidence from Italy

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

Tax audits are the main tool adopted by tax administrations to collect taxes. Their optimal number depends on two parameters, i.e. the enforcement elasticity of tax revenue with respect to the audit effort and the sum of private compliance costs and public administrative costs entailed by audits. In turn, the enforcement elasticity critically depends on audit selection criteria actually chosen by tax authorities. In this paper, we apply a machine learning approach to Italian data and we provide evidence that, in 2010 and 2011, audited taxpayers are those whose reporting behaviour in between the report year and the audit year has deviated from the business cycle. We use these audit criteria to match audited taxpayers to non-audited ones and we obtain an estimate of the enforcement elasticity that allows us to characterize the optimal number of tax audits as a function of the ratio between private compliance and public administrative costs.

Keywords: Optimal Tax Administration, Enforcement Elasticity of Tax Revenue.

JEL Numbers: H26; C55

*Corresponding author: Daniele Spinelli - daniele.spinelli@unimib.it, Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, Edificio U7 20126 Milan (Italy)

1 Introduction

The paper by Keen and Slemrod (2017) has set out a framework for analyzing optimal interventions by a tax administration and proposed a summary measure of their impact, the enforcement elasticity of tax revenue. With a flat tax rate, this is defined as the percentage change in reported income associated with a percentage change in the audit rate. It is a very convenient parameter that parallels the elasticity of taxable income, which is the common measure of the response to tax rates. By measuring both of these elasticities, the policy maker can find the optimal balance between the two tax instruments, tax rates and administrative measures, that can raise revenues.

However, the evidence on the magnitude of the enforcement elasticity of tax revenue is not abundant. The literature surveyed by Alm (2012), and quoted by Keen and Slemrod (2017), refers to the Mid Eighties and only looks at the direct effect of tax audits (in the US), i.e. the increase of revenues due to additional tax and penalties resulting from IRS examinations. This effect is limited to audited taxpayers and to audited returns. This direct component yields estimates of the enforcement elasticity that vary in the narrow range of 0.1 to 0.2.

But, as acknowledged by Alm (2012), including the indirect effect of tax audits can change the picture. Such an indirect effect, in turn, can be split in two parts. First, the impact of audits on post-audit returns issued by audited taxpayers, i.e. the impact of audits on subsequent tax compliance. Second, the impact of audits on post-audit returns issued by non-audited taxpayers whose compliance, through various possible channels, is affected by audits. Although the latter, also known as the spillover effect of audits, is conjectured by various studies, the recent literature focuses on the former and this is what we do in this paper.

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 estimates for the enforcement elasticity. 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. 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, 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 for taxpayers that are really risky, 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 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. We find that best predictors of the probability to be audited are personal income tax bases and VAT turnover values reported by audited taxpayers in the years lying between the year of the audited report and the year when the audit actually occurs. In particular, we provide evidence that the audit criteria are based on a comparison between the pattern of these reports and the business cycle.

Second, we use the audit criteria to identify taxpayers that were not audited but whose characteristics are very similar to audited ones. We apply a *Coarsened Exact Matching, CEM* 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.

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 derive the optimal number of audits as a function of the ratio between (private) compliance costs and (public) administrative costs. We find that, for a plausible range of values, the number of audits performed by the Italian tax authority is suboptimal.

This conclusion is broadly in line with that reached by Advani et al. (2022) for the UK. In particular, 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. However, their paper uses revenues from random, rather than operational, audits and looks only at the revenue effect, without making any attempt to evaluate the cost side of the enforcement activity.

2 Relationship with previous literature

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 associated additional compliance and administration 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.

This elasticity is conditional on the audit rules adopted by tax authorities. Although it is well known that cutoffs and risk scores are widely used, the exact formulation of audit rules is unknown (Andreoni et al., 1998) or only partially known [1] , although there are some exceptions.

Alm et al. (2004) 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 drive, and it exploits all the explanatory variables included in the data. This choice has at least two benefits: first, in the absence of a

[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

priori knowledge about the rules determining the audits, this approach makes it possible to use 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 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 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 is applied 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.

3 Data and institutional background

We analyze a perfectly balanced panel of Italian taxpayers using data from two sources, both released by the AE. 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 account for around one third of the entire Italian population. VAT registered taxpayers usually obtain their income mainly from self-employment or 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, gross tax, total amount of tax allowances, net tax.

The second source of data is the tax audit database. For each audit, it contains information on the amount of the preliminary adjustment, the audit 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 tax year (see Table 1 below). To analyze this database some relevant features of the institutional system should be taken into account.

In particular, it is important to distinguish between the *tax report year* and the *tax audit year*.

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. Personal incomes may derive from dependent work, self-employment, sole proprietorship and capital (shares in a partnership or in a corporation).

After reports are issued, they can be audited. The Italian revenue agency

(Agenzia delle Entrate, henceforth AE) can audit tax reports for up to five years (ordinary expiration period) after the end of the calendar year to which the declaration refers. Then, after five years, evasion can no longer be prosecuted unless it is the outcome of a fraud or a criminal act, in which case the expiration period may be longer. Audits generate an audit notice which contains the preliminary tax adjustment claimed by the AE. Note that an audit notice can refer to multiple taxes (for example, to both income and value added taxes), but in this paper we consider only adjustments referring to personal income tax. 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. A ‘year t’ audit overlaps with two tax years (t-1 and t) and with two tax reports (referring to tax years t-2 and t-1, respectively). Note that tax reports referring to year t are issued between May and September of year t+1, thus surely after a ‘year t’ audit.

Now, consider the distribution of audits in Table 1, where rows report the ’audit year’ and columns report the ’report year’.

