EIEF WORKING PAPER SERIES

EINAUDI INSTITUTE FOR ECONOMICS AND FINANCE

EIEF Working Paper 23/02

January 2023

# Refining Public Policies with Machine Learning: The Case of Tax Auditing

By

Marco Battaglini
(Bocconi University and EIEF)

Luigi Guiso
(EIEF)

Chiara Lacava
(Goethe University Frankfurt)

Douglas L. Miller
(Cornell University)

Eleonora Patacchini
(Bocconi University)

# Refining Public Policies with Machine Learning: The Case of Tax Auditing*

Marco Battaglini† Luigi Guiso‡ Chiara Lacava§
Douglas L. Miller¶ Eleonora Patacchini‖

We study how ML techniques can be used to improve tax auditing efficiency using administrative data without the need of randomized audits. Using Italy's population data on sole proprietorship tax returns and audits, our new approach addresses the challenge that predictions must be trained on human-selected data. There are substantial margins for raising revenue from audits by improving the selection of taxpayers to audit with ML. Replacing the 10% least promising audits with an equal number selected by our algorithm raises detected tax evasion by as much as 38%, and evasion that is actually paid back by 29%.

*Keywords*: tax enforcement, tax evasion, policy prediction problems.

*JEL classification*: C55, H26.

## 1 Introduction

Tax authorities routinely collect deep datasets on tax filers that can be used to identify auditing targets. This makes the choice of auditing strategy a prime candidate as an

*We are very grateful to the Italian Revenue Agency for granting us access to the data. We are solely responsible for the ideas expressed in the paper. We thank seminars participants at Cornell University, ETH Zurich and the Italian Presidency of the Council of Ministers for valuable discussions.

†Bocconi University and EIEF E-Mail: marco.battaglini@unibocconi..edu
‡EIEF E-Mail: luigi.guiso55@gmail.com
§Goethe University Frankfurt E-Mail: lacava@econ.uni-frankfurt.de
¶Cornell University E-Mail: dlm336@cornell.edu
‖Bocconi University E-Mail: eleonora.patacchini@unibocconi.edu

application for machine learning techniques (henceforth, ML), that can be deployed to exploit available information efficiently, consistently and transparently. While both tax authorities and researchers are aware of these opportunities, the opacity of the auditing processes followed by most tax authorities makes it unclear the extent to which they operate at the "production possibility frontier" or whether there are margins for improvements by a more efficient use of data.

In this paper, we exploit an exclusive dataset from the Italian Revenue Agency (henceforth, IRA) to explore whether ML techniques can be used to improve audit policies. The dataset includes the tax returns with relative audits and audits' results for the universe of non-incorporated small businesses in Italy from 2007 to 2012. The dataset, moreover, also includes information on whether the taxpayer appealed against the audit as well as all the statistical information available to the IRA concerning the tax file and filer.

The general idea behind ML techniques is to exploit data on policy outcomes (in our case audits) to train a predictive algorithm designed to achieve specific goals (maximizing the probability of finding evasion, for example). Ideally, after validating the analysis out of sample, the algorithm can be used to guide the policy (the choice of which files to audit, in our specific case). Even when detailed data is available, however, two challenges make the design and evaluation of policies with ML a difficult task. The first is what Kleinberg et al. (2018) have defined the *selective labels problem*: only outcomes of files that have been endogenously selected for treatment are observed. In our case, this problem is particularly serious if the IRA selects audits relying also on unobserved variables that may be relevant for audits' performance. The second problem is the *omitted payoff bias*. This refers to the fact that the objectives of the policymakers may be multidimensional and unobserved, so an auditing policy that is unsuccessful with respect to a narrow measure of success, may instead be justified when all the goals of the tax authority are considered. Our data allows to make progress in evaluating the benefits of improving the auditing process despite these two problems. The approaches we propose exploit specific features of auditing data common to several countries, namely the facts that the tax authority is severely limited in the number of audits and that currently unaudited files can occasionally be audited at a later date. Both of these features can be found in other environments, thus our strategies can be applied in other contexts as well.

We start our analysis by documenting the extent to which a ML algorithm can be used to identify audits that perform particularly poorly within the set of observed audits. Since we observe the universe of tax audits and relative outcomes, we can test

our ability of identifying the audits that perform poorly under a variety of criteria. Contrary to other types of policymaker, tax authorities have a narrow policy mandate and do not have significant latitude in deciding their policy goals. Still, their activity is driven by at least two goals: maximizing detected tax evasion; and maximizing the amount of evaded taxes recovered by the end of the auditing process.[1] The two goals may differ because some audits may appear to be promising in terms of detected evasion but much less so in terms of the amounts that can be recuperated since taxpayers may appeal against the audit. Our dataset allows us to assess both goals. We show that our ML algorithms can accurately rank audits in terms of both expected tax evasion and expected recovered evasion. More importantly, we show that the performance of the audits that are predicted as worst performing is extremely poor. Eliminating the bottom 10% of the audits, would induce a reduction in detected evasion of only 3%; even more importantly it would also induce a reduction of recovered evasion of only 2.6% implying that the omitted payoff bias problem, while important in principle, may not alter qualitative conclusions in practice.

Once we have identified the poorly performing audits, the next question is whether we can replace the worst performing audits with ex ante superior tax files. Here is where the selective labeling problem starts to bite. To address this question, we propose two complementary strategies. The first strategy relies on the longitudinal nature of our dataset to choose the replacing files. In Italy (like in many other countries), the IRA has five years to audit a tax file. While most files are never audited at all, some are audited in later years. This fact gives us a plausible counterfactual for which we can actually observe the true outcome of an audit: a file that is auditable but unaudited at $t$ is identical in all following periods in which it is auditable since it refers to a tax year preceding $t$. We can therefore replace an audited file we predict to be non-performing with a file that is available for auditing today but audited in following years. We find that replacing the worst predicted 10% of files with an equal number of the best unaudited files audited later yields an improvement of 38% in detected tax evasion. The set of unaudited files that are audited at a later date may of course be different from the general population. It is however unlikely that the IRA intentionally postpones the audit of good files. By not auditing a "good" file at time $t$, an agent of the IRA exposes the agency to the risk of never auditing it in the future (if overlooked by future agents) or to the risk of losing the ability to recuperate

[1]These statutory goals are well defined by directives from the Treasury and have been described to us in interviews by many officials from the IRA. Auditors do not have authority to deviate from these goals.

any evaded income since some companies may dissolve or go bankrupt before the IRA can document a claim on the firm balance sheet. Indeed, we document that files that remain unaudited for a few years but then are audited are not fundamentally different from other audited files, in terms of both detected and recovered tax evasion.

The second strategy attempts to bound from below the value of replacement without using ML for selecting the replacement files. The strategy relies on the fact that the IRA is severely constrained in terms of resources, so much that only about 3% of the sole proprietorships' files are audited (for comparison, in 2017 the audit coverage on personal income in France was 5%, and that of the EITC recipients in the U.S. was 6%). If the authority were to eliminate from a list of proposed audits the bottom 10% of files according to the predicted performance and replace them with an equal (to fulfill the resource constraint) number of files with random performance, would the replacement be worthwhile? It is reasonable to assume that for such a marginal substitution, the replaced files will not be very different from the average audited files. How bad should be the quality of a replacement relative to the audited files, to make such substitution undesirable? We show that replacing the bottom 10% of the worst expected files with average audited files would increase both detected tax evasion and recovered tax evasion by 6.7% and 7.6%, respectively. Indeed, the replacement would be advantageous even if the replacing files were significantly below average.

A common risk when using ML to guide policy decision is to inject unintended bias in the decisions. We assess the extent to which the replacements may introduce bias by comparing a large set of sensitive observable characteristics before and after the replacement. We however do not observe significant differences in terms of key demographics characteristics of the tax filer (gender, age and marital status) and of the nature of their business (family business status, years of activity and employees).

# 2 Related Literature

A significant and growing literature at the intersection between computer science and economics applies ML techniques to policy problems. For example, several papers present algorithms to detect tax evasion (Bonchi et al., 1999; Bots and Lohman 2003; Cleary, 2011; Hsu et al., 2015; Ruan et al., 2019; Wu et al., 2020; among others), insurance fraud (Bhowmik, 2011), and fraudulent financial statements (Kirkos et al., 2007).[2] These papers focus attention on the design of algorithms to predict a positive

[2]These works are constrained by much smaller datasets than ours, typically limited to a few thousand taxpayers.

outcome, limiting the evaluation of their performance to the quality of the out-of-sample predictions, thus ignoring the selective labels problem and omitted payoff bias. Therefore, they offer little guidance to the policymaker who is interested in deciding how to allocate a scarce resource (Athey, 2017). This question is as important for policy purposes as difficult to solve empirically because it requires information on counterfactual conditions. While this problem is mitigated in cases of random allocation of treatments (e.g. random audits),[3] policy interventions are almost never at random. This paper is the first to propose a solution to this question in the context of tax auditing. The strategies that we propose, however, can be applied to any allocation problem of a scarce resource where longitudinal data are available, a fraction of untreated units are treated at a later date, and their outcomes remain unchanged during the period. The importance of the selective labels problem for public policy application is highlighted by Lakkaraju and Rudin (2017), Jung et al. (2017), who rely on a "selection on observables" assumption to assess ML applied to judicial decisions to release or detain defendants while they await trial. Lakkaraju et al. (2017) and Kleinberg et al. (2018) rely on institutional features of these decisions. In particular, Kleinberg et al. (2018) leverage the quasi-random assignment of cases to judges of differential leniency: they use the algorithm's predictions for cases handled by lenient judges to predict the outcomes for defendants released by more stringent judges. The institutional features that enable the Kleinberg et al. (2018) solution to the selective labels problem may not be available in all policy prediction applications. A key contribution of our paper is to identify an alternative approach to the selective labels problem. Instead of relying on random shocks to propensity to observe the labels, we use the feature that some labels are only revealed later in time.

