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 machine learning 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 machine learning. Replacing the 10% least promising audits with an equal number selected by our algorithm raises detected tax evasion by as much as 39%, 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 from tax returns that can be used to identify audit targets. Consequently, the choice of auditing strategy is a prime

*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 Franco Peracchi, Edoardo Di Porto, Matteo Paradisi, and seminar and conference participants at the Italian Presidency of the Council of Ministers, Cornell University, ETH Zurich, University of Cambridge, Goethe University Frankfurt, the EUTO/IEB Workshop on the Economics on Taxation and the BSE Summer Forum in Public Economics for valuable discussions. This research was supported by a Cornell Center for Social Sciences Grant.

†Cornell University E-Mail: mb2457@cornell.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
‖Cornell University E-Mail: ep454@cornell.edu

candidate for applying machine learning techniques (henceforth, ML). The promise of these techniques is that they 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 audit selection 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 a novel dataset from the Italian Revenue Agency (henceforth, IRA) to explore whether ML techniques can be used to improve audit selection policies. The dataset includes tax returns from the universe of non-incorporated small businesses in Italy from 2007 to 2012. For these tax returns, we know whether or not they were audited, and the results of the audit, including information on whether the taxpayer appealed against the audit as well as all the statistical information available to the IRA concerning the tax return and filer.

The general idea behind ML techniques is to exploit data on realized outcomes to train a predictive algorithm. In our setting, the data is on audits that have occurred, and the outcome is, for example, detected tax evasion. Ideally, after validation procedures, the algorithm can be used to guide future policy (in our case, the choice of which returns to audit).

Even when detailed data is available, 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 tax returns that have been endogenously selected for audit are observed. In our setting, this would cause a problem if the IRA selects audits also relying on variables unobserved by the econometrician that are relevant for the audits' outcomes. The second problem is the *omitted payoff bias*. This refers to the fact that the policymakers' objectives may be multidimensional and unobserved, so an audit selection 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. In this paper, we make progress in evaluating the benefits of improving the audit selection process despite these two problems. Although our application is for one country and time period, the approaches we propose exploit common auditing data features. These features are that the tax authority is severely limited in the number of audits and that currently unaudited tax returns can occasionally be audited at a later date. Both of these features can be found in other environments, so our strategies can be applied in other contexts.

We start our analysis by documenting the extent to which an ML algorithm can

be used to identify audits that perform particularly poorly among the set of observed audits. Since we observe the universe of tax audits and relative outcomes, we can test our ability to identify the audits that perform poorly under various criteria. Contrary to other types of policymakers, tax authorities have a narrow policy mandate and do not have significant latitude in deciding their policy goals. The mission of the IRA is clearly set by the law that says that "The revenue agency is assigned the task of pursuing the maximum level of fulfillment of tax obligations both through assistance to taxpayers and through direct controls to combat non-compliance and tax evasion". Leaving aside the assistance to taxpayers, this directive translates into two goals: maximizing detected tax evasion and maximizing the amount of evaded taxes recovered.[1] The two goals may differ because while some audits may appear promising in detected evasion, the actual amounts that can be recuperated may be significantly lower, as taxpayers have the option to appeal against the audit. Our dataset allows us to assess both goals. We show that our ML algorithms can accurately rank audits based on both expected detected and expected recovered tax evasion. More importantly, we show that the audits that the ML algorithm predicts as having low evasion amounts result in low detected evasion. Eliminating the bottom predicted 10% of the audits would induce a reduction in detected evasion of only 3.1%. This would also induce a reduction of recovered evasion of only 2.8%. This suggests that the omitted payoff bias problem, while important in principle, may not alter qualitative conclusions in our setting.

Once we have identified the audits that detect zero or low evasion, the next question is whether we can replace them with tax returns with higher evasion levels. This is where the selective labeling problem starts to bite. To address this, we propose two complementary strategies. The first strategy relies on the longitudinal nature of our dataset to choose the replacement tax returns. In Italy (like in many other countries), the IRA has five years to audit a tax return. While most tax returns 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 tax return that is auditable but unaudited at t does not change for the following periods since it is based on income for a tax year preceding t. We can, therefore replace an

[1]The IRA mandate is defined by Legislative Decree July 30, 1999, n. 300, article 63. Our translation of the IRA mandate into actionable objectives was validated by IRA officials in direct consultations. Auditors do not have the authority to deviate from these goals. The law assigns to the IRA a few other tasks, listed in article 64. Namely, the management of the House Property Registry, the House Prices Observatory, and the task of providing estimates of property values to the public administrations when needed. These secondary tasks are irrelevant to the mandate specified in article 63.

audited tax return we predict to have low predicted evasion with a tax return that is available for auditing today but audited in the following years. We find that replacing the tax returns with the 10% lowest predicted evasion with an equal number of later audits with the highest predicted evasion yields an improvement of 39% in detected tax evasion.

The set of unaudited tax returns 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 tax returns with high predicted evasion. By not auditing a high-evasion tax return 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 any evaded income since some companies may dissolve or go bankrupt before the IRA can document a claim on the firm balance sheet. Empirically, we document that tax returns that remain unaudited for a few years but then are audited are not fundamentally different from other audited tax returns, in terms of both detected and recovered tax evasion.

The second strategy attempts to evaluate the replacements without using ML to select the replacement tax returns. The strategy relies on the fact that the IRA is severely constrained in terms of resources, so much so that only about 2% of the sole proprietorships' tax returns are audited (for comparison, in 2017, the audit coverage on personal income in France was 5%, and that of Earned Income Tax Credit recipients in the U.S. was 6%). If the authority were to eliminate from a list of proposed audits the bottom 10% of tax returns according to the predicted evasion and replace them with an equal (to fulfill the resource constraint) number of tax returns with random evasion, would the replacement be worthwhile? It is reasonable to assume that the replaced tax returns will not be different greatly from the average among audited tax returns for such a marginal substitution. We show that replacing the bottom 10% of predicted audited tax returns with average predicted evasion would increase detected tax evasion and recovered tax evasion by 6.5% and 7.4%, respectively.

The remainder of the paper is organized as follows. Section 2 revises the related literature. Section 3 introduces the institutional context. While Section 4 presents the statistical model and empirical strategy, Section 5 presents our policy experiments. Section 6 discusses extensions to the baseline policy experiments using alternative prediction models and considering both policy targets simultaneously. Section 7 concludes.

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).[2,3] These papers focus on the design of algorithms to predict an outcome of interest, restricting the evaluation of the algorithm performance to the quality of the out-of-sample predictions. Hence, they do not address the selective labels and omitted payoff bias problems. This limits the guidance that a policymaker interested in allocating scarce auditing resources can take from these studies (Athey, 2017). The selective labels problem is mitigated in cases of random allocation of treatments (e.g., random audits),[4] but policy interventions are almost never at random. We propose one solution to this question in the context of tax audit selection. This strategy 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.[5]

The importance of the selective labels problem for public policy applications is highlighted by Lakkaraju and Rudin (2017) and Jung et al. (2017), both studying how ML predictions can improve the judicial decisions to release or detain defendants while they await trial. To address the problem of missing labels for defendants whose judge's decision differs from the ML-guided decision, both papers adopt estimation methods relying on a "selection on observables" assumption. This assumption enables them to impute predicted values to observations with similar observable covariates that are

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

[3]In the context of Environmental Protection Agency (EPA) inspections, Hino et al. (2018) train an ML model to predict inspection failure probability. The model is then used to simulate reallocations of EPA resources.

[4]Recently, Ash et al. (2024) use randomized audits to evaluate the benefits of using an ML algorithm to predict corruption in Brazilian municipalities. In the context of gun violence prevention and energy consumption prediction, Bhatt et al. (2024) and Knittel and Stolper (2021) rely on randomized control trials.

[5]Using a sample of about 300,000 firms in Delhi, out of which 538 are known to be fraudulent, Mittal et al. (2018) deal with a different type of labeling bias problem; that is the case where only the audits that determine that a firm is fraudulent are observed. They predict the probability of being a fraudulent firm using an ML model where audits determining a firm is legitimate are labeled the same as unaudited firms, and exploiting the fact that the type of the firm (either fraudulent or non-fraudulent) is assumed to remain constant over time. The revenue-saving potential of the predictive model is then estimated on the firms at the top of the predicted ranking, which are the firms that are most likely to be fraudulent.

missing observed outcomes. To do so, they use either a two-step strategy combining an inverse propensity-score weighting and a logistic regression, or a regularized logistic regression model, respectively. In the same context of judicial decisions, Lakkaraju et al. (2017) and Kleinberg et al. (2018) follow a different approach and rely on the 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 the propensity to observe the labels, we use the feature that some labels are only revealed later in time.