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

	report_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

On the basis of previous discussion, an audit conducted in audit year 2012 may have an impact only on tax reports for 2011. On the contrary, audits conducted in audit years 2011 and 2010 have a wider potential impact. More precisely, audits conducted in 2010 are surely initiated *before* the tax reports for 2010 and for 2011 are issued, while audits conducted in 2011 are surely initiated before tax reports for 2011 are issued and *possibly* initiated before tax reports for 2010 are conducted. For these reasons, we shall focus the attention on 2010 audits and, for robustness checks, on 2011 audits.

Finally, note that 2010 audits amount at 2715 because they include 'multiple audits', i.e audits on more than one tax report. This difference in the intensity of treatment is not easily captured within our model, so that we focus on 2327 single audits, i.e audits conducted on a single taxpayer and on a single tax report.

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 exlcuding 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	min	max
Age	47.4486	11.839	46	18	107
Female	0.249	0.433	0	0	1
Number of workers	0.823	2.624	0	0	409
PIT tax base	29,933	76,724	16,350	0	25,418,359
VAT turnover	105,656	273,397	46,693	0	69,512,479
IRAP value of production	26,125	72,955	8,049	0	15,433,600

[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.

4 Methods

4.1 Random forest approach to identify the audit rule determinants

With the aim to identify the so-called audit rule we adopt the random forests (RF) approach, a machine learning method based on ensembles of de-correlated classification trees (Breiman, 1999, 2001; Hastie et al., 2009). The response variable of the RF is a binary indicator (Y_i) equal to 1 if the taxpayer's income related to fiscal years 2007, 2008 or 2009 has been target of a tax audit in 2010. We predict this event using, in principle, the 42 variables previously selected and listed by type in Table 2. Howevere, as a maching learningh method, RF is parameterized by a set of hyperparameters, which must be set appropriately by to maximize the usefulness of RF (Claesen and De Moor, 2015). In particular, 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. This random selection of the input variables reduces the correlation between the trees in the RF (Hastie et al., 2009). The detailed procedure is the following:

1. assuming that B is the number of trees we want to estimate, let m the subset of variables to be selected at every split in each tree, and n the minimum node size;

2. For b=1 to B

    (a) Draw a bootstrap sample of observations with replacement from the training dataset

    (b) Fit a classification tree on the bootstrapped data. Nodes are split iteratively by:

        i. from the p variable randomly select a m -dimensional subset of taxpayers' characteristics;

ii. select the variables and the cut-off points that maximize the node purity (measured by the Gini index);

iii. split each node into two daughter nodes.

3. 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 candidates is related to variance reduction through the introduction of node splits based on variables and criteria that otherwise would be overlooked (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 but low sensitivity. However, our main interest is to be able to predict the audited class. 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 hyperparamater 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 to 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 Gini index. We use the set of the most important variables as inputs in a matching approach explained below.

4.2 Identifying the causal effect of audit: average treatment effects and coarsened exact matching

This paper aims at identifying the impact of operational audits on taxpayers behavior. We consider *personal income tax base* related to fiscal years 2010 and 2011 as our outcomes of interest. Let us introduce Y_i^U, the potential outcome for the i-th taxpayer when the treatment is not assigned, and Y_i^T, the potential outcome for the same taxpayer receiving treatment. In addition, let τ be the set of the treated units. For each unit, only one between Y_i^T and Y_i^U can be observed based on the treatment assignment. The main quantity of interest becomes the average treatment effect on the treated (ATT, Equation 1).

$$ATT = \frac{1}{N_\tau} \sum_{i \in \tau} (Y_i^T - Y_i^U) \tag{1}$$

Considering the treatment units, Y_i^U is unobserved, and, on the framework of potential outcome, it must be estimated. In this case, the basic idea is to replace Y_i^U with a set of control units selected from the population of the untreated (i.e. those taxpayers for which only Y_i^U is observed) on the basis of observable pre-treatment characteristics $\boldsymbol{X}$ that affect both the outcome and the treatment assignment[3]. In this way the conditional independence assumption (CIA) is satisfied, meanning that treatment is conditionally

[3] In our estimations, Y is specified as the absolute value or as the logarithm of PIT taxbase and of VAT turnover.

independent from the observables $\boldsymbol{X}$. Such set of variables is obtained by the importance plot given by the RF: they are those used by the revenue agency to assign audits to taxpayers. The chosen conditioning strategy is based on the coarsened exact matching (CEM, Iacus et al. (2011)). This matching algorithm is based on creating strata by dividing numerical variables into discrete intervals (i.e. coarsening), whereas each class of categorical variables is an additional stratum. Each observation is classified according to the combinations of the strata. For each strata, the CEM calculates taxpayer-specific weights w_i. The weight of observation i in stratum s is defined as follows:

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

In the bottom case of Equation (2), N^U and N^U_s refer to the number of untreated taxpayers in the whole sample and in stratum s respectively. Similarly, N^T and N^T_s are the number of taxpayers in stratum s and in the whole sample that have been audited. Unmatched units receive zero weight. These weights are used to reweight untreated observations and replacing the quantity in Equation (1) with the weighted mean in Equation (3).

$$\widehat{ATT} = \frac{\sum_{i \in \tau} Y^T_i}{N^T} - \frac{\sum_{j \notin \tau} w_j Y^U_j}{\sum_{j \notin \tau} w_j} \tag{3}$$

This weighting scheme given by 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. Furthermore, CEM overcomes the need to control for observables. Lastly, CEM superimposes matched data to the area of common support. In addition, compared with stochastic matching methods, such as the commonly adopted propensity score matching (PSM), CEM does not suffer for several issues related with PSM, in particular the risk of an increase in imbalance between groups by pruning observations. 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).