# 3 Institutional Setting and Data

The taxpayers in our dataset are individuals who own a sole proprietorship, where no legal distinction is made between the enterprise and the sole owner. In most countries, this fiscal category is the subsample of taxpayers characterized by the highest evasion rate and accounts for a relevant portion of the total tax gap. We merge information from two different administrative records that the IRA shared uniquely with us: returns files and audit files. Records are at the individual level

[3]Recently, Ash et al. (2021) use randomized audits to evaluate the benefits of using a ML algorithm to predict corruption in Brazilian municipalities. In the context of gun violence prevention and energy consumption prediction, Bertrand et al. (2022) and Knittel and Stolper (2021) rely on randomized control trials.

and cover statements of incomes generated from 2007 to 2012, reported between 2008 and 2013, and audited between 2009 and 2014. Overall, our sample contains 19 million tax returns filed by almost 4.8 million taxpayers, and 405,115 audits. This database (the Tax Registry) is the one used by the IRA to select audits. It includes detailed information on all components of taxpayers' tax returns (including reported taxable income, turnover, liabilities and deductions) and characteristics of their business (sector, geographical location, years of activity, number of employees). The audit data contain information on whether and when a file was audited, as well as the amount of evaded tax assessed (if any).[4] Additionally, the IRA shared with us the exclusive information on whether the taxpayer is insolvent, that is when the audited taxpayer does not pay back the assessed evasion. These cases of insolvency can originate from an appeal against the audit or simply a failure to respond to the audit notification. In Italy, an average of 5% of audited taxpayers appeal against an audit, and another 37% of taxpayers neither pay nor appeal within due time after the audit notification. Both cases of insolvency trigger complicated processes, which last for many years, entail complex schemes of sanctions, and may involve several layers of judiciary. The probability of insolvency increases faster for higher levels of tax evasion, thus hampering seriously the ability of the IRA to recover evasion. Further details on our data and summary statistics are reported in the Online Appendix A.

# 4 The Machine Learning Algorithm

**Predicted variables.** We tailor our statistical model around the two goals of the IRA: detecting evasion and recovering the detected amount of evaded tax. More specifically, for each file our two main variables of interest are i) tax evasion, defined as the difference between the tax amount assessed during an audit and the tax paid (labeled as *TaxEva*); and ii) a proxy for the actual tax evasion recovered by the IRA (labeled as *TaxGot*). This is equal to tax evasion when the taxpayer pays back within due time, and zero when the taxpayer is insolvent.[5] Because the policy goals are in levels of detected and recovered evasion, we predict levels of evasion (rather than e.g. binary indicators for a certain level of evasion).[6]

[4]We identify and exclude audits initiated by authorities cooperating with IRA (3%) that might use other information or selection criteria.

[5]Because of the lengthy and complicated process triggered by an appeal against an audit mentioned before, the actual amount of recovered evasion is not available for many observations in our sample. In the observed time frame, 97% of the taxpayers who appeal never pay back their debt.

[6]We have also explored models using the inverse hyperbolic sine transformation of our predicted variables. Working with these did not produce better predictions, and we focus on predicting

**Model.** We use a random forest to separately predict tax evasion and recoverable tax evasion (Breiman, 2001). This allows for rich interactions among explanatory variables, and easily adapts to non-linearities. Furthermore, this algorithm is suitable for handling very rich databases, which in addition to a large sample size feature a large number of explanatory variables, since the predictive variables are not used simultaneously. Our random forests contain 1,000 trees each.[7] The tuning parameters that we choose include a minimum leaf size of 28 observations to be eligible for a split, and 2.5% of features eligible for consideration at each split.[8] We train separate models for *TaxEva* and *TaxGot*.

**Predictors.** We use a rich selection of variables to predict *TaxEva* and *TaxGot*. We include business characteristics (years of activity, the number and logarithm of the number of employees, dummy indicators for the presence of employees and for the taxpayer being self-employed) and the full selection of financial variables included in the Tax Registry. For example, the latter include the reported taxable income (both the value and its logarithm), a dummy indicating a positive reported taxable income, reported taxable income net of employees deductions, gross income, revenues, taxable revenues, total assets, total liabilities, net value of production, VAT taxable turnover, operating costs, amortized costs, and VAT transactions. Importantly, we were granted access to the highest level of disaggregation of the sector of activity (ATECO 5-digit code, i.e. 1,215 sectors) and geographical location (municipality level, i.e. 8,054 municipalities). However, capturing this information using fixed effects poses computational challenges, given the large number of sectors and municipalities. We use this information as follows. First, we propose a specification with 5-digit sector of activity fixed effects and geographical fixed effects at the province level (110 provinces) (specification *i*). In an alternative specification, we instead exploit the granularity of the geographical information contained in the data by building Mundlak-type predictors (Mundlak, 1978), defined as the average at the municipality for two key financial accounts, namely taxable income and turnover. We add these variables to 5-digit sector of activity fixed effects (specification *ii*). Next, we exclude 5-digit sector of ac-

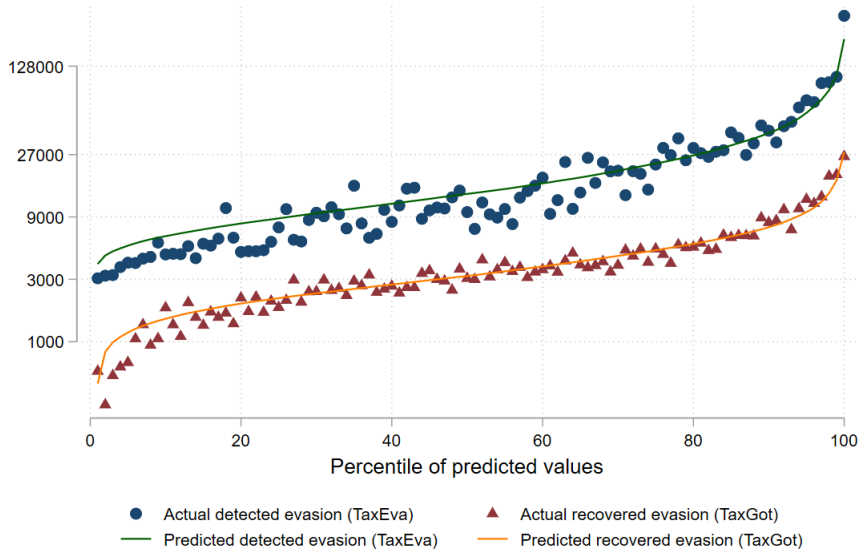
expected evasion in levels. We winsorize these variables at the top 1 percentile to prevent our model "chasing" extreme and idiosyncratic observations. To implement this, we consider the top 5% of the relevant variable, and estimate the parameters for a Pareto distribution to best match. We then use the conditional mean from this distribution of being in the top 1 percentile as the imputed value. Separately, the evaded tax can be negative if people pay a higher amount than the due tax. These instances are typically due to minor errors. We replace negative values of *TaxEva* or *TaxGot* with zeros.

[7]Increasing the number of trees further does not lead to a sizeable increase in the model fit.

[8]These were chosen based on a semi-structured grid search, using the random forest Out of Bag goodness of fit to guide the choice of tuning parameters.

tivity fixed effects and use Mundlak-type predictors at both municipality and 5-digit sector level (specification *iii*). We discuss the sensitivity of our predictions against alternative sets of predictors in the Online Appendix B. Because the performance of the different models is roughly equivalent, we use as baseline the more parsimonious version of the model, featuring more than 250 variables (specification *iii*).

**Samples.** We use files from years 2007, 2008 and 2009, for which we observe the complete fiscal cycle - that is the following 5 years in which they can receive an audit. By doing so, we get a representative sample of the composition of the files that are audited over time. To get a clean partition of the data, our randomized classification into training and testing samples occurs at the taxpayer (pseudo-)ID level; so that taxpayers in the training sample for one year are also in the training sample for all other years. Our training sample for the random forest prediction model consists of an 80% subset of the universe of audited taxpayers. To assess goodness of fit, we compare predictions for our testing sample, which consists of the 20% of audits from each of the fiscal years 2007-2009 excluded from the training sample. In total, our training sample contains 163,234 files, and our testing sample 40,869 files.




This figure reports the actual tax evasion (blue dots) and the recovered tax evasion (red triangles) of realized audits in the testing sample by percentiles of predicted values. The green and the orange line display the predicted tax evasion and the predicted recovered tax evasion, respectively. The sample includes files of fiscal years 2007-2012 that are audited by IRA.