3 Institutional Setting

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 (see Appendix A for further details). We merge information from two different administrative records that the IRA shared with us: tax return records and audit records. Records are at the individual level and cover statements of incomes generated from 2007 to 2012, reported between 2008 and 2013, and audited between 2009 and 2014. In Italy, similar to several other nations, the IRA has a five-year window to examine a tax return. In our primary analysis, we use the tax returns from 2007-2009, for which we can observe the entire audit window. We use the returns from 2010-2012 as an alternative testing sample.[6] Our sample contains around 19 million tax returns filed by almost 4.7 million taxpayers and 257,701 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

[6]See Section 4, and Appendix B.

tax return was audited and the amount of evaded tax assessed (if any).[7, 8] An audit typically spans approximately 2 to 5 months (interquartile range) and culminates in an assessment of any detected evasion, if present. However, the process of recovering unpaid taxes, once evasion is discovered, can be considerably lengthy. One advantage of our dataset lies in the IRA providing novel information regarding the taxpayer's response to the audit. Specifically, the IRA shared with us the information on whether the audited taxpayer does not pay back the assessed evasion. This case can originate from insolvency, an appeal against the audit, or simply a failure to respond to the audit notification. It triggers complicated processes, which last for many years, entail complex sanctions schemes, and may involve several layers of the judiciary. In Italy, an average of 7% of audited taxpayers appeal against an audit, and another 37% of taxpayers neither pay nor appeal within due time after the audit notification. IRA officials explained to us that the mandate of the IRA entails not only maximizing the identification of evasion but also maximizing the recovered amount from the identified evasion.[9] However, because of the challenge of accurately estimating the complete revenue, we adopt a cautious approach. In cases where taxpayers appeal, provide no response, or make partial payments following the audit, we conservatively set the value of the recovered evasion to zero (about 50% of the cases). Specifically, we tailor our statistical model around the two goals of the IRA: detecting evasion and recovering the detected amount of evaded tax. For each tax return, our two main outcome 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 delinquent. Further details on our data and summary statistics are reported in Appendix A.

[7] Audits typically target a single tax return. In our data, only 7% of taxpayers are audited more than once. In the paper, we treat each audited tax returns as independent. Results are unaffected if we add a control variable indicating that a taxpayer has been previously audited in our data. We focus on the specification without this control because, for tax returns earlier in time, we do not have a complete history of the audit. This means that that the indicator has a different meaning depending on the filing year.

[8] We identify and exclude audits initiated by authorities cooperating with IRA (3%) that might use other information or selection criteria. These authorities are the *Guardia di Finanza*, a military police force in charge of dealing with a wide range of criminal acts, including fiscal fraud, and the Customs Authority, in charge of monitoring tariffs and international trade taxes.

[9] Figure A1 shows that the probability of delinquency increases with higher levels of tax evasion, which seriously hampers the ability of the IRA to recover evasion.

4 The Machine Learning Algorithm

We propose an ML algorithm to create predictions for all tax returns, and then use these predictions to guide the audit selection process. In this section, we describe the prediction model we use. In the next section, we show how these predictions can be used to identify audits that perform particularly poorly and eventually replace them with audits with higher expected outcomes.

Model. We propose a random forest model $\varphi^k(\cdot)$ that uses a vector of predictors Z and yields for each tax return i a prediction of an outcome k :

$$\hat{y}_i^k = \varphi^k(Z_i). \tag{1}$$

This model allows for rich interactions among explanatory variables and easily adapts to non-linearities (Breiman, 2001). Our random forests contain 1,000 trees each.[10] 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.[11]

Predicted outcomes ($\hat{y}_i^k$)**.** We predict two outcome variables k, the detected and recovered evasion levels. We train a different model for each of them. We winsorize each outcome variable at the top 1 percentile to prevent our model chasing extreme and idiosyncratic observations.[12] Specifically, we consider the top 5% of the variable and estimate the parameters of a Pareto distribution to fit the distribution of those values. We then use the estimated distribution to compute the conditional mean of being in the top 1 percentile and impute this conditional mean to all outcomes in the top 1%.

Predictors (Z_i)**.** 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 the taxpayer being self-employed), and the full selection of financial variables included in

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

[11] 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.

[12] Two additional features of our analysis guard against mistaken conclusions driven by outliers. First, we evaluate the models and perform our policy experiments using the testing sample. This sample is completely different from the data used to train the model. If the predictions were tainted by overfitting to extreme values in the training data, this would result in a very bad fit in the testing sample. Second, regarding the risk of outliers among the predictors, the random forest prevents assigning excessive weight to idiosyncratic values of some predictors by selecting different features available for each branch of each tree in the forest and averaging over those trees. In addition, we have also explored models using the inverse hyperbolic sine transformation of our predicted variables. Working with these specifications did not produce better predictions, and we focused on predicting expected evasion in levels.

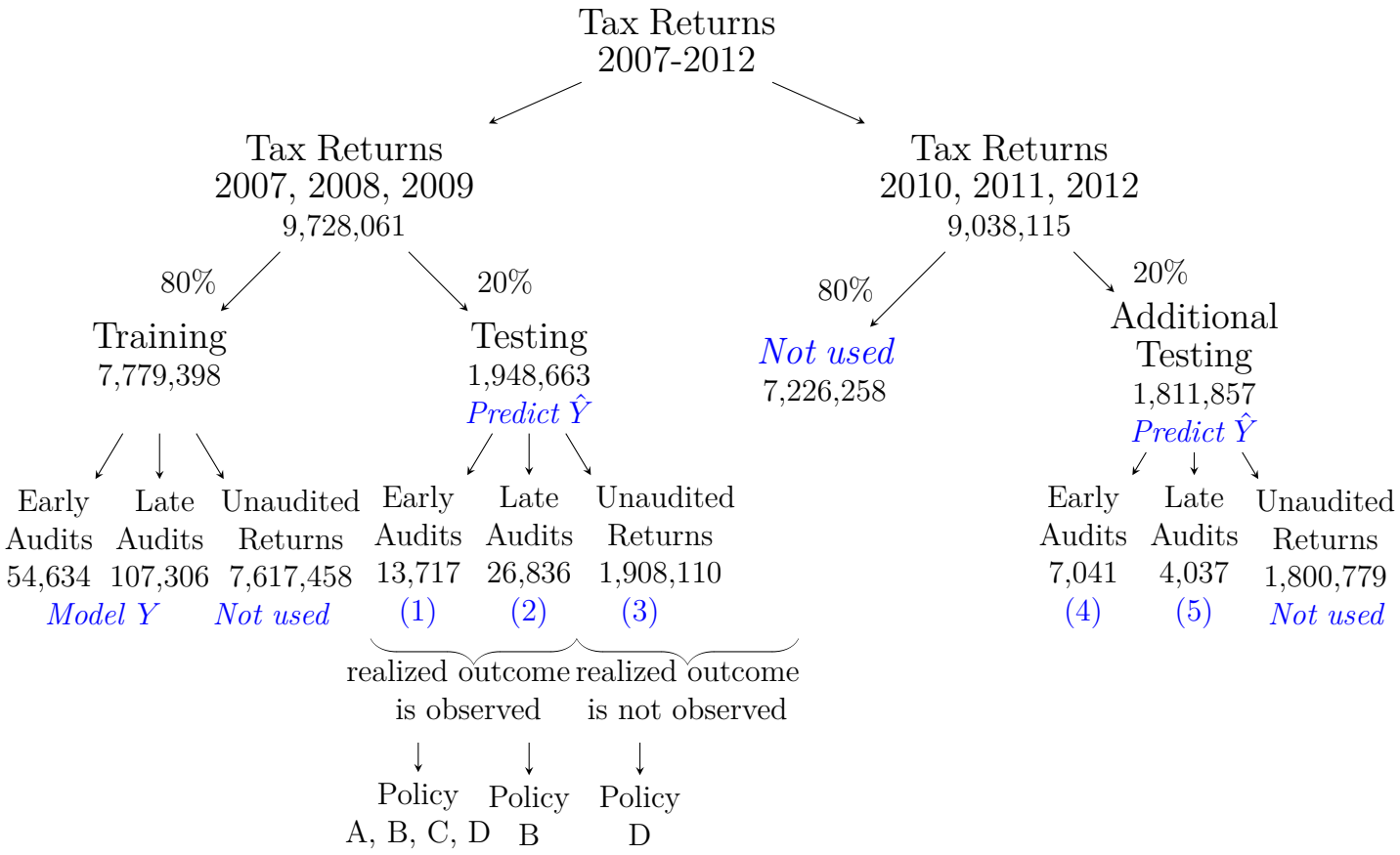
the Tax Registry. These include the reported taxable income (both the value and its logarithm), a dummy indicating a positive reported taxable income, reported taxable income net of employees deductions, gross income, revenues, taxable revenues, total assets, total liabilities, the net value of production, VAT taxable turnover, operating costs, amortized costs, and VAT transactions.

We include geographical fixed effects at the province level (110 provinces) and 2-digit-level sector of activity fixed effects. 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. Instead, we use this information as follows. First, we explored a specification with the 5-digit sector of activity fixed effects and geographic 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: taxable income, and turnover. We add these variables to the 5-digit sector of activity fixed effects (specification *ii*). Next, we exclude the 5-digit sector of activity fixed effects and use Mundlak-type predictors at both municipality and 5-digit sector levels (specification *iii*), while keeping fixed effects at the province level and at the 2-digit sector level. We discuss the sensitivity of our predictions against alternative sets of predictors in Appendix B. Because the performance of the different models is roughly equivalent, we use as a baseline the more parsimonious version of the model, featuring roughly 250 variables (specification *iii*).[13]

Sample. In our primary analysis, we use tax returns from 2007, 2008, and 2009, for which we observe the complete fiscal cycle - the following 5 years in which they can receive an audit. By doing so, we get a representative sample of the composition of the tax returns audited over time.[14] To get a clean partition of the data, our randomized

[13]In principle, random forest models can handle categorical variables, thus reducing the dimensionality of the predictor set. However, in our case, the orderings of values of our categorical variables (identifiers for geographical location and sector of activity) do not have an economic interpretation or systematic statistical relation. While there are several possible pre-processing strategies, they require additional assumptions on how to order the variables. Our final model specification iii) is a compromise between complexity and parsimony. In this specification, we reduce the dimensionality of the model by using Mundlak controls (group-level average of taxable income, and of turnover) based on the municipality identifier variable and the 5-digit sector of activity, but we keep the province and aggregate sector of activity as dummies.