5 Results

5.1 Audit rule determinants

As anticipated, the application of several RF allows us to compare the accuracy measure of each RF. The results of the RF-OOB accuracy can be observed in Figure 1 and summarized in Table 4. Figure 1 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 1: 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 1 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 this 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 1: Accuracy measures for the RF

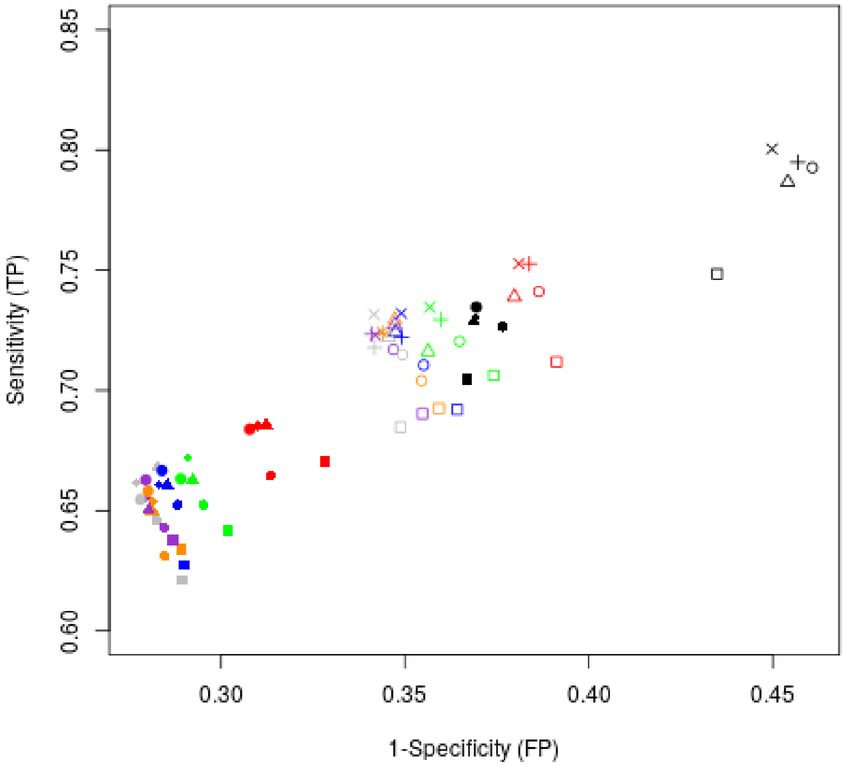


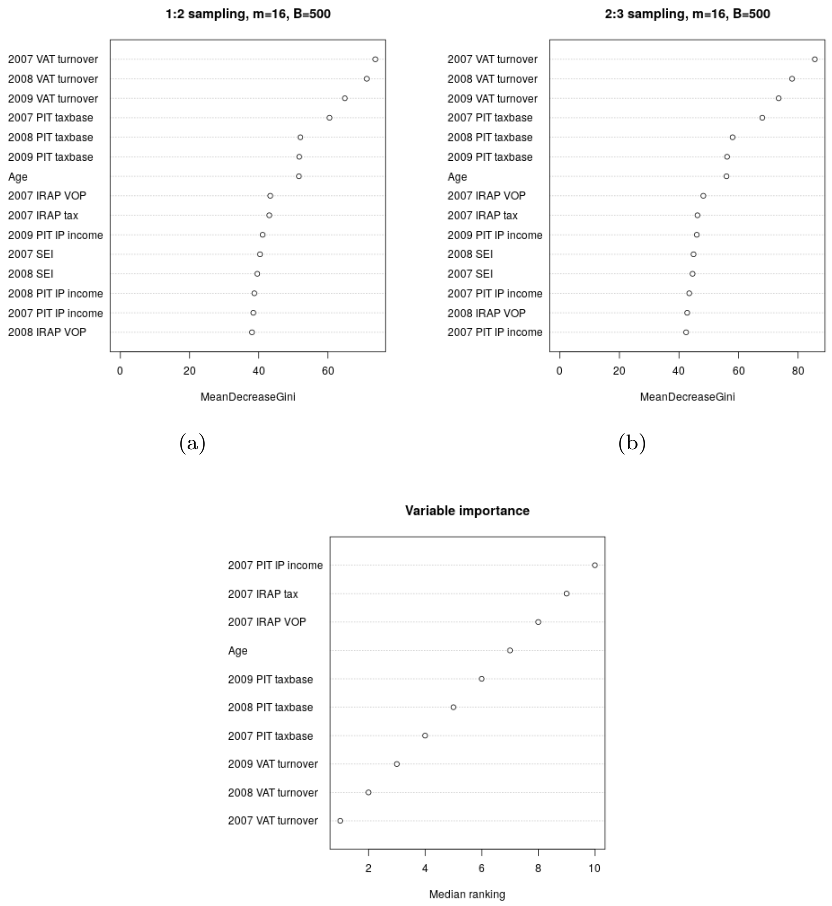
Table 4: Accuracy measures for the RF

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.	0.6847	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
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 2 we can appreciate the variable importance in our RFs [4]. The top left panel represent the most accurate 1:2 sampling RF, whereas the top right panel is related to the most accurate 2:3 sampling RF. The bottom panel summarizes all the 35 RFs by reporting the top ten variables in terms of median importance ranking. The first six rows of top and central panels show that the larger gain in terms of decrease of Gini index is based on the *personal income tax base* and *vat turnover* related to fiscal years preceding the audit (l4 denotes,2007, l3 denotes 2008 and l2 denotes 2009). Such variables contribute to node purity to a larger extent along with *age*. This means that these measures are likely to be used by the revenue agency for selecting the audit candidates.