Figure 1
Model Fit

**Out-of-sample assessment of prediction model.** To measure the model per-

formance, we compare predicted outcomes with outcomes assessed during the audit. In Figure 1, the curved green line reports the average predicted evasion at each percentile of predicted evasion and the blue dots report the average evasion detected by audits that were actually implemented. The curved orange line shows the average predicted *TaxGot* for each percentile of predicted *TaxGot*, and the red triangles report the average actual *TaxGot* for that percentile.

The model has a remarkable out-of-sample performance. The Online Appendix B provides a more rigorous analysis and explores additional tests of the validity of the prediction model. We show that its performance persists well for different testing samples, and that the random forest model does better than a straightforward OLS regression approach.

# 5 Policy Experiments

While the ML algorithm is very accurate in ranking the audited files in terms of their expected outcome, the gain of replacing the files that are predicted as poorly performing with files of superior performance is unclear. This depends on the expected outcome of unaudited files, for which we only have a ML prediction. To bypass this problem and quantify a lower bound on the benefit of using ML to refine the auditing process, we propose two types of exercises. First, we use the prediction model to calculate gains from a "discarding" exercise, where the audits with the worst predicted outcomes are simply discarded (Policy *A*). Audits have both administrative, human and economic costs on the target taxpayer (e.g. due to psychological distress and/or interference with the business). Performing audits with zero or minimal results is a waste of resources even without replacing them by superior audits. Second, we use the predictions for a "discard and replace" exercise, where these audits are discarded, and then replaced with audits from an alternative donor pool (Policy *B*). We propose two strategies to select the replacements: Policy *B.1* chooses replacements by using our ML algorithm; Policy *B.2* substitutes with random files. As for Policy B.1 we propose two alternatives. The first, Policy *B.1a* relies on the longitudinal dimension of our data, that include three complete fiscal cycles, i.e. for incomes produced in 2007, 2008, and 2009. Our data feature the universe of files in those years that were audited over the five-year cycle. In our exercise, the pool of audits eligible for discarding are thus the 2007, 2008, and 2009 files in the testing sample, whose audits occurred 1, 2, or 3 years after filing. The donor pool for replacement audits is 2007, 2008, and 2009 files in the testing sample, whose audits occurred 4 or 5 years after filing. These files in the

donor pool are a valid counterfactual because of three institutional features: they i) were available to audit at the time of decision for the files in our consideration sample, ii) had their tax file information locked in and so there is no risk of this information changing, and iii) were ultimately actually audited, and so have been selected into audit and have observable outcomes. In Policy *B.1a* we replace the discarded files with the best files from this pool according to the algorithm. By revealed preference, files that are available for an audit and overlooked in a given year are considered less promising by the IRA. Therefore, our suggested replacement procedure produces a lower bound of the expected gains. As a benchmark, in Policy *B.1b*, we replace the files with the best files (based on the predicted outcome) that were available for that year, even if they were never audited, and we impute their expected return using the ML prediction.

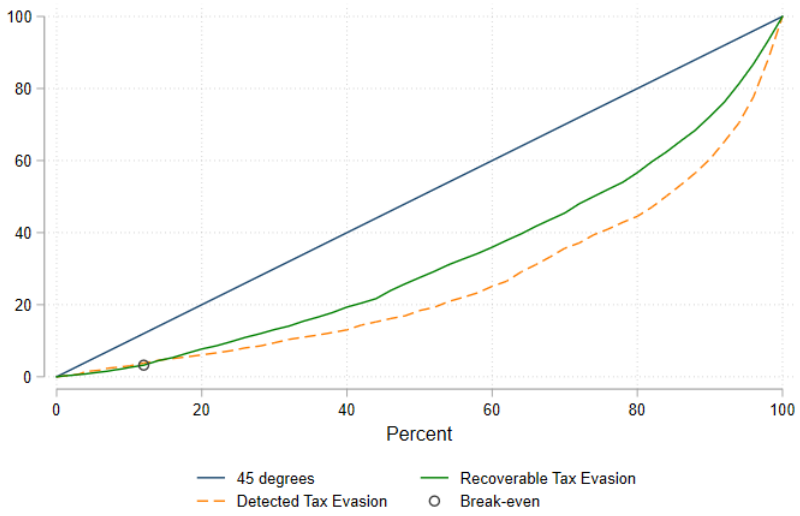
Policy B.2 instead does not require the use of a ML algorithm for the selection of the replacements but substitutes the discarded audits with randomly selected files from the pool of all audited files (Policy *B2.b*). The assumption here is that the authority has locally constant returns to auditing: if a small fraction of the audits is removed, the agency can replace them with a similar average marginal return. This assumption is motivated by the fact that the authority is severely constrained in terms of resources, thus able to audit only a tiny fraction of files. This problem resembles that of a top university selecting prospective students: the tiny fraction of accepted applicants is not dissimilar to the next 5% of rejected candidates. We also repeat the exercise by replacing with random files from the pool of later audits used in Policy B.1a (Policy *B2.a*).

## 5.1 Results

**Policy A: Discarding audits with low ex ante promise.** We consider how much *TaxEva* and *TaxGot* would be lost if the worst $X\%$ of audited files (sorted by the predicted outcome) were not audited. Because we observe the actual outcomes for these files, this is a straightforward direct calculation, and results in the Lorenz-type curves in Figure 2.[9] We present Lorenz-type curves for both *TaxEva* (solid line) and *TaxGot* (dashed line). For each percentage of discarded files, the curves show the percentage reduction in each outcome, respectively. The forty-five degree line depicts the loss in the aggregate outcome resulting from a random discarding of audits. The

[9] A Lorenz curve shows the share lost by discarding the lowest $X\%$ of *actual* outcomes. The Lorenz-type curves we present instead show the share lost by discarding the lowest $X\%$ of *predicted* outcomes.

"cost" of discarding audits with lowest-predicted outcomes is very small. Indeed, discarding the worst 10% of audits is associated with less than a 3% loss of detected tax evaded and 2.6% of recovered tax evasion.




This figure reports in the $y$-axis the percentage of the total amount of tax evaded and recovered tax evaded lost by not auditing in each office a given percentage of audits with the lowest predicted tax evaded and the lowest predicted recoverable tax evaded, respectively.

Figure 2
Losses from not auditing the files with lowest predicted evasion

Conversations with IRA officials informed us that an internal assessment of the cost per audit is around 1,700 euros.[10] This amount is similar to that reported as the cost to the United States' IRS at $2,278 per audit (Government Accountability Office, 2012). Our model predicts that each audit in the lowest 12% of audits recovers less than this amount of *TaxGot*. We plot a dot on the figure to indicate this break-even point. Of this 12%, the 50% generates exactly zero *TaxGot*. Eliminating audits below the break-even would reduce recovered tax evasion by 1,904,746 euros and costs by 2,789,700 euros (number of audits times 1,700 euros). Even ignoring the human and economic costs of audits on the target taxpayers, this policy would increase the net recovery by 884,954 euros.[11]

**Policy B: Replacing discarded audits under the ML guidance *vs.* at random.** One natural way to assess the opportunity cost of an audit is to focus on

[10]The IRA conducts routine audits, with standard costs.

[11]The basic monetary cost to the IRA is a lower bound to the actual cost. Actual costs also include the costs borne by the taxpayer in complying with the audit. Additionally, there may be social costs. For example, our data reveals that audited taxpayers are more likely to close their business during the three years following an audit than taxpayers who are not audited.

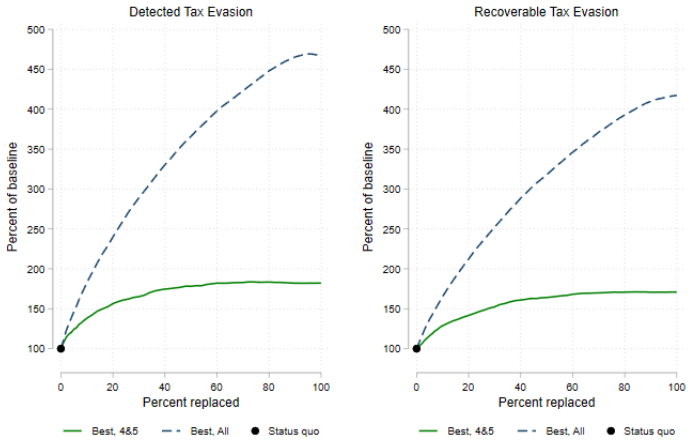
the benefit of alternative audits that could have been performed but were not. This idea leads to the modeling exercise of dropping the "worst" audits and then replacing them with an equal number of unaudited files. For Policy B.1, we sort the unaudited files according to the outcome (*i.e. TaxEva* or *TaxGot*) predicted by our model. Then we identify the best $N$ files from the set of unaudited files, where $N$ is the number of discarded files.