[14]Our sample was not strongly affected by the Great Recession. For sole proprietors in Italy, a notable decrease in turnover occurred during a subsequent period, aligning with the sovereign debt



Measures of fit: (1) + (2) + (4)+ (5)

Policies:
Policy A: discard from (1)
Policy B: discard from (1), replace from (2)
Policy C: discard from (1), replace with average from (2)
Policy D: discard from (1), replace from (2) + (3), use $\hat{Y}$ to impute Y for (3)

Fig. 1. Partition of tax returns data into samples used for predictions and policy exercises. Notes: This figure illustrates the samples used for predictions and for policy exercises. At each node, the sample size is indicated. Early audits are those occurring 1-3 years after the tax filing year, and late audits are those occurring 4-5 years after.

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 prediction model consists of

crisis that impacted European countries after 2011.

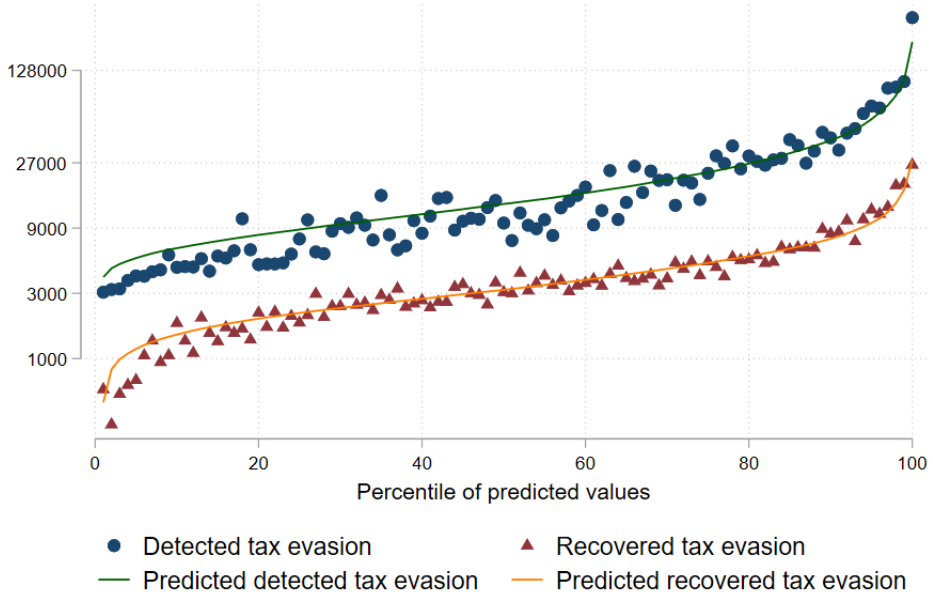


Fig. 2. Model fit.
Notes: This figure reports the detected tax evasion (dots) and the recovered tax evasion (triangles) of realized audits in the testing sample by percentiles of predicted values. The green and the orange lines display the predicted detected and recovered tax evasion, respectively. The sample includes tax returns of income produced in years 2007-2012 that are audited by IRA. The y-axis is represented on a logarithmic scale.

an 80% subset of the universe of audited taxpayers. To assess the goodness of fit, we compare predictions with observed outcomes for our testing sample, which consists of the 20% of returns from each of the fiscal years 2007-2009 that were excluded from the training sample. Our training sample contains 161,940 tax returns, and our testing sample has 40,553 tax returns. We illustrate our data partitioning in Figure 1.

Out-of-sample assessment of prediction model. To measure the model performance, we compare predicted outcomes with outcomes assessed during the audit in our testing sample. In Figure 2, the curved green line reports the average predicted *TaxEva* at each percentile of predicted *TaxEva* and the blue dots report the average *TaxEva* 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 random forest algorithm can predict the actual levels of both outcomes quite well along the entire distribution.[15]

[15]Figure 2 is computed using audited testing returns, pooling the years 2007-2009 and 2010-2012. The data partitioning is illustrated in the gray box of Figure 1.

Table 1
Comparison across prediction models.

A. Alternative Methods						
Model	N. predictors	N. obs.	Out-of-sample R-squared		Out-of-sample RMSE	
			TaxEva	*TaxGot*	*TaxEva*	*TaxGot*
Random Forest†	255	40,553	0.131	0.083	90,032	13,213
OLS	255	40,553	0.060	0.068	93,629	13,324
LASSO postselection	255	40,553	0.052	0.057	94,004	13,403

B. Augmented OLS and LASSO Specifications						
Model	N. predictors	N. obs.	Out-of-sample R-squared		Out-of-sample RMSE	
			TaxEva	*TaxGot*	*TaxEva*	*TaxGot*
OLS interactions	310	40,553	0.054	0.068	93,908	13,321
OLS polynomials	275	40,553	0.070	0.062	93,105	13,364
LASSO interactions	310	40,553	0.057	0.060	93,769	13,381
LASSO polynomials	275	40,553	0.042	0.054	94,538	13,423

Notes: This table reports measures of fit computed on the testing sample. † indicates the baseline specification. The training sample includes 161,940 tax returns audited between 2009 and 2014. The LASSO estimator selects 89 non-zero predictors for detected tax evasion and 19 non-zero predictors for recovered tax evasion. The LASSO estimator with interactions selects 130 non-zero predictors for detected tax evasion and 21 for recovered evasion; the LASSO estimator with polynomials selects 74 non-zero predictors for detected and 16 for recovered evasion.

In Table 1 (Panel A, first row), we present the out-of-sample R-squared and RMSE for both variables as summary metrics of predictive accuracy of the random forest.[16] The R-squared values are 0.131 for $TaxEva$ and 0.083 for $TaxGot$, while the RMSE values are high (90,032 for $TaxEva$ and 13,213 for $TaxGot$) compared to the means of the outcome variables. The modest R-squared values are unsurprising given the characteristics of the outcome variables, which exhibit high dispersion, significant right-skewness, and a considerable number of zero values. Figure 2 shows that the random forest model predicts the conditional averages very well. Nevertheless, Table 1 reveals that there exists a notable amount of inherent noise around these averages, contributing to the observed R-squared and RMSE values. Despite the relatively low out-of-sample R-squared, in the next section we find that the model predictions can be used to improve the audit selection policy outcomes substantially.

Appendix B provides additional tests of the validity of the prediction model. We show that its performance persists well for different testing samples, and on returns filed in later years, and that the fit is similarly good for both early and late audits.

Fit of random forest versus linear models. In Table 1, Panel A, we also report the predicting performance of an OLS prediction model and a LASSO postse-

[16]These measures of fit are computed using the same samples as Figure 2 (see the grey box in Figure 1).

lection model estimated using the baseline set of predictors.[17]

The R-squared of detected tax evasion predicted using the random forest model (RF hereafter) is twice the one predicted using the OLS and LASSO models, with values of the RMSE pointing towards a qualitatively similar improvement in performance. The R-squared of recovered tax evasion using the RF model is approximately 1.2 and 1.5 times that of the R-squared for OLS and LASSO, respectively; with similar values for the RMSE across models.

The most natural explanation for the better predictive fit of the RF model is that the RF model allows for nonlinear functions of the predictors. To explore this possibility, we identify the ten most important predictors in our OLS models as those variables with the highest t-statistic. We then consider alternate models that include additional features such as quadratic and cubic terms in these important predictors, or alternatively, include interaction terms between these important predictors. We then re-estimate OLS and LASSO with these enhanced feature sets. Table 1, Panel B, shows the results of this investigation. Overall, the performance of these alternate models is still far from the accuracy achieved by the RF. In the best case (OLS polynomial for $TaxEva$), the R-squared is 0.070, which is still almost half of the one for the RF model (0.131). In Section 6, we explore how these differences in predictive fit and other pertinent features translate into gains concerning the audit selection policy.

Relevance of predictors. For completeness, to understand how the RF predictions depend on the different predictors, we compute the Shapley values for each observation, following Štrumbelj and Kononenko (2010) and Lundberg and Lee (2017). We then summarize these values using Shapley shares (as in Joseph, 2020), and present mean Shapley shares as a measure of feature importance. The main predictors for both outcome variables are different types of reported income (total, gross, taxable) and income source (real estate, professional, compensations), turnover, operating costs, and taxes (on purchases and imported intermediate goods, credit, withholdings). We provide more details in Appendix C.

[17] We estimate LASSO linear models using a five-fold cross-validation selection method. This estimator selects 89 and 19 predictors with non-zero coefficients for detected and recovered tax evasion, respectively. We then use postselection coefficient estimates for predictions, to compare with our other prediction models. Penalized coefficient estimates show similar predictive performance (R-squared equal to 0.055 for $TaxEva$ and 0.052 for $TaxGot$).

5 Policy Experiments

We consider four policy experiments to quantify the benefit of using ML to refine the audit selection process . Let Ω_c be the set of tax returns reporting the income produced in year $c \in \{2007, 2008, 2009\}$. We describe them below. They are also illustrated in Figure 1. Let us denote the year in which income is produced (c) as a cohort. Each tax return in Ω_c is assigned to a local tax authority office $o \in O$, which decides whether to audit it. Let $\Omega_{o,c}$ denote the subset of Ω_c containing the tax returns of cohort c that can be audited by office o, with element $i \in \Omega_{o,c}$. The set $\Pi_{o,c} \subset \Omega_{o,c}$ is the set of audited tax returns and contains the subsets $\underline{\Pi}_{o,c}$ and $\overline{\Pi}_{o,c}$, which are the sets of tax returns of cohort c audited by office o by the third year after filing (early audits) or after (late audits), respectively. Each office selects the audits according to a mapping rule that is unknown to the researcher but constant over time.[18] The total amount of outcome $k \in \{TaxEva, TaxGot\}$ obtained by an office from early audits is $Y^k = \sum_{o \in O} \sum_{c=2007}^{2009} \sum_{i \in \underline{\Pi}_{o,c}} y_i^k$, where y_i^k indicates the outcome value for tax return i.