[4] ’Irf’ stands for personal income tax, ’redimpo’ for its tax base, and ’red’ for a type of income included; ’ira’ stands for irap, ’iva’ stands for vat and ’volaff’ stands for VAT-turnover; ’eta’ is age

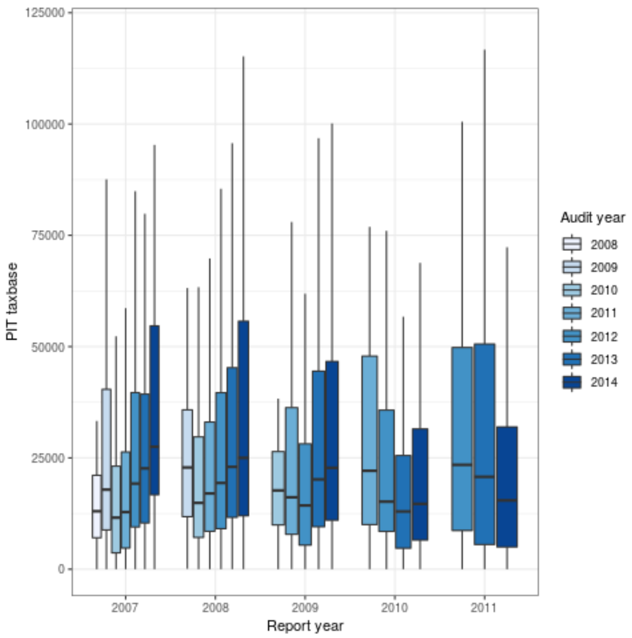


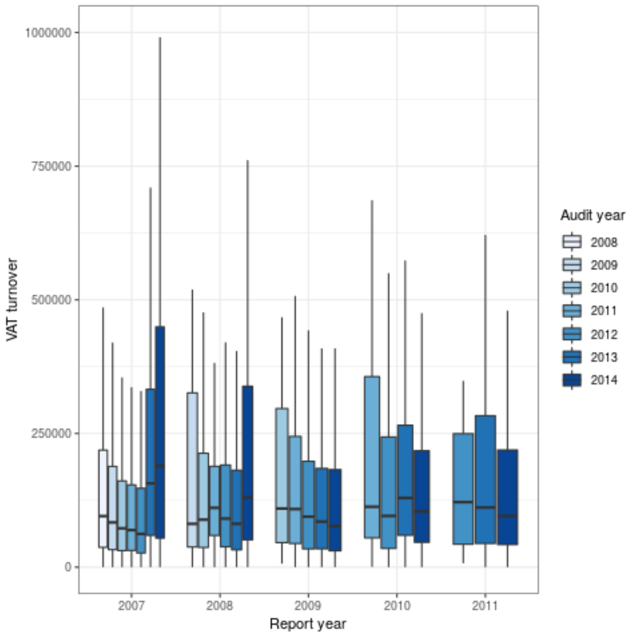
(a)

(b)

(c)

Figure 2: Figure (a) and (b) shows the summary of variable importance: RF with maximum CCR. Figure (c) presents the median ranking in variable importance. (Legend: VOP=value of production, SEI=self-employment income, IP=immovable property)





The boxplot reports, on the horizontal axis, the report year and, on the vertical axis, the average of PIT taxbase (first graph) and that of VAT turnover (second graph). Mean values are reported only for audited taxpayers, and only for report years preceeding the audit year. These graphs visualize the variability of the tax reports observed by the tax authority when it selects taxpayers to be audited.

Tax reports referring to 2007 and 2008 represent two-thirds of the audited reports in our sample. Our interpretation of the audit rule followed by the tax authority for these two audit cohorts is based on the business cycle. In nominal terms [5] the Italian GDP reached a peak in 2008, declined in the following two years and bounced back up in 2011, though at a level lower than that reached in 2008. The graph provides some evidence for the hypothesis that the tax authority has focussed on taxpayers reporting in a way not consistent with the business cycle.

Consider, in particular, audit years 2010 and 2011, i.e. the cohorts on which we focus in this paper because we can observe the reaction of taxpayers to audits. Both these cohorts of audited taxpayers reported a VAT turnover in 2007 and 2008 that, on average, is *lower* than that reported in 2009 and 2010. Also, he average PIT taxbase reported in 2007 and in 2008 by taxpayers reported in 2011 is, again, lower than that reported in the crisis years, 2009 and 2010. The only (partial) exception to this suspicious pattern is that concerning the PIT taxbase by taxpayers audited in 2010: for this cohort it is still true that the 2007 average is lower than the 2009 average, but the latter is lower than the 2008 average.

Now consider taxpayers audited in 2012 and 2013. Although we do not consider them in our analysis, it is still instructive to observe the pattern of average reports. Again, the majority of tax reports audited in these cohorts are from tax years 2007 and 2008. However, these taxpayers do report, in 2007 and 2008, on average, values of both the PIT taxbase and of VAT turnover that are

[5] Currently in Italy tax bases are not adjusted for inflation.

higher than those they report in 2009 and 2010. Thus, the audit rule appears to be different from that followed previously. An insight can be derived from the fact that average values they report in 2011 are higher than those reported in 2007 and 2008, despite in 2011 the Italian economy had not fully recovered the loss generated by the 2009-2010 crisis. However, a more complete view of the audit rule would require to observe also 2012 reports.

In sum, the selection of audited tax reports from tax years 2007 and 2008 seem to be based on an apparent inconsistency between the reporting behaviour and the business cycle. For 2010 and 2011 audit cohorts, this inconsistency lies in average 2007 and 2008 reports that are *lower* than those issued in the following two years, i.e. during the economic crisis. For 2012 and 2013 audit cohorts, the inconsistency may be related to the fact that 2007 and 2008 reports are *lower* than those issued in 2011, a year where the economic recovery from the crisis was still incomplete[6].

5.2 Application of CEM

The CEM matched 1372 treated taxpayers to 39764 untreated (Table 5). Such observations are in the common support area of the matching variables.