When considering potential replacements, we explore two variations. First, we deal with the selective labels problem by choosing the replacements among files from the same fiscal year as the discarded ones that were audited by the same office 4 or 5 years after filing (Policy B.1a). The solid-line in Figure 3, Panel A, shows very substantial gains of Policy B1.a of this "discard and replace" exercise on *TaxEva* (left panel) and *TaxGot* (right panel). The effects are expressed in percentage of the aggregate amount of *TaxEva* and *TaxGot* that is obtained by the IRA in the actual audits in our sample (*status quo*). At a 10% discard rate, this policy would increase the status quo aggregate *TaxEva* by 38% and the status quo aggregate *TaxGot* by 29%.

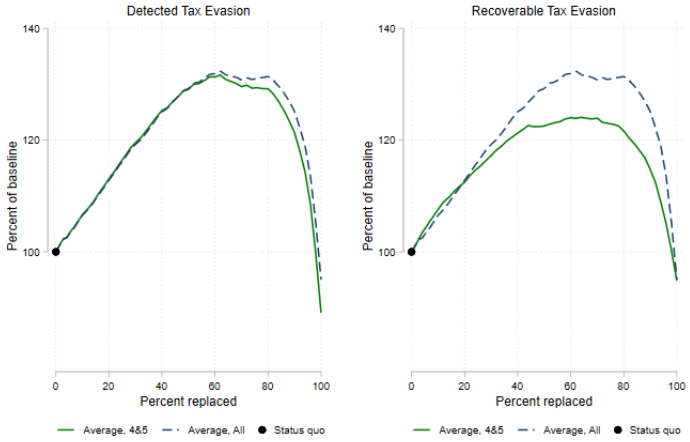
Second, we consider a more speculative approach for replacing discarded files (Policy B.1b). Namely, we choose the best files from the full set of unaudited files from the same fiscal year and office as those discarded, which in addition to files audited later includes the much larger set of files never audited. We use the predictions of our model to impute what the outcome would be. This strategy relies on the accuracy of the out-of-sample predictions from our prediction model at the top of the distribution. Differently from Policy B.1a we cannot verify this directly. Hence, this exercise potentially provides an upper bound to the gains from replacement. In Figure A2 we show that the large gain in replacement from this policy is mainly driven by the broader pools, since these larger pools have greater scope to identify promising files. As shown by the dashed line in Figure 3, Panel A, the theoretical gains of Policy B1.b are significantly higher than the gains in Policy B.1a, which is based on realistic counterfactuals. These gains amount to 83% of the status quo for *TaxEva* and 65% for *TaxGot*. Because Policy B.1b is more speculative, the actual gains are likely to be in between Policy B.1a and Policy B.1b.

Finally, to isolate the benefits of a ML-guided discarding from those of a ML-guided selection of replacements, we compare the gains of Policy B1 with Policy B2, where we consider replacing with an "average audit" taking as reference the office average. For ease of illustration, we replace the discarded files with an average file from the pool of audits, rather than showing Montecarlo simulations using random

*Panel A: Policy B.1 - Replace with the highest ranked files*




*Panel B: Policy B.2 - Replace with average files*




This figure reports on the $y$-axis the percentage gains of discarding a given percentage of files with the lowest predicted value of the target variable and replacing them with the same number of files *(i)* with the highest predicted value (panel A) or *(ii)* randomly selected (panel B). The $x$-axis reports the discarded percentage. The target variable is tax evaded in the left figures and recoverable tax evasion in the right figures. All values are reported relatively to the status quo total revenue of audits set at 100 and represented by a dot.

Figure 3
Gains from "discard and replace" exercise

drawings from this pool.[12]

In Figure 3, Panel B, the solid line represents Policy B.2a, that is when we consider the average among the "later audits" pool. Limiting the replacement to the lowest 10% leads to more modest but still significant gains: an overall 6.7% improvement relative to the status quo in *TaxEva* and 7.6% in *TaxGot*. The dashed line denotes

[12]Montecarlo simulations in the Online Appendix C.3 yield similar results.

Policy B.2b, that is when we allow for files audit at any age. The improvements are similar: 6.5% for *TaxEva* and 7.5% for *TaxGot*. This confirms that the average quality of files audited later is similar to that of the files that are audited at any time. These exercises show that the range of possible gain may be wide, but even adopting the most conservative measurements, any gain would be a meaningful improvement with respect to the status quo.

As within the IRA the selection of audits typically occurs at the local office level, our baseline results consider only replacements within the same local office.[13] However, in principle there could be greater gains from being able to replace from a broader pool. In the Online Appendix C.1, we consider replacement from files drawn within higher level IRA offices, i.e. the same province, the same region, or from anywhere in the country. While most of the gains are captured by replacing simply within the same office, reallocation at the higher level increases the performance further: with reference to the gains in *TaxEva* after a 10% replacement under Policy B.1a, reallocating within office implies a 38% gain, while reallocating at the province, region and country level increases *TaxEva* by 42%, 48% and 50%, respectively. This exercise suggests that some current organizational choices may be suboptimal under the machine selection.

## 5.2 Discussion: Policy Goals

**Tradeoffs between targeting *TaxEva* vs. *TaxGot*.** A key challenge in any policy prediction problem is that the specific goal of the policymaker is usually unknown. This is the "*omitted payoff bias*" problem. In our context, we were given direct information and thus designed algorithms tailored to best predict *each of the two* policymaker goals: the *TaxEva* and *TaxGot* outcomes. The next question is whether there are tradeoffs between these two measures. First, we note that a policy of discarding audits based on predictions for one measure can result in dramatically reduced improvements for the other measure. The Online Appendix D shows that the dominant predictors in the random forest differ by some extent among *TaxEva* and *Tax*Got. Also, among our pooled testing sample, the correlation coefficient between predicted *TaxEva* and predicted *TaxGot* is only 0.55. This implies that targeting tax files to discard (and/or replace) on the "wrong measure" can reduce the value of the exercise.

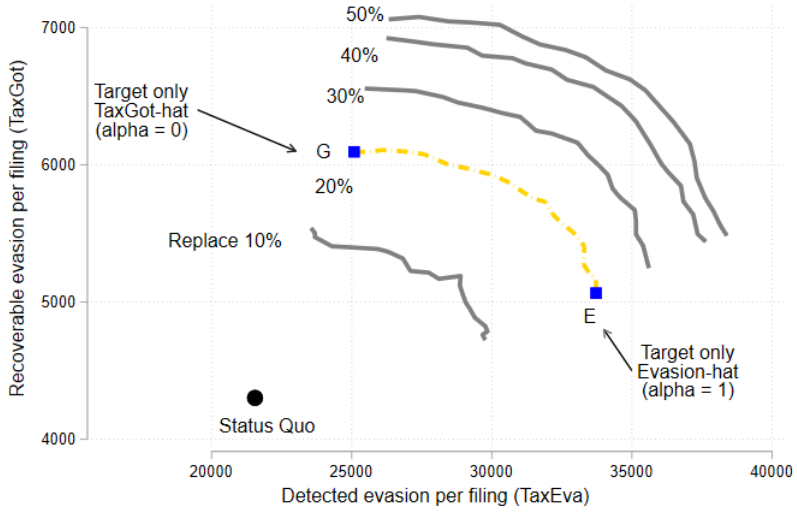
To gain insight into the tradeoffs across these outcomes, we model the auditors'

[13] There are 288 offices.

utility from choosing a given file into audit as a combination of the predicted *TaxEva* and predicted *TaxGot,*

$$U\left(\widehat{TaxEva}, \widehat{TaxGot}\right) = \widehat{TaxEva}^{\alpha} \cdot \widehat{TaxGot}^{1-\alpha}. \tag{1}$$

For each file, we compute this utility for a range of values $\alpha \in [0, 1]$. For each value of $\alpha$, we rank them according to their utility and implement our "discard and replace" exercise, where we replace with the best available files from the donor pool of files audited at later ages (Policy B.1a). Each line in Figure 4 shows the average *TaxEva* and *TaxGot* per file by discarding and replacing a given percentage of files, as indicated on the left end of each line. The dashed line corresponds to a 20% discard rate. When the utility of a file depends only on predicted *TaxEva* ($\alpha = 1$), our model results in point $E$; when it depends only on predicted *TaxGot* ($\alpha = 0$) our model results in point $G$. The other points on the line represent intermediate degrees of weighting the two policy goals.




This figure reports the detected tax evasion per file ($x$-axis) against the recoverable evasion per file ($y$-axis) after discarding the files with the lowest predicted utility and replacing them with the same number of files with the highest predicted utility at the office level (Policy B.1a). Each line reports a different percentage of discard-and-replace values and each point along aline represents different combinations of utility weights on the two policy goals, namely maximizing detected tax evasion and recoverable tax evasion. The status quo levels are represented by a dot.

Figure 4
Tradeoffs between detected and recovered tax evasion

We highlight two features of these results. First, and in line with Figure 3, Figure 4 shows that there is great scope for improving the selection of files. Second,

Figure 4 shows that when the policy objective is a simple function as in (1), even targeting either measure may dramatically improve both outcomes, compared to status quo (represented by a dot). However, there is some tradeoff between targeting only *TaxGot* versus targeting only *TaxEva*. Under a 20% replacement rate, targeting *TaxEva* only ($\alpha = 1$) produces a 56% increase over status quo in *TaxEva*, and a 18% increase in *TaxGot*. On the other hand, targeting *TaxGot* only ($\alpha = 0$) produces a 16% increase over status quo in *TaxEva*, and a 42% increase in *TaxGot*. In this sense, the tradeoffs are roughly symmetric across the two measures, in terms of percentage increase over status quo.