We begin our analysis by ranking the tax returns in a set $S \in \left\{\Omega_{o,c}, \Pi_{o,c}, \underline{\Pi}_{o,c}, \overline{\Pi}_{o,c}\right\}$ by their predicted outcome. Let us denote with $q_S^{k,x}$ the x percentile of $\hat{y}_{i \in S}^k$ for $i \in S$, and $\underline{N}_S^{k,x}$ the number of tax returns in set S whose predicted outcome is lower than $q_S^{k,x}$. In particular, consider the number of early audits whose predicted outcome is lower than $q_{\underline{\Pi}_{o,c}}^{k,x}$, denoted $\underline{N}_{\underline{\Pi}_{o,c}}^{k,x}$, and indicate with $T_S^{k,x}$ the highest threshold t such that there are at least $\underline{N}_{\underline{\Pi}_{o,c}}^{k,x}$ returns with prediction $\hat{y}_i^k$ higher than t in S:

$$T_S^{k,x} = max(t \text{ such that } |S^{\hat{y}_i^k \geq t}| \geq \underline{N}_{\underline{\Pi}_{o,c}}^{k,x})$$

where the modulus function indicates the number of elements in a set, and we denote $S^{\hat{y}_i^k \geq t} := \{i \in S \text{ s.t. } \hat{y}_i^k \geq t\}$.

Policy A: Discarding audits with low ex-ante promise, no replacement. In the first experiment (*Policy A*), we use the ranking based on the predictions to calculate the loss from a "discarding" exercise, where the x percent of early audits with the lowest predicted outcome are discarded. The total outcome of Policy A is:

$$Y_A^{k,x} = \sum_{o \in O} \sum_{c=2007}^{2009} \sum_{i \in \Phi_{o,c}^A} y_i^k. \tag{2}$$

[18] Battaglini et al. (2019) provide a description of the determinants of the audit selection by the IRA.

where for future reference we denote $\Phi^A_{o,c} = \underline{\Pi}_{o,c}^{\hat{y}_i^k \geq q^{k,x}_{\Pi_{o,c}}}$. We compute the loss from Policy A as one minus the ratio $Y_A^{k,x}/Y^k$ for each percentile x, and we illustrate this ratio as it changes with x in a Lorenz-type curve. Given that audits have administrative, human, and economic costs on the target taxpayer (e.g., due to psychological distress and/or interference with the business) as well as for the IRA, cutting audits with zero or minimal outcome might lead to a net gain for the x percent of discarded audits because the loss from not auditing is lower than the cost of the audit (even without replacing them with audits with higher predicted evasion). More specifically, we consider how much *TaxEva* and *TaxGot* would be lost if the x% of audited tax returns with the lowest predicted outcome were not audited. Because we observe the actual outcomes for these tax returns, this is a straightforward calculation and results in the Lorenz-type curves in Figure 3.[19] We present Lorenz-type curves for both *TaxEva* (green line) and *TaxGot* (orange line). For each percentage x of discarded tax returns, the curves show the percentage reduction $(100 \cdot (1 - Y_A^{k,x}/Y^k))$ in each outcome, respectively. For example, the green line shows that if we were to discard the lower 40% audits based on their machine-learning predicted tax evasion, the actual amount of tax evasion that would be lost is around 20%. The forty-five-degree line depicts the percentage reduction in each outcome (the loss) resulting from a random discarding of audits - discarding 40% of audits at random yields a 40% reduction in total evasion detected and total evasion recovered. It is linear with a slope of one because discarding at random involves no selection. The average value of randomly discarded audits regarding evaded income (recovered evasion) is independent of x. The Lorenz-type curves show that the loss in terms of each outcome of discarding audits with the lowest-predicted outcomes is very small. Indeed, discarding the worst 10% of audits is associated with less than a 3.1% loss of detected tax evaded and 2.8% of recovered tax evasion.

The IRA conducts routine audits at standard costs. Conversations with IRA officials informed us that an internal assessment of the cost per audit is around €1,700. This amount is similar to that reported as the cost to the United States IRS at $2,278 per audit (Government Accountability Office, 2012). Sorting by predicted *TaxGot*, the lowest-ranked 9% of audits recover less than this amount. We plot a dot on the figure to indicate this break-even point. Of this 9%, 80% generates exactly zero *TaxGot*. Eliminating audits below the break-even would reduce recovered tax

[19] 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.

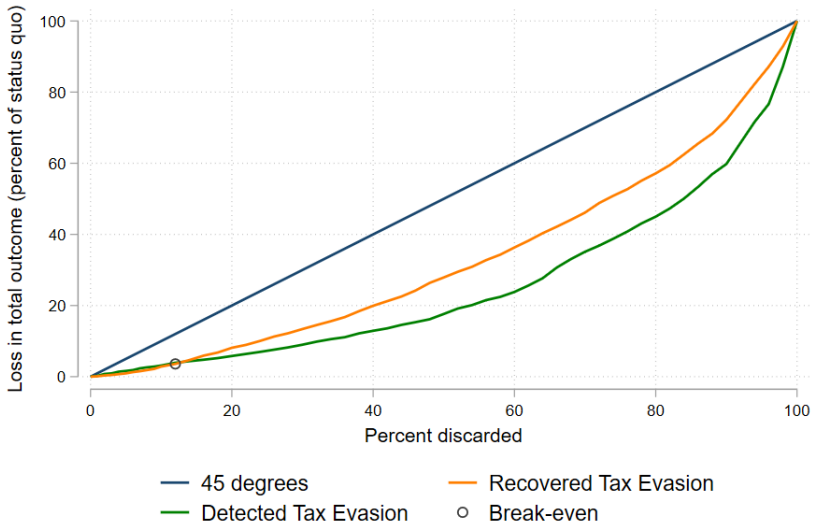


Fig. 3. Loss from discarding the lowest predicted evasion audits.
Notes: This figure reports in the y-axis the percentage of the total amount of tax evaded (green line), and recovered tax evaded (orange line) lost by not auditing in each office a given percentage of audits with the lowest predicted tax evaded and the lowest predicted recovered tax evaded, respectively.The x-axis reports the discarded percentage. All values are reported relatively to the status quo total tax evaded or recovered tax evaded by actual audits conducted by the third year, set at 100 and represented by a dot.

evasion in the testing sample by €1,289,386 and costs by €2,097,800 (number of audits times €1,700). Even ignoring the human and economic costs of audits on the target taxpayers, this policy would increase the net recovery by €808,414. Given that our testing sample is a random 20% extraction of audits over three years, this implies an annual net recovery for the full population of approximately €1,347,357 (our estimate multiplied by 5 and divided by three years).[20] In our following experiments, we focus on strategies to replace the discarded audits with others. In those experiments, the total number of audits and thus total costs are held constant.

Policy B: Discarding audits with low ex-ante promise, replacement using ML guidance. In the second experiment (*Policy B*), we use the ranking based

[20]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. Boning et al. (2023) suggest that the average cost of an individual audit (which might differ from the cost of auditing a sole proprietorship tax return) is roughly $6,000 with all costs accounted for. Higher audit costs imply both a higher proportion of discarded returns and a greater avoided cost per discarded return. Therefore, in our context, underestimating the audit costs would mean that our results represent the minimum potential gains.

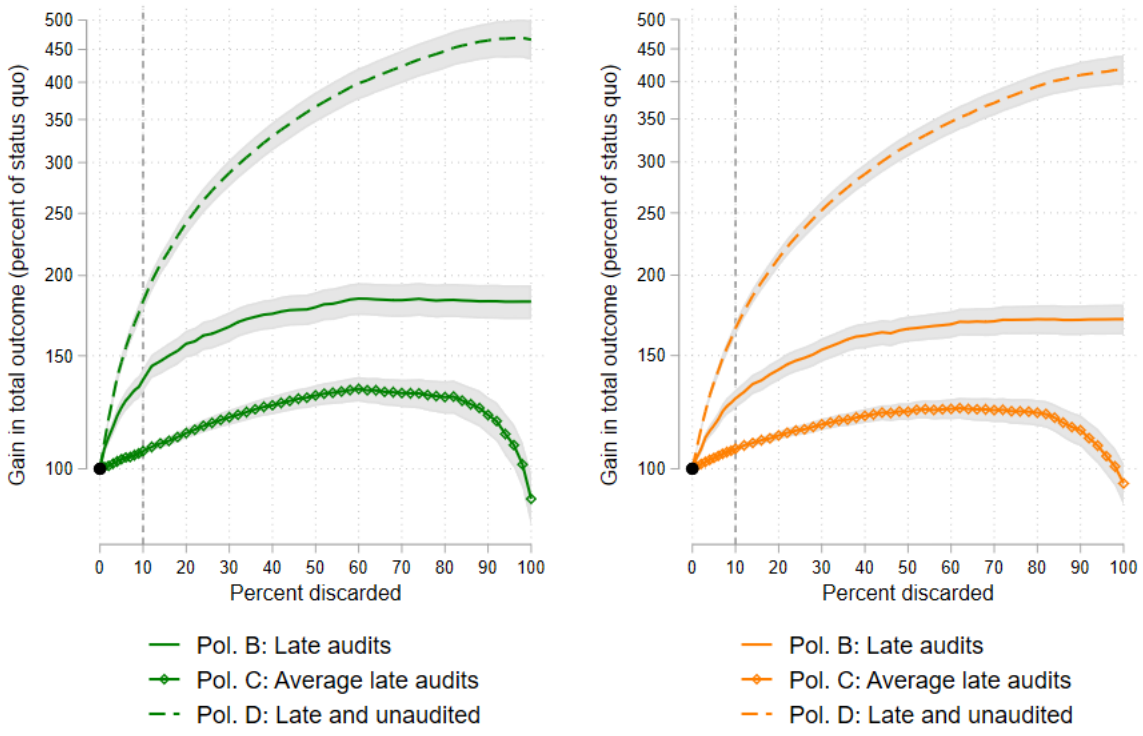


Fig. 4. Gain from discard and replace policies.
Notes: This figure reports on the y-axis the gain in total outcome of discarding a given percentage of early tax audits with the lowest predicted outcome and replacing them with an equal number of tax returns *(i)* with the highest predicted outcome among those audited later (Policy B); or *(ii)* with outcome equal to the average outcome of a late audit (Policy C); *(iii)* with the highest predicted outcome among those audited later or never audited (Policy D). The x-axis reports the discarded percentage. The vertical dashed line indicates the 10% level of discarding. The outcome variable is detected tax evasion in the left panel and recovered tax evasion in the right panel. All values are reported relatively to the status quo total outcome of early audits set at 100 and represented by a dot. The y-axis is represented on a logarithmic scale. The gray-shaded areas display 95% confidence intervals of the gain estimates computed performing a bootstrap, resampling returns, and re-calculating the policy gains for each of the percentiles discarded.