Table 5: CEM results

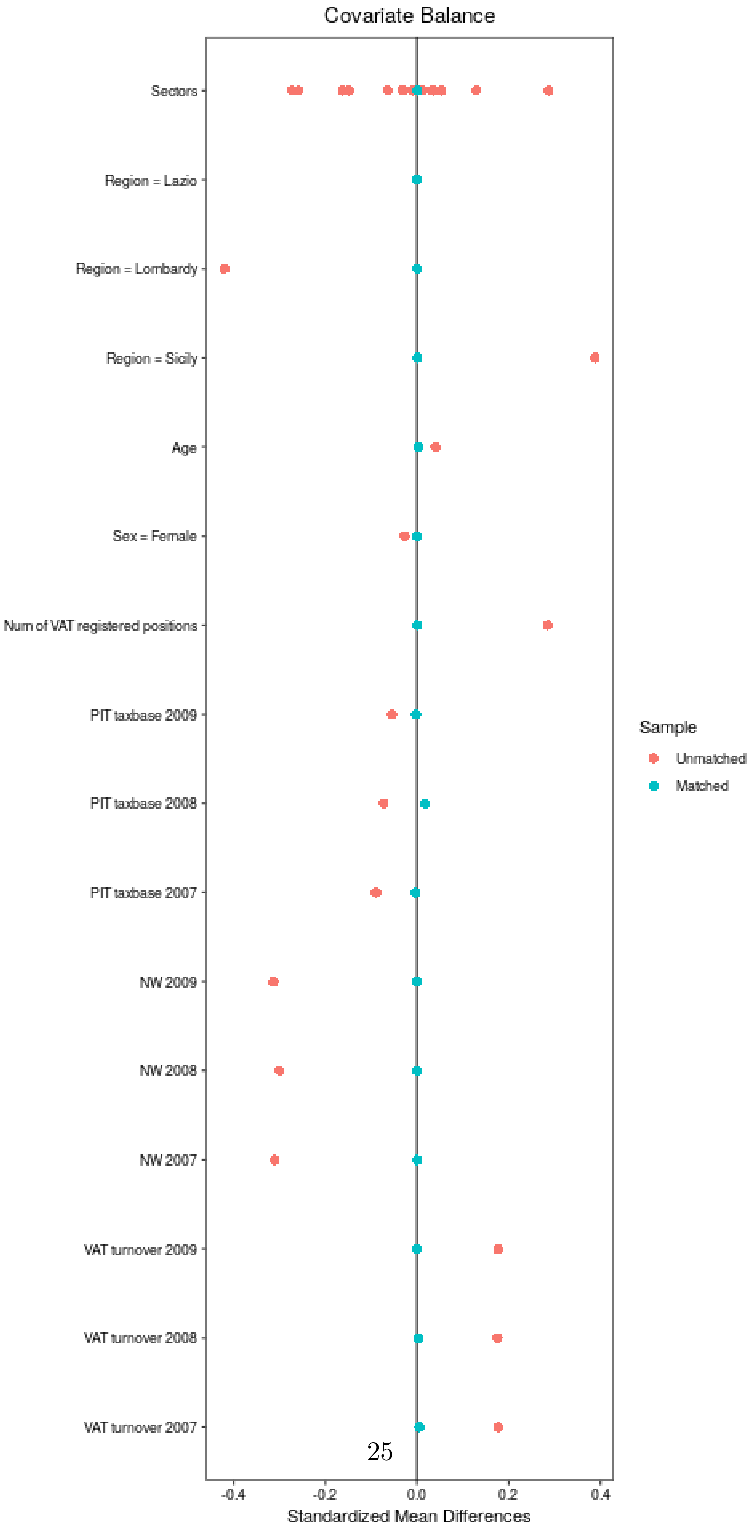
	Unmatched	Matched	Proportion Matched
Untreated	620150	39764	6 %
Treated	955	1372	59%

Covariate balance for our matching strategy is depicted in Figure 3.

[6] Our interpretation is broadly consistent also with descriptive evidence available for 2014 cohorts. This is composed by taxpayer whose audited tax reports are mainly from 2009. Indeed, these taxpayers report in 2007 and 2008 are perfectly consistent with the business cycle: both the PIT taxbase and the VAT turnover in 2007 and 2008 are higher than the following ones from the 2009-2011 period, with the 2011 reports being slightly higher than those from 2010. Thus, from our data there is no reason to audit these taxpayers which suggest that they might have been selected on their reports from periods that we do not observe here, i.e. from 2012 and 2013 tax years

Concerning time invariant taxpayer characteristics such as sector, presence of employees, region of activity and sex, CEM approach achieves a perfect covariate balance: the sample composition in terms of the above-mentioned variables is equal between the treated and untreated. This means that in post-CEM comparison mean difference between treated and control over all the taxpayers' characteristics included in the matching procedure is very close to zero.

Figure 3: Balance plot for matching variables (2010 treatment)



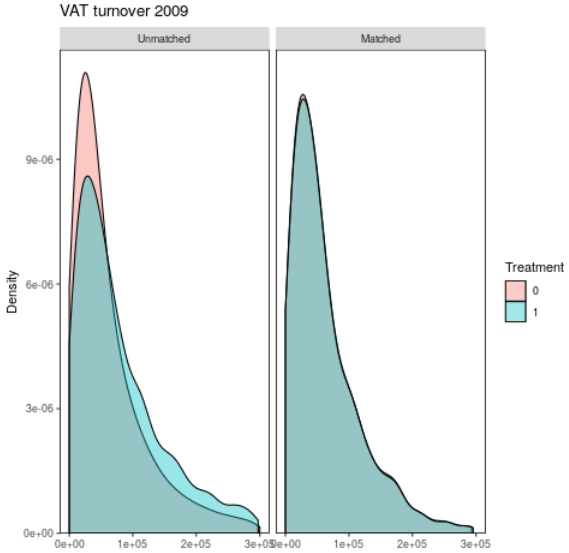
The largest imbalance is related to 2009 PIT taxbase; however, the post matching mean difference between treated and untreated is not significant (Table 6). As shown in Table 6 (top panel), prior to CEM the audit determinants were significantly different between the treated and the untreated. The order of magnitude was the thousand Euro for PIT taxbase and the hundred of thousand Euro for the VAT turnover. Following the CEM, such differences shift to hundreds Euro and thousands Euro respectively, while p-values range from 16% to 99%. Our matching strategy also improves distributional balance: this means that the untreated and treated are not only not statistically different in mean but also in distributional terms. This result is visible in Figure 4: the overlapping of the density plot largely improved from left to right panel. This also guarantees that the common support assumption is satisfied (Cunningham, 2021).