**Additional policy goals.** Pursuing the two main goals of reducing tax evasion and insolvency rate could have the unintended consequence of concentrating the audits among specific groups of taxpayers. For example, the policy may suggest to over-pick replacement files belonging to a specific sector or income class characterized by higher evasion rates. However, this could unintentionally hinder the deterrence effects for other groups of taxpayers. In the Online Appendix E, we repeat the discard and replace exercise by choosing replacement files only within deciles of taxable income or within business sectors. We show that, when replacing the files with the lowest 10% of predicted *TaxEva*, the gains of Policy B1.a and Policy B1.b (38% and 83%, respectively) are reduced to 22% and 49% when replacements of discarded files are limited to files in the same income decile as the discarded ones, and to 19% and 51% when replacements are limited to files in the same business sector. This evidence indicates that there are considerable gains from replacement even when keeping the composition of the audited files by key characteristics broadly unchanged.

To further investigate whether the new selection of files is undesirable along additional margins, we compare the average observable characteristics of the discarded and replaced files under these different replacement schemes. Table A3 in the Online Appendix shows that the ML algorithms select replacement returns filed by taxpayers with similar demographic characteristics to the ones who filed the discarded returns and managing similar businesses.

# 6 Concluding Remarks

This paper exploits an exclusive data-set from the IRA to explore the extent to which ML can be used to improve audit selection. We use the prediction model to calculate gains from a "discarding" exercise, where the audits with the worst predicted outcomes are discarded, and a "discard and replace" exercise, where these audits are discarded

and then replaced with audits from an alternative donor pool. The discard-only exercise shows that ML can be reliably used to identify poorly performing audits: in an out-of-sample analysis, we find that the audits with the lowest 12% outcome recover less than the material cost of performing the audit.

The "discard and replace" exercise is more delicate since it faces the key challenge that the actual outcomes for most files are typically not observed, because they did not receive an audit. We propose novel solutions to this challenge. Because the IRA has 5 years to audit a file, we develop a methodology where unaudited files audited at a later stage are used as counterfactuals. This allows to use files that were available at the time of the audit selection, but were neglected for inferior choices. Since these files were later audited, we can use later audits to assess their "value" and the performance of the replacement. We find that even if we restrict this replacement to files from the same office, very sizable performance improvements are feasible: at a 10% discard rate, selecting the replacements using ML from this pool yields an improvement of 38% of TaxEva, and of 29% of *TaxGot*. Allowing the replacements to be selected in the larger pool of unaudited files (and using the predicted value to evaluate them) yields much larger improvements: at a 10% discard rate, selecting the replacements using ML from the larger pool yields an improvement of 83% of *TaxEva*, and of 65% of *TaxGot*.

As a lower-bound, we also evaluate the potential improvements if the replacements are selected at random from the pool of audited files, both limiting the pool of replacements to files that were later audited and to the larger pool of audited files. The idea is that if only a small fraction of audited files is discarded and replaced, the tax agency can at least replicate its "average" performance with the replacements. In this case, the improvements are naturally smaller, but still significant: at a 10% discard rate, selecting the replacements at random yields an improvement of about 7% of *TaxEva*, and of 8% of *TaxGot*, independently of whether the pool is restricted or not.

A natural concern is whether selecting replacements with ML algorithm injects unintended bias in the selection. However, we do not observe significant differences in terms of key demographics characteristics of the tax filer and of the nature of their business.

# References


Ash, E., S. Galletta, and T. Giommoni (2021). A Machine Learning Approach to Analyze and Support Anti-Corruption Policy. *CESifo Working Paper Series 9015*.

Athey, S. (2017). Beyond Prediction: Using Big Data for Policy Problems. *Science 355*(6324), 483–485.

Bertrand, M., M. Bhatt, C. Blattman, S. B. Heller, and M. Kapustin (2022). Predicting and Preventing Gun Violence: An Experimental Evaluation of READI Chicago. mimeo.

Bhowmik, R. (2011). Detecting Auto Insurance Fraud by Data Mining Techniques. *Journal of Emerging Trends in Computing and Information Science 2*(4), 156–162.

Bonchi, F., F. Giannotti, G. Mainetto, and D. Pedreschi (1999). A Classification-Based Methodology for Planning Audit Strategies in Fraud Detection. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 175–184.

Bots, P. and F. Lohman (2003). Estimating the Added Value of Data Mining: A study for the Dutch Internal Revenue Service. *International Journal of Technology and Policy Management 3*(3/4), 380–395.

Breiman, L. (2001). Random Forests. *Machine Learning 45*(1), 5–32.

Cleary, D. (2011). Predictive Analytics in the Public Sector: Using Data Mining to Assist Better Target Selection for Audit. *Electronic Journal of e-Government 9*(2).

Government Accountability Office (2012). IRS Could Significantly Increase Revenues by Better Targeting Enforcement Resources. *GAO-13-151*.

Hsu, K., N. Pathak, J. Srivastava, G. Tschida, and E. Bjorklund (2015). Data Mining Based Tax Audit Selection: a Case Study of a Pilot Project at the Minnesota Department of Revenue. In *Real world data mining applications*, pp. 221–245. Cham: Springer.

Jung, J., C. Concannon, R. Shroff, S. Goel, and D. G. Goldstein (2017). Simple Rules for Complex Decisions. *Stanford University Working Paper*.

Kirkos, E., C. Spathis, and Y. Manolopoulos (2007). Data Mining Techniques for the Detection of Fraudulent Financial Statements. *Expert Systems with Applications 32*(995–1003).

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.

Knittel, C. R. and S. Stolper (2021). Machine Learning about Treatment Effect Heterogeneity: The Case of Household Energy Use. *AEA Papers and Proceedings 111*, 440–44.

Lakkaraju, H., J. Kleinberg, J. Leskovec, J. Ludwig, and S. Mullainathan (2017). The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (275-284).

Lakkaraju, H. and C. Rudin (2017). Learning Cost-Effective and Interpretable Treatment Regimes. *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS) 20th*.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica 46*(1), 69–85.

Ruan, J., Z. Yan, B. Dong, Q. Zheng, and B. Qian (2019). Identifying Suspicious Groups of Affiliated-Transaction-Based Tax Evasion in Big Data. *Information Sciences 477*, 508–532.

Wu, Y., B. Dong, Q. Zheng, R. Wei, Z. Wang, and X. Li (2020). A Novel Tax Evasion Detection Framework via Fused Transaction Network Representation. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 235–244.

# Online Appendix

# Refining Public Policies with Machine Learning: The Case of Tax Auditing

Marco Battaglini  Luigi Guiso  Chiara Lacava

Douglas L. Miller  Eleonora Patacchini

## A Data: Further Details

Our data includes the universe of files and audits to sole-proprietorship taxpayers. This is the category of tax filers that contributes the most to aggregate tax evasion in Italy, as well as in other countries. According to estimates of the Italian Treasury Ministry, in 2018 in Italy, small business (mostly registered as sole-proprietorship business) account for 60% of the total evasion detected from firms' reporting, an amount equal to 8.9 billion euros (Ministero dell'Economia e delle Finanze, 2019). The U.S. Internal Revenue Service (IRS) estimates that the lost federal tax revenue due to underreported individual income was 197 billion dollars in 2001 (18% of the individual income tax liability; U.S. Department of the Treasury, 2006). Johns and Slemrod (2010) report that in the U.S. 57% of self-employed income is misreported, in contrast to only 1% of wages and salaries. Similarly, Artavanis et al. (2016) document that in Greece, evasion by the self-employed accounts for large losses in the public budget.

Table A1 shows summary statistics. The sample includes 18,923,474 tax files of income produced in years 2007-2012 by 4,731,693 sole-proprietorship taxpayers. Among these tax returns, 403,691 returns filed by 304,726 different taxpayers receive an audit. The probability of receiving an audit over the five years after filing is close to 3% and similar across sectors, except for business operating in agriculture that have a much lower audit rate (1.2%).