on the predictions for a "discard and replace" exercise, where we first discard the x percent early audits with the lowest predicted outcome as in Policy A, and then we "replace" them with an equal number of late audits with the highest predicted outcome. The total outcome of Policy B is:

$$Y_B^{k,x} = \sum_{o \in O} \sum_{c=2007}^{2009} \sum_{i \in \Phi_{o,c}^B} y_i^k,$$

$$with\ \Phi_{o,c}^B = \underline{\Pi}_{o,c}^{\hat{y}_i^k \geq q_{\underline{\Pi}_{o,c}}^{k,x}} \bigcup \widetilde{\Pi}_{o,c}^{\hat{y}_i^k \geq T_{\widetilde{\Pi}_{o,c}}^{k,x}} \tag{3}$$

where we define:

$$\widetilde{\Pi}_{o,c} = \overline{\Pi}_{o,c} \bigcup \underline{\Pi}_{o,c}^{\hat{y}_i^k < q_{\underline{\Pi}_{o,c}}^{k,x}}.$$

Essentially, to construct the set $\widetilde{\Pi}_{o,c}^{\hat{y}_i^k \geq T_{\overline{\Pi}_{o,c}}^{k,x}}$, we combine the potential replacements from $\overline{\Pi}_{o,c}$ with the set of returns discarded from $\underline{\Pi}_{o,c}$ (i.e., $\underline{\Pi}_{o,c}^{\hat{y}_i^k < q_{\underline{\Pi}_{o,c}}^{k,x}}$). Then, we choose from this union the set of replacements; which are selected based on the returns with the property that $\hat{y}_i^k \geq T_{\widetilde{\Pi}_{o,c}}^{k,x}$. In this way, we ensure that the replacement's predicted value is not lower than the predicted value of any discarded return.

For each percentile x, we compute the gain associated to the Policy B as a ratio $Y_B^{k,x}/Y^k$, and we illustrate how the gain changes as a function of x. This exercise relies on the longitudinal dimension of our data. Our data feature the universe of the tax returns that were audited over a five-year cycle. In our exercise, the pool of audits eligible for discarding (i.e., $\underline{\Pi}_{o,c}$) consists of the set of tax returns filed in the years $c = \{2007,2008,2009\}$, whose audits occurred 1, 2, or 3 years after filing returns. The donor pool for replacement audits (i.e., $\overline{\Pi}_{o,c}$) consists in the set of tax returns filed in the same years and offices, whose audits occurred 4 or 5 years after filing returns. These tax returns in the donor pool are a valid counterfactual because of three institutional features: *i)* they were available to audit at the time of decision for the tax returns in our consideration sample, *ii)* they had their tax return information locked in and so there is no risk of this information changing, and *iii)* they were ultimately actually audited, and so have been selected into audit and have observable outcomes.[21, 22] By revealed preference, tax returns that are available for an audit and

[21] Examining observable characteristics of tax returns audited at different ages, we find that tax returns with higher declared turnover and taxable income are audited sooner. However, conditional on our measure of predicted evasion, we find no systematic difference in actual observed evasion (either detected or recovered) between early and late audits. This is shown in Figure A2 in Appendix B. We interpret this as evidence that there is no systematic difference in the selection on unobservables between early and late audits. This is further confirmed by conversations with IRA representatives, which reveal that criteria for selection do not depend on the years since filing returns.

[22] In our context, we want to measure how the total outcome would change if we discard some returns and replace them with some others. Here, the counterfactual is the value of the replaced returns. For unaudited returns, this is not directly observable, but for those returns audited later, we know the value. If these late audits had been audited earlier (replacing the discarded returns),

overlooked in a given year are considered less promising by the IRA. The solid-line in Figure 4 shows very substantial gains of Policy B 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 aggregate *TaxEva* by 39% and the aggregate *TaxGot* by 29%, compared to status quo. In aggregate monetary terms, this amounts to an increase in detected evasion of 175 million euro and an increase in recovered evasion of 28 million euro per year. We account for the uncertainty from the sample of returns in our data by displaying the estimates with 95% bootstrapped confidence intervals.[23]

Since the IRA selection of audits typically occurs at the local office level, our baseline results consider only replacements within the same local office.[24] However, in principle, there could be greater gains from being able to replace from a broader pool. In Appendix E, we consider replacements from tax returns drawn within higher-level IRA offices, i.e., the same province, the same region, or anywhere in the country. While most of the gains are captured by replacing simply within the same office, reallocation at a higher level increases the total outcome further. With reference to the gains in *TaxEva,* after a 10% replacement under Policy B, reallocating within office implies a 39% gain, while reallocating at the province, region, and country level increases *TaxEva* by 43%, 48%, and 50%, respectively. This exercise suggests that some current organizational choices may be suboptimal.

Policy C: Discarding audits with low ex-ante promise, replacement with an average audit. The third experiment (*Policy C*) does not require the use of an ML algorithm for the selection of replacements. In this exercise, we discard the x percent of early audits with the lowest predicted outcome and replace them with an equal number of tax returns with outcomes equal to the average outcome among the tax returns of cohort c audited at any time by office o. The total outcome of Policy C is:

---

the outcome assessed by the audit would be the same.

[23]To compute these confidence intervals, we perform a bootstrap with 200 replications by re-sampling returns (and their associated predictions). We then re-calculate our discard and replace exercise for each of the percentiles discarded. We compute the confidence interval half-widths as 1.96 times the percentile-specific standard deviation of the bootstrapped estimates. In this exercise, we sample the original predicted evasion alongside the actual evasion amounts. As such, these confidence intervals capture uncertainty from the idiosyncratic aspects of which returns can be discarded or replaced. However, they do not capture the uncertainty from the prediction stage of our procedure. The computational requirements of the RF prevent us from nesting estimation of the prediction model within each bootstrap replication.

[24]There are 288 offices.

$$Y^{k,x}_{C,o,c} = Y^{k,x}_{A,o,c} + \sum_{o \in O} \sum_{c=2007}^{2009} \left(\underline{N}^{k,x}_{\Phi^A_{o,c}} \cdot E(\hat{y}^k_i | i \in \Pi_{o,c}) \right). \quad (4)$$

The assumption here is that the authority has (at least locally) constant returns to auditing: if a small fraction of the audits is removed, the revenue agency can replace them with a similar average return. This assumption is motivated by the fact that the authority is severely constrained in terms of resources, and so can audit only a tiny fraction of tax returns. 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. For each percentile x, we calculate the gain of Policy C as a ratio $Y^{k,x}_C / Y^k$, and illustrate how it changes as a function of x.[25] To isolate the benefits of an ML-guided discarding from those of an ML-guided selection of replacements, we compare the gains of Policy B with Policy C.

In Figure 4, the solid line with diamonds represents $Y^{k,x}_C / Y^k, \forall x$, in Policy C, that is when we consider the average among the late audits.[26] Compared to Policy B, this leads to a more modest but still meaningful improvement with respect to the status quo. For example, replacing the lowest 10% produces an overall 6.5% improvement relative to the status quo in *TaxEva* and 7.4% in *TaxGot*. These correspond to an additional annual 29 million euro of $TaxEva$ and 7 million euro of $TaxGot$.[27]

Policy D: Discarding audits with low ex-ante promise, replacement from all returns using ML guidance, ignoring the selective label problem. Finally, we consider an exercise where the selective labels problem is ignored (*Policy D*). In this exercise, the early audits discarded as in Policy A are replaced with the tax returns available for audit with the highest predicted outcome. Here, the set of candidate replacements includes the large number of those that were never audited. The total outcome obtained by Policy D uses the predicted outcome for the non-audited tax returns selected as replacements:

[25] We have also explored an alternative version of Policy C, where, after discarding the early audits with the lowest predicted evasion, we replace them with an equal number of late audits selected randomly. Monte Carlo simulations in Appendix D show that this variation yields similar results.

[26] The line is roughly unchanged if we consider the average of all tax returns because the average quality of tax returns audited 1 to 3 years after filing is similar to that of the tax returns that are audited 4 or 5 years after filing.