Table 6: Balancing of pre-treatment covariates. Pre-CWM and Post-CWM

	Mean (St. Dev.)			
	Pre-Matching			
	Untreated		Treated	
PIT taxbase 2009 *	29,329	(81,063)	26,295	(55,769)
PIT taxbase 2008 ***	30,221	(77,967)	25,940	(58,505)
PIT taxbase 2007 ***	30,131	(81,459)	25,305	(53,862)
VAT turnover 2009 ***	102,871	(251,946)	215,858	(640,293)
VAT turnover 2008 ***	109,708	(272,808)	241,192	(749,861)
VAT turnover 2007 ***	105,925	(278,723)	234,880	(727,726)
	Post-Matching			
	Untreated		Treated	
PIT taxbase 2009 ^{ns}	17,949	(53,136)	17,828	(41,301)
PIT taxbase 2008 ^{ns}	18,436	(53,125)	19,455	(60,344)
PIT taxbase 2007 ^{ns}	18,771	(46,349)	18,574	(50,670)
VAT turnover 2009 ^{ns}	72,424	(85,871)	72,415	(85,913)
VAT turnover 2008 ^{ns}	76,136	(94,061)	78,103	(90,883)
VAT turnover 2007 ^{ns}	74,637	(96,588)	78,358	(94,017)

t-Test for mean difference between untreated and treated:
$*p < 0.1, **p < 0.05, ***p < 0.01$, ns not significant

Figure 4: Distributional balance



5.3 ATT estimates

We now apply equation (3) using as $Y's$ 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 *asbolute 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 [7].

The ATT estimates in Table 7 shed some light on the taxpayers' response to the audit.

When we consider *asbolute differences*, the audits are associated to a positive but not significant variations of 2010 PIT taxbase, and to a positive

[7] ATT estimates of semi-elasticities are obtained using a Poisson Pseudo Maximum Likelihood, PPML, estimator

and significant variation of 2011 PIT taxbase.

Both coefficients on absolute VAT turnover variations, for 2010 and for 2011, are positive.

However, here we are interested in relative differences more than in absolute ones, also because the latter are influenced by the business cycle. Thus, we focus here on ATT estimates of *semi-elasticities*. On average, in 2010, i.e in the tax report following the audit, audited taxpayers reported a value of the PIT taxbase that was 6.4% higher than that reported by matched unaudited taxpayers.

The semi-elasticity of reported income almost doubled in 2011, reaching a value of 12.9%. A similar pattern, although with different values, is shown by VAT turnover, whose increase in percentage terms is not significant in 2010 but significant, and equal to 9.8% in 2011.

Table 7: ATT for 2010 audits

	2010 PIT taxbase	2011 PIT taxbase	2010 VAT turnover	2011 VAT turnover
Differences	1,260.393	2,595.778*	5,358.915**	$9,355.653^{***}$
	(1,296.560)	(1,465.496)	(2,527.782)	(2,584.468)
Semi-Elasticity	0.064***	0.129***	0.007	0.098*
	(0.0002)	(0.0002)	(0.005)	(0.050)
	Legend: ATT (St. Err.)			
	*p<0.1; **p<0.05; ***p<0.01			

As a robustness check, we repeat the same approach to 2011 audits. After checking for the balance (see Tables 6 and 9 in the Appendix) we again tested for absolute differences and for semi-elasticity of PIT taxbases and VAT turnover of taxpayers audited in 2011 with respect to reports made by matched unaudited taxpayer. The pattern, displayed in Table 8, is consistent with our previous analysis: 2011 audits have a clear impact on tax reports that are surely issued after the audits, i.e those referring to tax year 2011. while the impact is less clear on 2010 tax reports because, on average, they might well have been issued before the audit was conducted.

Table 8: ATT for 2011 audits

	2010 PIT taxbase	2011 PIT taxbase	2011 2010 VAT turnover	2011 VAT turnover
Difference	172.597	1,774.537***	1,796.175	4,448.776***
	(406.426)	(424.303)	(1,200.057)	(1,227.203)
Elasticity	0.012***	0.114***	0.011***	0.041***
	(0.0002)	(0.0002)	(0.00005)	(0.0001)

Legend: ATT (Std. Error)
*p<0.1; **p<0.05; ***p<0.01

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 that, to evaluate the optimal number of audits, it is crucial to take such impact into account.

6 Optimal number of audits

Keen and Slemrod (2017) show that when a piecewise-linear tax schedule is applied the condition for α to be an optimal level of enforcement (audits in our case) is the following:

$$E(T, \alpha) = \alpha((c_\alpha/v') + a_\alpha)/T \qquad (4)$$

where $E(T, \alpha)$ is the elasticity of the tax revenue, T, with respect to α, c_α and a_α are the first order derivatives of private compliance costs, c, and public administrative costs, a, with respect to α and v' is the marginal social utility of an additional Euro of public spending.

On the left-hand side, $E(T, \alpha)$ is the sum of a *direct effect*, i.e. the increase in additional taxes collected with the audit, and of an *indirect effect*, i.e. the increase in additional taxes reported by the audited taxpayer after the audit.