Table A1
Summary Statistics - Audited Tax files

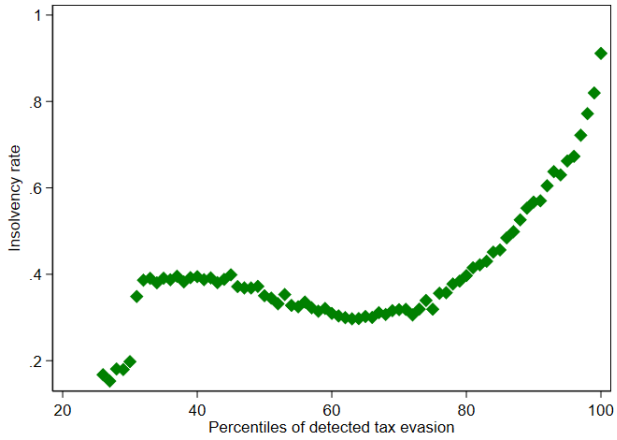
| | mean | median | sd | 10th pct | 90th pct |
|---|---|---|---|---|---|
| Audit rate | 0.029 | 0 | 0.169 | 0 | 0 |
| Agriculture | 0.012 | 0 | 0.107 | 0 | 0 |
| Trade | 0.032 | 0 | 0.175 | 0 | 0 |
| Construction and Manufacturing | 0.033 | 0 | 0.179 | 0 | 0 |
| Private services | 0.032 | 0 | 0.175 | 0 | 0 |
| Health, education, recreational services | 0.030 | 0 | 0.170 | 0 | 0 |
| Insolvency rate | 0.425 | 0 | 0.494 | 0 | 1 |
| Agriculture | 0.302 | 0 | 0.459 | 0 | 1 |
| Trade | 0.431 | 0 | 0.495 | 0 | 1 |
| Construction and Manufacturing | 0.570 | 1 | 0.495 | 0 | 1 |
| Private services | 0.379 | 0 | 0.485 | 0 | 1 |
| Health, education, recreational services | 0.213 | 0 | 0.410 | 0 | 1 |
| Appeal rate | 0.050 | 0 | 0.217 | 0 | 0 |
| Agriculture | 0.046 | 0 | 0.210 | 0 | 0 |
| Trade | 0.052 | 0 | 0.223 | 0 | 0 |
| Construction and Manufacturing | 0.045 | 0 | 0.208 | 0 | 0 |
| Private services | 0.051 | 0 | 0.221 | 0 | 0 |
| Health, education, recreational services | 0.040 | 0 | 0.196 | 0 | 0 |
| Positive evasion | 0.742 | 1 | 0.438 | 0 | 1 |
| Taxable income | 22,783 | 13,040 | 52,997 | 0 | 49,390 |
| Detected tax evasion ($TaxEva$) | 14,520 | 1,914 | 78,111 | 0 | 21,655 |
| Recovered tax evasion ($TaxGot$) | 3,436 | 0 | 12,556 | 0 | 7,952 |
| Years of activity | 13.276 | 11 | 10.412 | 0 | 29 |
| N. employees | 0.830 | 0 | 3.152 | 0 | 2 |
| Turnover | 85,047 | 33,985 | 2,571,116 | 3,014 | 168,474 |

*Notes.* The audit rate is calculated for complete fiscal cycles (2007, 2008, 2009 files). Financial accounts are expressed in euros. Detected and reported tax evasion are reported after winsorization.

Among audited files, the insolvency rate is 42.5%, and varies across sectors. About 57% of audited taxpayers operating in constructions and manufacturing (e.g. small construction firms, plumbers, artisans, bakers) do not pay back the detected evasion. Business providing services in trade and private services (e.g. lawyers, hairstylists, coffee shop owners, architects) have an insolvency rate of 43% and 38%, respectively. Finally, business in agriculture are insolvent 30% of the time and those providing health, education and recreational services (e.g. physicians, dentists) are insolvent 21% of the times. The probability that the taxpayer appeals the audit is on average 5%, with low variation across sectors: the appeal rate ranges between 4% -for health, education and recreational services- and 5.2% -for trade activities.

Audits detect evasion in 74% of the cases. The average detected tax evasion is 14,520 euros, with a quite dispersed distribution (min: 0, max: 21,884,085 before winsorization). Evasion is a substantial share of the taxable income declared: on average it amounts to 63% of the taxable income. The average recovered tax evasion is 3,436 euros.

The average audited taxpayer has been in operation for 13 years, only 24% of the business have employees and those with employees on average employee 3.4 workers. The average reported turnover is 85,047 euros with relevant heterogeneity (standard deviation: 33,985 euros; 90th percentile 168,474 euros), partly reflecting differences across industries.




This figure reports the insolvency rate for the percentiles of positive detected tax evasion.

Figure A1
Distribution of the insolvency rate by percentiles of detected tax evasion

Figure A1 shows the relationship between the detected tax evasion and the insolvency rate, as measured by the ratio between the number of audited tax files for which no payment is received in due time among those who are found with positive evasion. The figure shows a marked non linearity: the insolvency rate is much higher for files with high evasion.

# B Model Performance: Further Details

In this section, we compare the performance of our model with a standard linear (OLS) prediction model and investigate its robustness when using different sets of predictors. In Table A2, Panel A, we compare the predicting performance of our random forest model and a linear probability prediction model estimated using the baseline set of predictors. We measure the fit of the prediction model by comparing the predicted evasion and recovered evasion of files in the testing sample that were actually audited, with the realized values. We report the out-of-sample $R$-squared and the RMSE for both variables as a measure of fit. Overall, the $R$-squared values are low. This is not surprising, given the nature of the outcome variables: they are

highly right-skewed, and additionally have many 0 values. As Figure 1 shows, the random forest model predicts the conditional average very well. For these data, there is an inherently large degree of noise around the conditional average, which is reflected in the $R$-squared values. As we show in the paper, our prediction models can greatly improve detected evasion, even given these modest $R$-squared values. The table shows that the random forest model improves significantly the prediction of both outcomes: the $R$-squared of detected tax evasion doubles and the $R$-squared of the recovered tax evasion using the random forest model is 1.5 times that of the OLS model. This suggests that allowing for nonlinear functions of predictors is important to better predict detected and recovered tax evasion.

Table A2
Comparison across models

*A. OLS vs Random Forest*

| Model | N. predictors | N. obs. | Out-of-sample R-squared | | Out-of-sample RMSE | |
|---|---|---|---|---|---|---|
| | | | *TaxEva* | *TaxGot* | *TaxEva* | *TaxGot* |
| OLS | 255 | 40,869 | 0.041 | 0.055 | 95,329 | 13,498 |
| Random Forest (baseline) | 255 | 40,869 | 0.131 | 0.083 | 90,746 | 13,297 |

*B. Random Forest: Alternative Sets of Predictors*

| Predictors | N. predictors | N. obs. | Out-of-sample R-squared | | Out-of-sample RMSE | |
|---|---|---|---|---|---|---|
| | | | *TaxEva* | *TaxGot* | *TaxEva* | *TaxGot* |
| 5 dgt sector FE | 1,467 | 40,869 | 0.131 | 0.082 | 90,728 | 13,305 |
| 5 dgt sector FE + Mundlak's municipality | 1,469 | 40,869 | 0.131 | 0.082 | 90,732 | 13,303 |
| Mundlak's municipality & 5-dgt sector (baseline) | 255 | 40,869 | 0.131 | 0.083 | 90,746 | 13,297 |
| no Mundlak's controls | 251 | 40,869 | 0.127 | 0.081 | 90,952 | 13,310 |
| no detailed financial accounts | 150 | 40,869 | 0.092 | 0.080 | 92,750 | 13,319 |
| no province and 2-dgt sector FE | 20 | 40,869 | 0.072 | 0.064 | 93,772 | 13,433 |

*C. Random Forest: Alternative Testing Samples*

| Testing sample | N. predictors | N. obs. | Out-of-sample R-squared | | Out-of-sample RMSE | |
|---|---|---|---|---|---|---|
| | | | *TaxEva* | *TaxGot* | *TaxEva* | *TaxGot* |
| Estimation years, 2007–2009 (baseline) | 255 | 40,869 | 0.131 | 0.083 | 90,746 | 13,297 |
| Post-estimation years, 2010–2013 | 255 | 11,177 | 0.127 | 0.081 | 71,759 | 15,267 |

Another key decision that characterizes the prediction model is the preferred set of predictors. We use as our baseline set of predictors the full set of variables reported in the tax returns, combinations of those variables, and dummy variables at the province and at the 2-digit sector level (100 and 21 variables, respectively). In addition, we exploit the granularity of the geographical and sectoral information contained in the data using Mundlak-type predictors (Mundlak, 1978), defined as the average at the municipality and at 5-digit sector level of two key financial accounts, namely taxable income and turnover. In Table A2, Panel B, we test the sensitivity of our algorithm to alternative sets of predictors. First, as discussed above, we show that by substitut-

ing the Mundlak's controls with the 5-digit sector fixed effects (with or without the Mundlak's municipality variables) leads to a similar prediction accuracy. Second, we evaluate the explanatory power of different types of variables by eliminating different sets of controls in turn. When removing the Mundlak's controls the fit of the model changes only slightly, for both outcomes. When removing detailed financial variables there is instead an important reduction in the fit of the model for tax evasion, while the improvement for recovered tax evasion proxy is minor, suggesting that the full set of financial accounts does not help much predicting insolvency.[14] On the other hand, both tax evasion and insolvency shows a strong sectorial and geographical dimension: when removing sector and province fixed effects (last row) the performance of the model for both variables is dramatically reduced.[15]