[27] The line for Policy C decreases for high replacement rates because the very last audits discarded are high-value audits, which are then replaced with an "average" audit which is worse. For Policy B, we allow the discarded audits to be available in the replacement pool; so that the line of Policy B should never go down. The turning point (where the slope for Policy C changes from positive to negative) indicates where the discarded audits are the same as the "average" replacement audit.

$$
\begin{aligned}
Y_D^{k,x} &= \sum_{o \in O} \sum_{c=2007}^{2009} \sum_{i \in \Phi_{o,c}^D} \tilde{y}_i^k, \\
&with\ \Phi_{o,c}^D = \underline{\Pi}_{o,c}^{\hat{y}_i^k \geq q_{\underline{\Pi}_{o,c}}^{k,x}} \bigcup \left[\widetilde{\Omega}_{o,c}\right]^{\hat{y}_i^k \geq T_{\widetilde{\Omega}_{o,c}}^{k,x}}, \\
\tilde{y}_i^k &= y_i^k\ if\ i \in \Pi_{o,c} \\
\tilde{y}_i^k &= \hat{y}_i^k\ if\ i \notin \Pi_{o,c}
\end{aligned} \tag{5}
$$

and where we define:

$$
\widetilde{\Omega}_{o,c} = \left[\Omega_{o,c} - \underline{\Pi}_{o,c}^{\hat{y}_i^k \geq q_{\underline{\Pi}_{o,c}}^{k,x}}\right].
$$

The set $\widetilde{\Omega}_{o,c}$ consists of all possible tax returns i excepting the early returns that are not discarded. Then, we select from this set the highest-predicted returns to replace those discarded, $\hat{y}_i^k \geq T_{\widetilde{\Omega}_{o,c}}^{k,x}$.

Similarly to Policies B and C, we display the gain associated with Policy D as a ratio $\tilde{Y}_D^{k,x}/Y^k$ for each percentile x. The dashed line in Figure 4 depicts the gains that are obtained under Policy D. This is when the discarded tax returns are replaced by the tax returns with the highest predicted outcomes from the full set of audited and unaudited tax returns from the same cohort and office as those discarded. This strategy not only considers a larger pool of donors but also relies on the accuracy of the predictions of our model to impute the value of the never-audited tax returns. Differently from Policy B, we cannot verify this directly, so we consider these results as speculative. The dashed lines in Figure 4 show that in our context, this practice would lead to estimated gains that are extremely large. These gains amount to 83% of the status quo (or 373 million euro) for *TaxEva* and 65% (or 63 million euro) for *TaxGot*.

Taken together, these exercises show that the range of possible gains from ML guidance of audit selection may be wide, depending on the fraction discarded and the thought experiment behind replacement. In all cases, the gains would be a meaningful improvement with respect to the status quo.

6 Extensions

Gains of random forest versus linear models. To explore how much of the gains depend on the specific prediction model, we repeat the policy experiments us-

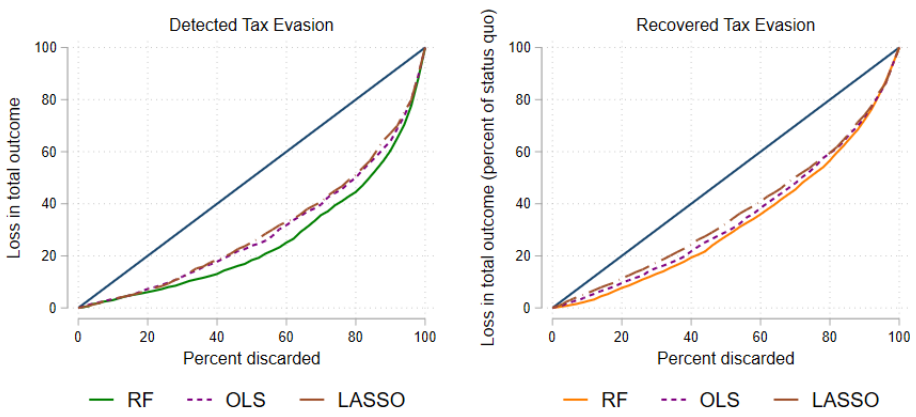


Fig. 5. Loss from discarding by prediction model.
Notes: This figure reports on the y-axis the loss in total outcome of discarding in total outcome of discarding a given percentage of early tax audits with the lowest predicted outcome variable by prediction model. The loss reported is computed by using the predictions of a random forest model in the solid line, of a LASSO model in the long dashed line, of an OLS model in the short dashed line. The x-axis reports the discarded percentage. The outcome variable is detected tax evasion in the left panel and recovered tax evasion in the right panel. All values are reported relatively to the status quo total outcome of early audits set at 100 and represented by a dot.

ing the predictions of OLS and LASSO models. Figure 5 displays Lorenz-type curves reporting the percentage reduction ($Y_A^{k,x}/Y^k$) in $TaxEva$ (left panel) and $TaxGot$ (right panel) when discarding the audited tax returns with the lowest predicted outcome, separately by prediction model. For $TaxEva$, we find that the RF is better at ranking tax audits relative to the OLS and LASSO models since the curve is further away from the 45-degree line. For $TaxGot$, the difference between the curves is much smaller. Figure 6 compares the gain from discarding the x% early tax audits with the lowest predicted evasion and replacing them with an equal number of late tax audits with the highest predicted evasion, under different prediction models. The RF model (solid lines) returns a higher total gain $Y_B^{k,x}/Y^k$ than both the LASSO and OLS models (dashed and dotted lines, respectively) for both variables. By discarding-and-replacing 10% of audits, total $TaxEva$ increases by 31% using either OLS or LASSO, compared to 39% using the RF model. The gains for $TaxGot$ are more similar to

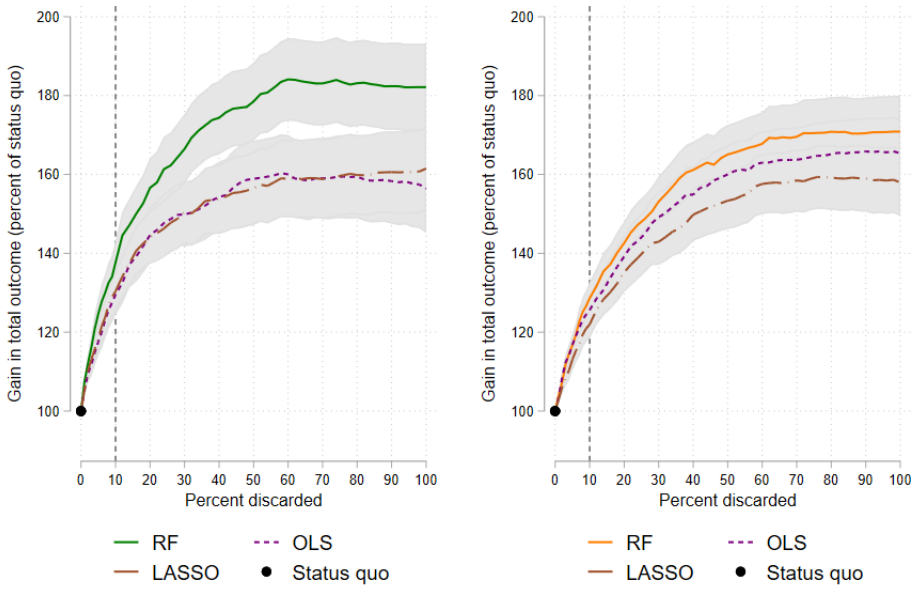


Fig. 6. Gain from Policy B discard and replace by prediction model.
Notes: This figure reports on the y-axis the gain in total outcome of discarding a given percentage of early tax audits with the lowest predicted outcome variable and replacing them with an equal number of tax returns with the highest predicted outcome among those audited later (Policy B) by prediction model. The gain reported is computed by using the predictions of a random forest model in the solid line, of a LASSO model in the long dashed line, of an OLS model in the short dashed line. The x-axis reports the discarded percentage. The vertical dashed line indicates the 10% level of discarding. The outcome variable is detected tax evasion in the left panel and recovered tax evasion in the right panel. All values are reported relatively to the status quo total outcome of early audits set at 100 and represented by a dot. The gray-shaded areas display 95% confidence intervals of the gain estimates computed performing a bootstrap, resampling returns, and re-calculating the policy gains for each of the percentiles discarded.

those of the RF (26% using OLS and 23% using LASSO compared to 29% using RF).

We further explore the sources of these gains. As mentioned in Section 4, a first source of difference may be that the RF model allows for nonlinear functions of the predictors. However, when repeating our policy exercises with OLS and LASSO models augmented by polynomials or interaction terms (see Section 4 and Table 1, Panel B), the "discard and replace" results do not change very much compared to the baseline (see Appendix Figure A7). We take this as evidence against the hypothesis that the main source of the gains in the RF model is its ability to include nonlinearities. Another candidate reason for the difference is that the statistical model for OLS and LASSO allows extrapolation beyond the support of the data, potentially