The direct effect elasticity, that we denote as E_{de} can be computed using data on adjusted taxbases and adjusted taxes, including sanctions. The indirect effect is more easily estimated looking at the taxbase elasticity, as we did in Section 5.3 and then computing the associated elasticity of the tax revenue. These two elasticities are linked as follows

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

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, $T_z \geq \overline{T}$.

If one assumes, following Keen and Slemrod (2017), that both $c(\alpha)$ and $a(\alpha)$ are homogeneous functions then we can finally write that the optimal level of enforcement must satisfy the following condition

$$E(z,\alpha)\frac{T_z}{\overline{T}} + E_{de} = \frac{a}{T} + \frac{c}{v'T} \tag{6}$$

From Section 5.3 we know that the average indirect effect of audits for Italy can be estimated, considering only PIT revenues, as $E(z,\alpha) = 9.65\%$ which is the average estimate of elasticity of additional income reported in 2010 and 2011 by taxpayers audited in 2010, with respect to matched unaudited ones.

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 11 and 12). 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.

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 the direct effect elasticity is equal to $10,6\%$, very close to the estimate of 10% provided for the US by Alm (2012) [8]. Summing up, the LHS of (6) amounts at $9.65\% x 1,06 + 10.6\%) = 20,83\%$.

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 Keen and Slemrod (2017) for the US.

However, 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.

From equation (6) we can easily calculate optimal number of audits as a function of the ratio between compliance and administrative costs as

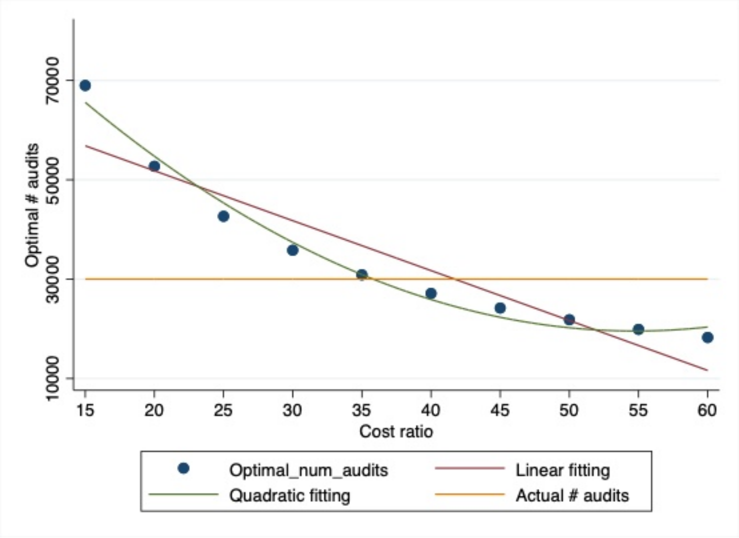
$$N = E(z, \alpha) \frac{T_z}{\overline{T}} \frac{T}{A} \frac{v'}{v' + r} \tag{7}$$

[8] 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.

where N is the number of audits, T is tax revenue, A is the average cost of an audit (obtained by dividing the administrative cost of the Italian revenue agency by the number of audits) and r is the cost ratio, i.e. the ratio between private compliance costs, c, and public administrative costs, a.

By assuming that each audit has a constant cost and that $v' = 1,2$, and knowing that in Italy approximately $30,000$ of audits are conducted every year, the optimal number of audits is an decreasing function of the cost ratio.

Figure 5: Optimal number of audits as a function of the cost ratio



In Figure 5 the blue dots represents the optimal number of audits in Italy, whereas the red line and the green one are respectively the linear and quadratic fit. The quadratic line is the better fitting for the points and it reaches the higher value, i.e slightly less than $70,000$ audits per year, when the cost ratio is equal to the US one. For higher values of the cost ratio the optimal number of audits decreases. It is equal to the observed one, $30,000$ i.e. the orange line, when the cost ratio equals 35, which is more than twice times larger than the US cost ratio. In sum, even taking into account the complexity and uncertainty of Italian administrative procedures, it appears reasonable to conclude that the actual number of audits in Italy is sub-optimal.

7 Concluding Remarks

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 between theory and practice is even more remarkable given that tax administration issues have finally gained the importance they deserve in the international policy context. Even a superficial look at the publications issued by the Oecd *Forum on Tax Administration, FTA* reveals, however, that tax authorities around the world are not interested in random auditing but, on the contrary, that they are keen on searching ways to take advantage of the Big data revolution so to better target their audit policies and make them more cost effective.

This is why we believe that our paper, which is the first to our knowledge to implement optimal tax administration rules using real-world data, can contribute to bridge the gap between the public finance literature and the tax administration practices and, by doing so, to characterize the latter as a real alternative to more traditional ways of raising revenues, e.g increases of tax rates.

A last remark concerns the distributional impacts, that are not explictly included in optimal tax administration conditions, whilst they are included in formulae used for optimal tax rates. It is known that, when there are audit capacity constraints, audits should focus only on reports lower than a threshold, so that, if the threat of an audit is credible enough to prevent tax evasion, audits tend to create a *regressive bias* (see Andreoni et al. (1998)). The evidence provided in this paper seems to go in the same direction, suggesting that tax authorities tend to focus on taxpayers whose tax reports deviate from the business cycle. The existence of such a regressive bias could clearly limit the attractiveness of tax enforcement policies from a social welfare perspective.

8 Appendix

8.1 Application of CEM for 2011 audits

The CEM matched 2789 treated taxpayers to 28823 untreated (Table 9) for which the common support assumption is satisfied. In relative term the proportion of matched among the untreated is lower than what shown in Table 5 (37% vs 59%). Figure 6 shows that the standardized mean difference between the treated and the untreated is null for all the variables, hence the matched data are perfect balanced after the application of CEM. Table 10 shows that the difference between the treated and untreated subsamples are statistically insignficant after weighting the observation by the CEM weights. Prior to matching, the PIT taxbase of fiscal year 2009 and all the variables related to VAT turnover were significantly different between the treated and untreated.