# C Discard and Replace Exercise: Further Evidence

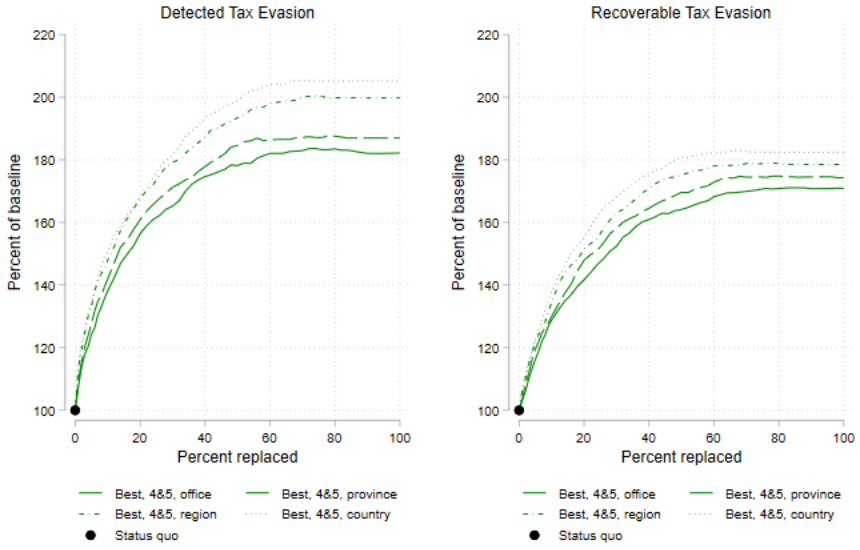
## C.1 Expanding the geography of the pool of available replacements

In this section, we repeat our baseline discard-and-replace exercise (Policy B.1a), allowing discarded and replacement files to come from a broader pool of audits. In other words, we envision a scenario in which higher-level organizational units of the IRA can impose replacement of files across tax offices. We consider replacement within the province, the region, and the whole country. For example, when the province is chosen as the organizational unit, we model dropping the 10% of audits with the lowest predicted outcome (among those that were audited 1-3 years after filing) within the province, and replacing them with an equal number of audits on the files with the highest predicted outcome among later audits (i.e. among those that were audited 4 or 5 years after filing) within the province. Results from these exercises are presented in Figure A2. We show two panels, one for each outcome, and use different line patterns for different replacement pools. The solid lines show the results from the baseline discard-and-replace Policy B.1a within an office, and

[14] A basic set of financial variables is included in all models. This set includes the reported taxable income (both the value and its logarithm), a dummy equal to one if the reported taxable income was positive, reported taxable income net of employees deductions, gross income, revenues, total assets, total liabilities, net value of production, VAT taxable turnover, total taxable revenues, operating costs, amortized costs, VAT transactions.

[15] We also experimented with the use of lagged variables as additional controls. Results show that the performance remains roughly unchanged. The lagged structure of the variables however substantially reduces the sample sizes.

gives the same results as in Figure 3, whereas dashed, dashed-dotted or dotted lines show the results for using province, or region as the organizational units (with region outperforming province), or the whole country. The picture shows that, for both outcomes, expanding the organizational level for discard-and-replace can improve the outcomes, but only modestly.



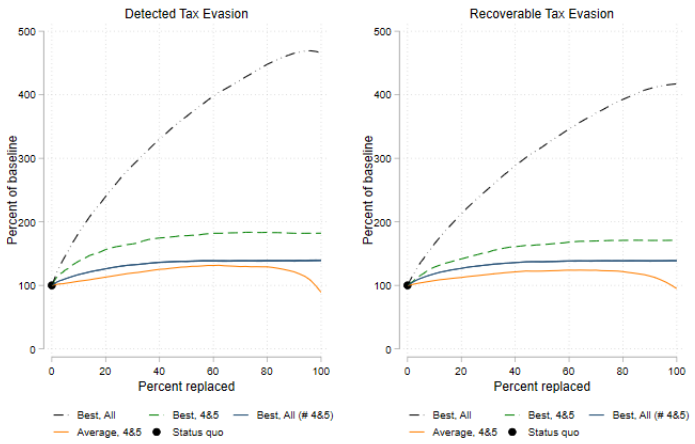

This figure reports the percentage gains of discarding the files with the lowest predicted tax evasion (left panel) or with the lowest predicted recoverable tax evasion (right panel), and replacing them with the same number of files with the highest predicted value at the local office (solid line) provincial office (dashed line), regional office (dashed-dotted line), and centralized-country level (dotted line). All values are reported relatively to the status quo total revenue of audits set at 100 and represented by a dot.

Figure A2
Discard and Replace at different organizational levels

## C.2 Expanding the size of the pool of available replacements

The gain in replacement from the main exercise with the broader pool (Policy B.1b) may be partly due to a higher probability of replacement with extreme values. To test this conjecture, we repeat our B.1b replacement exercise allowing for a random sample from a broader pool (which includes unaudited files) but with the same sample size as the pool of replacement files audited four and five years after filing. That is, we keep the replacement pool the same size as Policy B.1a. Figure A3 reports on this exercise when repeating the random draw 100 times. The 95% confidence interval is presented as a shaded blue region (however there is little variation across the 100 draws, so in

the figure this looks like a thick blue line). The figure shows that when limiting the size of the pool to the same size as Policy B1.a, Policy B1.b does worse than Policy B.1a. This is perhaps unsurprising since the files chosen to be audited (the 1a pool) are positively selected compared to the overall pool of files. The magnitude of this selection effect is moderate. On the other hand, the magnitude of the difference between the "full size" B.1b and the "small size" B.1b is very large. This indicates that the success of the B.1b Policy is driven by having a large pool of files from which to choose the best.



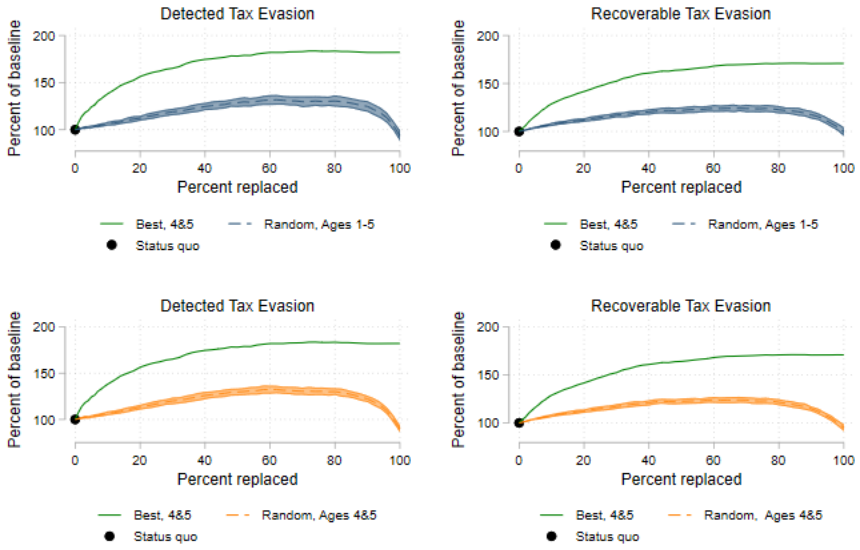

This figure reports on the $y$-axis the percentage gains of discarding a given percentage of files with the lowest predicted value of the target variable and replacing them with the same number of files with the highest predicted value using different replacement pools: *(i)* the pool of later audits (baseline Policy B1.a, green dashed lines); *(ii)* the pool of all unaudited files (Policy B1.b, black dashed lines); *(iii)* a random pool of all unaudited files that has equal size of pool *(i)* (blue solid lines). The orange solid line reports the gains of replacing the discarded percentage with files with outcome equal to the average value of those of pool *(i)* (Policy B2.a). The $x$-axis reports the discarded percentage. The target variable is tax evaded in the left figures and recoverable tax evasion in the right figures. All values are reported relatively to the status quo total revenue of audits set at 100 and represented by a dot.

Figure A3
Changing the size of donor pool impacts the gains

## C.3 Montecarlo experiment

In Section 5.1, we consider replacing with the average file from either all those files audited (Policy B.2b) or the average of those audited 4-5 years after filing (Policy B.2a). Because the "average file" is not a feasible direct choice, in this section we consider replacement with a random subset of actual files. We conduct 100 simulations, and in each one we draw a random subset of files as replacements. The 95%

confidence intervals for this exercise are reported in Figure A4. This figure shows that the results of these random draws are similar to "replacing with the average" shown in Section 5. It also shows that there is not a lot of variability across the random draws.