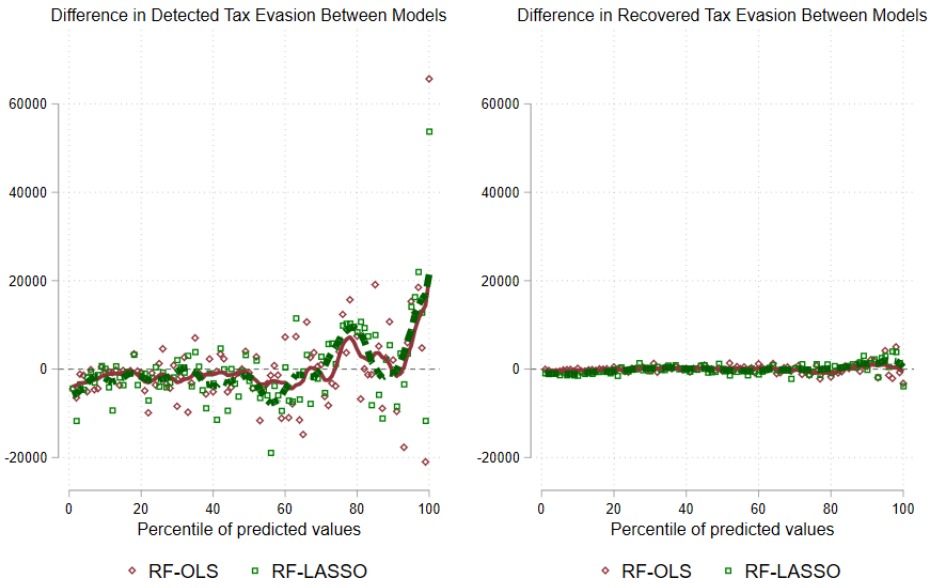


Fig. 7. Differences in model fit of OLS and LASSO relative to RF.
Notes: This figure reports the difference between the detected tax evasion (left panel) and the recovered tax evasion (right panel) of realized audits in the testing sample by percentiles of values predicted by the random forest and an alternative prediction model. Diamonds and squares denote the differences between the set of values predicted in the p-th percentile by the random forest and the OLS and LASSO, respectively. The solid and dotted lines depict the relationship fitted using a local linear regression with a triangle kernel and a bandwidth of five.

leading to extreme predictions. RF, however, avoids extrapolating predicted values beyond the training data. The handling of extrapolation by these models may interact with our "discard and replace" exercise, which depends on identifying returns with the most extreme predicted values. Such exercises could be particularly sensitive to extrapolation errors.

To examine the possibility of different extrapolation biases, we consider the difference in prediction quality when tax returns are sorted by percentiles. We compute the mean actual evasion for observations in the first percentile of OLS-based predicted evasion, and compare this to the mean actual evasion for observations in the first percentile of RF-based predicted evasion. We make this comparison percentile-by-percentile. Figure 7 shows these differences for RF-OLS percentiles and RF-LASSO percentiles. For example, the first diamond tells us that the 1% lowest RF-predicted evasion returns have, on average, 4,471 Euro less actual evasion per return than the lowest 1% OLS-predicted evasion returns. Therefore, when discarding returns with

low predicted evasion, using the RF-selected returns will result in less actual evasion discarded. On the other end of the graph, the top 1% RF-predicted evasion returns have an average 65,667 Euro per return higher than their OLS counterparts. This shows that replacing with RF-based predictions will lead to much more evasion being detected. Figure 7 shows a few additional patterns of interest. First, the patterns for RF compared to LASSO are similar to those for RF compared to OLS. Second, the percentile-by-percentile differences are noisy. We add a local linear estimate (using a triangle kernel and a bandwidth of five percentiles) to smooth out the noise. Third, while RF does better both at the highest- and lowest-predicted percentiles, the magnitude of improvement is greater for the highest percentiles. This evidence indicates that the improvement in RF primarily comes from finding better replacements. We speculate that this, in turn, results from OLS (and LASSO) being subject to greater extrapolation errors for top-predicted evasion returns. Our conjecture is consistent with the observation that the gains for $TaxGot$ are much smaller. Because large evaders on the right tails of the distribution of $TaxEva$ often do not pay back the money ($TaxGot$=0), the distribution of $TaxGot$ features a lower number of extreme observations.[28]

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, as discussed in the introduction, the ultimate goal of the IRA is explicitly articulated by law. Following the advice of IRA officials, we have translated the legal mandate into the two objectives of maximizing detected and recovered evasion, and we have designed algorithms tailored to best predict *each of the two* policymaker goals, which we have labeled *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 reduced improvements for the other measure. Appendix C shows that the dominant predictors in the RF differ somewhat between *TaxEva* and *TaxGot*. Also, among our pooled testing sample, the correlation coefficient between predicted *TaxEva* and predicted *TaxGot* is only 0.55. This implies that targeting tax returns to discard (and/or replace) based on the "wrong measure" can reduce the value of the exercise.

To gain insight into the tradeoffs across these outcomes, we model the auditors' goal as trying to maximize predicted *TaxEva,* subject to a constraint on the minimum

[28]The coefficient of variation for the distribution of $TaxEva$ is 5.4 versus 3.7 for $TaxGot$, confirming a higher spread (compared to the mean) of the first distribution.

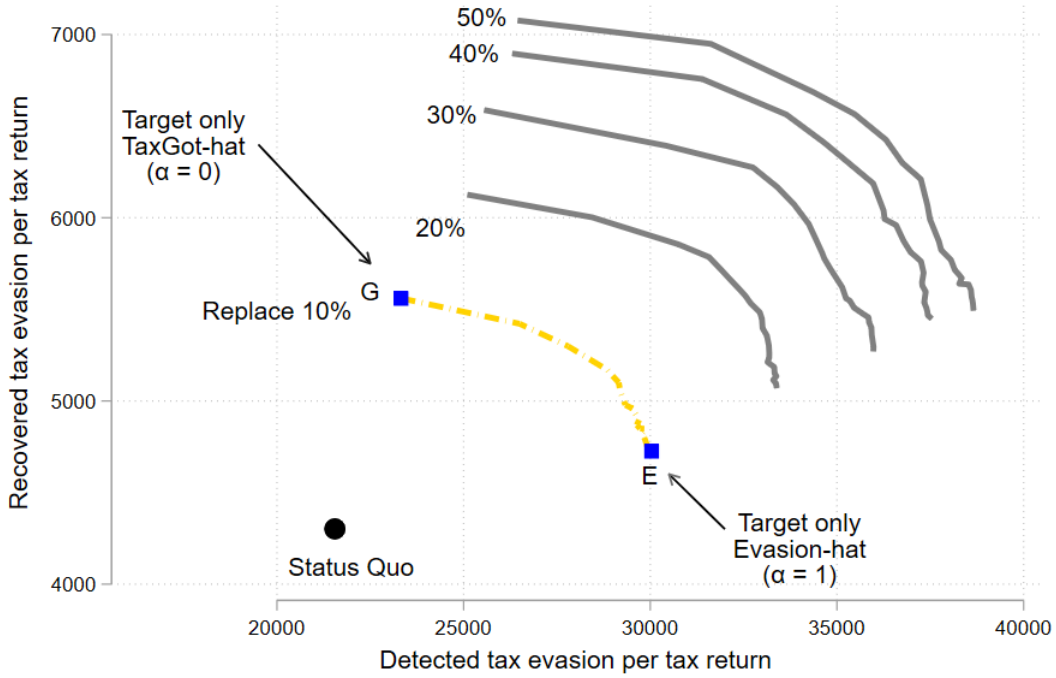


Fig. 8. Tradeoffs between detected and recovered tax evasion.
Notes: This figure reports the detected tax evasion per tax return (x-axis) against the recovered evasion per tax return (y-axis) after discarding the tax returns with the lowest predicted utility and replacing them with the same number of tax returns with the highest predicted utility at the office level (Policy B). Each line reports a different percentage of "discard and replace" values, and each point along a line represents different combinations of utility weights on the two policy goals, namely maximizing detected tax evasion and recovered tax evasion. The status quo levels are represented by a dot.

amount of predicted *TaxGot*. We do this by selecting the set I of replacement audits from the donor set (i.e., $\overline{\Pi}_{o,c}$ for policy B) in order to maximize the Lagrangian function:

$$\mathcal{L}(\lambda, I) = \sum_{j \in I} \hat{y}_j^{TaxEva} + \lambda \sum_{j \in I} \hat{y}_j^{TaxGot} \tag{6}$$

where λ is a parameter that calibrates how much importance we give to $TaxGot$ as a criterion for selection. The predictions for this choice problem come from our prediction model: $\hat{y}_j^{TaxEva} = \varphi^{TaxEva}(Z_j)$, $\hat{y}_j^{TaxGot} = \varphi^{TaxGot}(Z_j)$. A value $\lambda = 0$ corresponds to the case in which we ignore $TaxGot$; the larger is λ, the more importance we assign to $TaxGot$ as a criterion for selection. Maximizing $\mathcal{L}\left(\lambda, \overline{\Pi}_{o,c}\right)$ corresponds to selecting the replacement returns among the top returns according to the criterion:

$$\kappa\left(\hat{y}_j^{TaxEva}, \hat{y}_j^{TaxGot}\right) = \alpha \cdot \left(\hat{y}_j{}^{TaxEva}\right) + (1-\alpha) \cdot \left(\hat{y}_j{}^{TaxGot}\right) \tag{7}$$

for $\alpha = \frac{1}{1+\lambda} \in [0,1]$. As we change λ in $[0,\infty)$ (or equivalently α in $[0,1]$), we trace the feasible values of $\sum_{j\in I} \hat{y}_j^{TaxEva}$ and $\sum_{j\in I} \hat{y}_j^{TaxGot}$ and thus the trade-off between the two objectives. Here, no constraint on the minimum for predicted *TaxGot* corresponds to $\alpha = 1$, and a constraint of pursuing the maximum possible predicted *TaxGot* corresponds to $\alpha = 0$.

For each tax return, we compute the index in (7) for a range of values $\alpha \in [0,1]$. For each value of α, we rank them according to their value $\kappa\left(\hat{y}_j^{TaxEva}, \hat{y}_j^{TaxGot}\right)$ and implement our "discard and replace" exercise, where we replace with the best available tax returns from the donor pool of tax returns audited at later ages (Policy B). Each line in Figure 8 shows the average *TaxEva* and *TaxGot* per tax return by discarding and replacing a given percentage of tax returns, as indicated on the left end of each line. The dashed line corresponds to a 10% discard rate. When the criterion for selection of a tax return 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.