	Unmatchted	Matched	Proportion Matched
Untreated	625802	28823	4 %
Treated	4827	2789	37%

Table 9: CEM results for 2011 audits

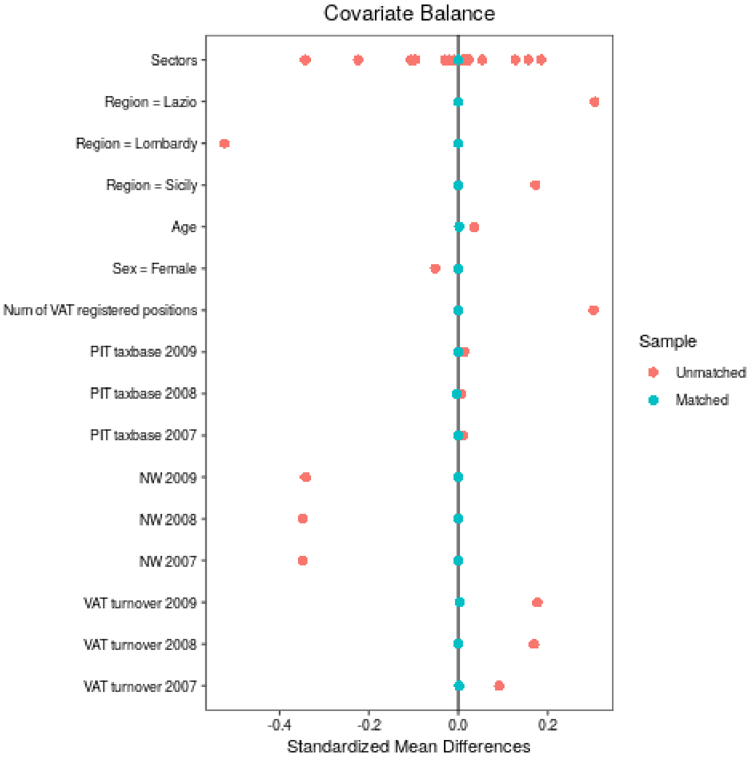


Figure 6: Balance plot for matching variables (2011 treatment)

	Pre Matching			
	Untreated		Treated	
PIT taxbase 2009 *	29,300	(80,334)	30,955	(125,007)
PIT taxbase 2008 ^{ns}	30,195	(77,209)	31,160	(124,062)
PIT taxbase 2007 ^{ns}	30,101	(80,811	31,246	(120,526)
VAT turnover 2009 ***	102,472	(252,300)	171,670	(391,276)
VAT turnover 2008 ***	109,254	(272,850)	188,892	(469,580)
VAT turnover 2007 ***	105,418	(265,612)	188,923	(909,611)
	Post Matching			
	Untreated		Treated	
PIT taxbase 2009 ^{ns}	13,516	(18,450)	13,473	(19,236)
PIT taxbase 2008 ^{ns}	13,754	(18,995)	13,207	(18,295)
PIT taxbase 2007 ^{ns}	12,873	(16,685)	12,738	(16,720)
VAT turnover 2009 ^{ns}	60,314	(58,846)	61,411	(60,283)
VAT turnover 2008 ^{ns}	62,827	(58,926)	62,940	(58,998)
VAT turnover 2007 ^{ns}	59,784	(60,413)	61,525	(59,220)

t-Test for mean difference between untreated and treated:
$*p < 0.1, **p < 0.05, ***p < 0.01$, ns not significant

Table 10: Balancing results treatment 2011

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.

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

PIT taxbase reported by 2010 audit cohort after the audit				
vingtile	2010		2011	
	min	max	min	max
1	0	0	0	0
2	0	0	0	0
3	1	275	1	996
4	279	2410	1038	3206
5	2422	4375	3290	4659
6	4437	5857	4661	6389
7	5908	7213	6432	7615
8	7232	8461	7632	8789
9	8542	9920	8792	9663
10	9924	10733	9696	11043
11	10739	12064	11047	12030
12	12065	13047	12032	13409
13	13048	14417	13440	14674
14	14454	16350	14738	16718
15	16372	18206	16721	19830
16	18250	21539	19838	23467
17	21608	27867	23493	29520
18	27938	38109	29649	39625
19	38151	66040	39934	69618
20	66061	1127332	71020	1295559

Table 11: PIT taxbase, post audit

PIT taxbase reported by 2010 audit cohort before the audit						
	2007		2008		2009	
ventile	min	max	min	max	min	max
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	1	223	189	2569	362	432
5	235	1558	2649	4668	1083	3691
6	1642	3333	4673	6387	3924	6366
7	3342	4789	6408	7918	6426	8212
8	4809	6544	8063	9051	8714	10018
9	6552	7817	9069	10107	10231	11576
10	7878	9095	10114	11103	11652	12012
11	9100	10345	11114	12799	12214	12800
12	10346	11244	12856	13708	12808	13689
13	11250	12623	13723	15094	14002	14564
14	12663	14365	15256	17054	14716	15821
15	14400	15983	17168	18938	16037	17962
16	15985	18930	18985	21343	17984	20812
17	19111	23128	21407	26423	20974	25557
18	23212	34461	26496	34815	30314	40535
19	34541	62542	34934	53898	47542	72594
20	63041	993152	54841	503921	106672	147465

Table 12: PIT taxbase, pre audit

Beer, S., M. Kasper, and B. Erard (October 2019). Do audits deter or provoke future tax noncompliance? evidence on self-employed taxpayers.

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.

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.

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.

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*.

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.

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.