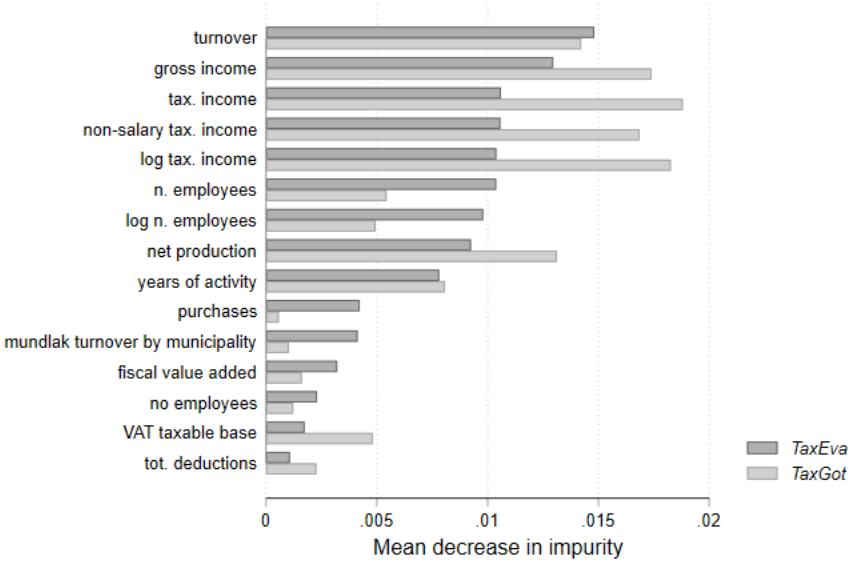

This figure reports on the $y$-axis the percentage gains of discarding a given percentage of files with the lowest predicted value of the target variable and replacing them with the same number of files with the highest predicted value using 100 random samples of the pool of later audits (first row, alternative implementation of Policy B.2a) and the pool of all unaudited files (second row, alternative implementation of Policy B.2b). The $x$-axis reports the discarded percentage. The target variable is tax evaded in the left figures and recoverable tax evasion in the right figures. All values are reported relatively to the status quo total revenue of audits set at 100 and represented by a dot.

Figure A4
Policy B.2 - Replacement with random draws

# D Predictors of Different Policy Objectives

In this section, we explore which variables are the main drivers of our prediction model. More specifically, we report the mean decrease in impurity (Gini importance) as described in Breiman (2001). At each split in each tree, the classification is performed by selecting, out of all splits of the candidate variables, the split that minimizes the Gini impurity. By doing so, we can measure how each predictor decreases the impurity of the split. The mean decrease in impurity is the sum over the number of splits (across all trees) that include the predictor, proportionally to the number of samples it splits. Figure A5 displays the predicting importance of the

predictors in our baseline model with the highest mean decrease in impurity, separately for each target variable. Notice that the main predictors for both variables are the different accounts reporting income (net, gross, business income excluding salary income), turnover, net production and VAT credits. The most relevant cost entries in predicting detectable and recoverable tax evasion are the purchases and imports relevant to the determination of the VAT tax base. Moreover, the years of activity and the business size (measured as the number of employees and its log) are demographic characteristics useful in predicting evasion. Interestingly, there is high variation in the predictive power of some predictors for the two target variables. This underlines the tradeoff between multiple targets when building a prediction model and complements our discussion in Section 5.2.



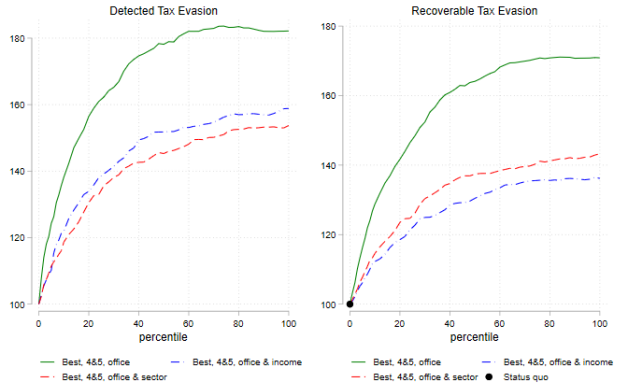

This figure reports the importance of predictors using mean decrease in impurity. The fifteen predictors with the highest maximum mean decrease in impurity in detected and recovered evasion are displayed.

Figure A5
Dominant predictors of tax evasion

# E Additional Policy Objectives

In this section, we consider a "constrained" discard and replace exercise. We focus on Policy B.1a, replacing discarded audits from the pool of files audited 4 or 5 years after file. The constraints we explore are either (1) income-decile constraints, or

(2) business sector constraints. To implement these constraints, we treat each file-year-by-office-by-income-decile as its own discard-and-replace pool. That is, if we are targeting a 10% discard rate, we do so for each file-year-office-decile pool, among those files that were audited 1-3 years after filing. And we replace within the same file-year-office-decile pool, using files audited 4-5 years after filing.




This figure reports on the $y$-axis the percentage gains of discarding a given percentage of files with the lowest predicted value of the target variable and replacing them with the same number of files selected with the highest predicted value (i) within the same office (baseline Policy B1.a, solid line); (ii) within the same office and income decile (dashed-dotted line); and (ii) within the same office and sector of activity (dashed line). The $x$-axis reports the discarded percentage. The target variable is tax evaded in the left figure and recoverable tax evasion in the right figure.

Figure A6
Discard and Replace within income deciles and sectors

Figure A6 shows the result of this exercise. The top green line is the unconstrained exercise Policy B.1a reported in the main text. The blue dash-dot line shows results when constraining replacements within the same income deciles, and the red dashed line shows results constraining replacements to be within the same sector. These lines show a similar profile of the gains of reallocation under the sector and income decile constraint to one another. They are each about one half as effective as the unconstrained exercise; but still represent a substantial improvement over status quo.

Table A3 shows that the proposed policies are not associated with a selection of files that differ systematically along additional margins. The table reports the average of a set of observable characteristics of the status quo selection of audits, and of the replacement files under the alternative versions of Policy B.1a replacement schemes mentioned above (i.e. the baseline replacements within office, replacements within office and sector, and within office and income decile), and the status quo (actual) selection of audits. Results for different target variables are reported in different

Table A3
Characteristics of files when discarding 20% and replacing with the best later audits

| | | *Panel A* Target: TaxEva | | | *Panel B* Target: TaxGot | | |
|---|---|---|---|---|---|---|---|
| | Status quo | Office | Office/Sector | Office/Income | Office | Office/Sector | Office/Income |
| Woman | 0.226 | 0.199 | 0.217 | 0.212 | 0.202 | 0.213 | 0.227 |
| Age | 2.695 | 3.049 | 2.985 | 2.980 | 3.058 | 2.982 | 2.984 |
| Married | 0.681 | 0.679 | 0.678 | 0.677 | 0.717 | 0.710 | 0.698 |
| Family business | 0.102 | 0.102 | 0.102 | 0.104 | 0.133 | 0.130 | 0.127 |
| Has employees | 0.425 | 0.495 | 0.469 | 0.464 | 0.504 | 0.486 | 0.479 |
| N. employees | 1.907 | 2.861 | 2.409 | 2.392 | 2.712 | 2.427 | 2.406 |
| Years of activity | 13.000 | 12.842 | 12.906 | 12.853 | 13.865 | 13.470 | 13.471 |
| *Sectors*: | | | | | | | |
| Agriculture | 0.036 | 0.035 | 0.036 | 0.037 | 0.041 | 0.036 | 0.041 |
| Trade | 0.334 | 0.314 | 0.334 | 0.325 | 0.315 | 0.334 | 0.338 |
| Construction & Manufacturing | 0.236 | 0.290 | 0.236 | 0.273 | 0.207 | 0.236 | 0.218 |
| Private services | 0.367 | 0.336 | 0.367 | 0.342 | 0.399 | 0.367 | 0.374 |
| Health & Education services | 0.027 | 0.025 | 0.027 | 0.024 | 0.038 | 0.027 | 0.029 |
| Taxable income | 26,370 | 36,199 | 31,438 | 28,847 | 47,256 | 35,660 | 30,127 |
| Turnover | 211,931 | 322,852 | 280,666 | 263,051 | 326,900 | 288,665 | 264,621 |

*Notes.* This table reports the mean characteristics of files (as indicated in the first column) after discarding and replacing 20% audits under Policy B.1a. Colums represent alternative sample selection, depending on the target variable and on the replacement pool. Column 2 ("Status quo") reports mean for the actual selection. "Office" indicates replacement with a later file in the same IRA office, "Office/Sector" and "Office/Income" allow replacement only with files of the same sector of activity and income decile, respectively. Financial accounts are expressed in euros.

panels (*TaxEva* in panel A, *TaxGot* in panel B). We find that the replacement tax files selected following the ML predictions are filed by taxpayers with similar demographic characteristics to the ones who filed the discarded returns irrespective of the replacement scheme (e.g. gender, age and marital status). Business characteristics are balanced too: the discarded and the replacement pools involve businesses that are comparable in terms of family business status, presence of employees, and years of activity. Perhaps interestingly, the sectorial composition of replacements and discard is not strikingly different even if balancing across sectors is not targeted (first column in both panels), which guarantees that deterrence effects mediated through the sectors are similar to those in the status quo. As expected, the striking difference between the replacement sample and the actual status quo sample emerges with respect to the financial accounts, since the level of evasion targeted by the algorithm is increasing in the business turnover. However, these differences are largely attenuated when replacing within the same income decile (third column in both panels), while maintaining balance in all other characteristics.

# References

Artavanis, N., A. Morse, and M. Tsoutsoura (2016). Measuring Income Tax Evasion using Bank Credit: Evidence from Greece. *The Quarterly Journal of Economics 131* (2), 739–798.

Johns, A. and J. Slemrod (2010). The Distribution of Income Tax Noncompliance. *National*

*Tax Journal 63* (3), 397–418.

Ministero dell'Economia e delle Finanze (2019). Nota di aggiornamento al DEF.

U.S. Department of the Treasury (2006). A Comprehensive Strategy for Reducing the Tax Gap. Office of Tax Policy.