We highlight two features of these results. First, and in line with Figure 4, Figure 8 shows that there is great scope for improving the selection of tax returns. Second, Figure 8 shows that when the policy objective is a simple function as in (7), even targeting either measure may dramatically improve both outcomes, compared to the status quo (represented by a dot). However, there is some tradeoff between targeting only *TaxGot* versus targeting only *TaxEva*. Under a 10% replacement rate, targeting *TaxEva* only ($\alpha = 1$) produces a 39% increase over the status quo in *TaxEva*, and a 10% increase in *TaxGot*. On the other hand, targeting *TaxGot* only ($\alpha = 0$) produces a 8% increase over status quo in *TaxEva*, and a 29% increase in *TaxGot*. In this sense, the tradeoffs are roughly symmetric across the two measures, in terms of percentage increase over the status quo.

Algorithmic Fairness and Vertical Equity. The adoption of a ML-guided selection in the policy exercises might inadvertently result in the disproportionate reallocation of audits towards tax returns in specific income groups and sectors. This would generate concerns around the algorithm's fairness as highlighted by Black et al. (2022). Additionally, the reallocation could imply lower deterrence effects in those income groups or sectors less targeted by the algorithm. To address these concerns, in Appendix E, we repeat the "discard and replace" exercise by choosing replacement tax

returns only within deciles of taxable income or within business sectors. We show that when replacing the tax returns with the lowest 10% of predicted *TaxEva*, the gains of Policy B and Policy D (39% and 83%, respectively) are reduced to 18% and 10% when replacements of discarded tax returns are limited to tax returns in the same income decile as the discarded ones, and to 18% and 15% when replacements are limited to tax returns in the same business sector. This shows that there are considerable gains from replacement, even when keeping the composition of key characteristics broadly unchanged for the audited tax returns.

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

7 Concluding Remarks

This paper explores the extent to which ML can be used to improve audit selection. We use a prediction model to calculate gains from a "discarding" exercise, where the audits with the lowest 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 9% predicted outcome recover less than the material cost of conducting the audit.

The "discard and replace" exercise poses a significant challenge as actual outcomes for most tax returns are typically not observed, because they did not receive an audit. We propose a novel solution to address this challenge. Because the IRA has a 5-year-window for auditing a tax return, we devise a methodology where we focus on the selection by the third year of audit eligibility, and we use as counterfactuals the tax returns unaudited by the third year but audited at a later stage. This allows for the use of tax returns that were available at the time of the audit selection but were neglected for predictably inferior choices. Since these tax returns were later audited, we can use the observed outcome of the audit to assess the gains from the replacement. Even if we restrict replacements to tax returns from the same office, we find substantial improvements over the status quo in terms of the detected and

recovered evasion. At a 10% discard rate, selecting the replacements using ML from this pool yields an improvement of 39% in TaxEva, and of 29% in *TaxGot*. Allowing replacements to be selected from the larger pool of unaudited tax returns 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% in *TaxEva*, and of 65% in *TaxGot*.

As a lower-bound assessment, we also evaluate the potential improvements when replacements are selected at random from the pool of audited tax returns, both limiting the pool of replacements to tax returns that were later audited and to the larger pool of audited tax returns. The idea is that if only a small fraction of audited tax returns is discarded and replaced, the tax agency can at least select replacements that have detected or recovered evasion equal to the average value among the audits. Although the improvements are naturally smaller in this case, they remain significant: at a 10% discard rate, selecting the replacements at random yields an improvement of about 7% in *TaxEva*, and of 8% in *TaxGot*, independently of whether the pool is restricted or not.

While in theory, it is possible that implementing ML-guided audit selection would prompt taxpayers to adjust their behavior, our observations on enhancing current policies remain relevant for at least two key reasons. First, the audit selection policy (and any changes to it) is unobserved by the taxpayers. Thus, even assuming full rationality and forward-looking expectations, taxpayers would take a substantial amount of time to learn and adapt to the new policy. It is important to notice that our strategy leaves the average audit probability and its distribution across geographical locations (288 local offices of the tax authority) unchanged. We modify the composition of the audited taxpayers, prioritizing those identified by the ML algorithm as more likely to evade taxes.[29] Secondly, and arguably most crucially, our analysis does not imply that the IRA should only update its policies using existing data once. Instead, the policy should continuously monitor behavior and interactively adjust over time. The results presented above should be interpreted as an indication of improvement opportunities.

Although our application of ML tools is focused on sole proprietors, it can be extended to other taxpayer categories. This extension is possible because tax return

[29]If taxpayers adjust their behavior in response to the average audit probability, we expect no reaction; if they act strategically, evaders who now understand that they are more likely to be included in the list of audits should evade less (report more) in hopes that of avoiding an audit. If so, our estimates of the benefits of using ML techniques to direct audit selection represent a lower bound of the equilibrium effect once taxpayers' responses (if any) are factored in.

and audit procedures are uniform across all types of taxpayers. Any tax return, regardless of the entity generating the income, can undergo audit within 5 years of filing. Therefore, the approach developed in this study can be adapted for use with all tax returns and in various countries with comparable audit systems.

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.

Ash, E., S. Galletta, and T. Giommoni (2024). A Machine Learning Approach to Analyze and Support Anti-Corruption Policy. *American Economic Journal: Economic Policy*.

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

Battaglini, M., L. Guiso, C. Lacava, and E. Patacchini (2019). Tax Professionals and Tax Evasion. *NBER Working Paper 25745*.

Bhatt, M., S. B. Heller, M. Kapustin, M. Bertrand, and C. Blattman (2024). Predicting and Preventing Gun Violence: An Experimental Evaluation of READI Chicago. *The Quarterly Journal of Economics 139*(1), 1–56.

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

Black, E., H. Elzayn, A. Chouldechova, J. Goldin, and D. E. Ho (2022). Algorithmic Fairness and Vertical Equity: Income Fairness with IRS Tax Audit Models. *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1479–1503.

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.

Boning, W. C., N. Hendren, B. Sprung-Keyser, and E. Stuart (2023). A Welfare Analysis of Tax Audits Across the Income Distribution. *NBER Working Paper 31376*.

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*.

Guyton, J., P. Langetieg, D. Reck, M. Risch, and G. Zucman (2021). Tax Evasion at the Top of the Income Distribution: Theory and Evidence. *NBER Working Papers 28542*.

Hino, M., E. Benami, and N. Brooks (2018). Machine Learning for Environmental Monitoring. *Nature Sustainability 1*(10), 583–588.

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. Springer.

Johns, A. and J. Slemrod (2010). The Distribution of Income Tax Noncompliance. *National Tax Journal 63*(3), 397–418.

Joseph, A. (2020). Parametric Inference with Universal Function Approximators. *Bank of England Staff Working Paper* (784).

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*.

Lundberg, S. and S.-I. Lee (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems 30*, 4765–4774.

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

Mittal, S., O. Reich, and A. Mahajan (2018). Who is Bogus?: Using One-Sided Labels to Identify Fraudulent Firms from Tax Returns. *COMPASS '18: Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, 1–11.

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.

Shapley, L. (1953). A Value for N-Persons Games. *Contributions to the Theory of Games 2*, 307–317.

Štrumbelj, E. and I. Kononenko (2010). An Efficient Explanation of Individual Classifications using Game Theory. *Journal of Machine Learning Research 11*(1), 1–18.

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

U.S. Government Accountability Office (2009). Limiting Sole Proprietor Loss Deductions Could Improve Compliance but Would Also Limit Some Legitimate Losses. *GAO-09-815*.

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.

APPENDIX

A Data: Further Details

Our data includes the universe of tax returns and audits to Italian 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 from 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 (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
Summary statistics - Audited tax returns.

	mean	median	std. dev.	10th pct.	90th pct.
Audit	0.021	0	0.143	0	0
Agriculture	0.007	0	0.083	0	0
Trade	0.024	0	0.154	0	0
Construction and manufacturing	0.024	0	0.154	0	0
Private services	0.022	0	0.146	0	0
Health, education, recreational services	0.017	0	0.127	0	0
Delinquency	0.437	0	0.496	0	1
Agriculture	0.340	0	0.474	0	1
Trade	0.433	0	0.496	0	1
Construction and manufacturing	0.556	1	0.497	0	1
Private services	0.397	0	0.489	0	1
Health, education, recreational services	0.238	0	0.426	0	1
Appeal	0.067	0	0.251	0	0
Agriculture	0.073	0	0.261	0	0
Trade	0.067	0	0.250	0	0
Construction and manufacturing	0.060	0	0.237	0	0
Private services	0.071	0	0.257	0	0
Health, education, recreational services	0.070	0	0.255	0	0
Positive evasion	0.765	1	0.424	0	1
Taxable income	22,633	12,992	52,522	0	49,072
Detected tax evasion (*TaxEva*)	20,053	3,704	91,794	0	31,310
Recovered tax evasion (*TaxGot*)	4,451	0	14,273	0	10,853
Years of activity	13.279	11	10.408	0	29
N. employees	0.825	0	3.121	0	2
Turnover	84,560	33,918	2,580,718	3,017	167,648

Notes: The sample includes 18,766,176 tax returns of 4,721,593 sole-proprietors for income years 2007-2012. 257,701 returns filed by 199,259 different taxpayers are audited. Audit, delinquency, and appeal are calculated considering tax returns with complete fiscal cycles (2007, 2008, 2009 tax returns, number of observations: 9,728,061). Financial accounts are expressed in euro. Detected and recovered tax evasion are reported after winsorization.

Table A1 shows summary statistics. The sample includes 18,766,176 tax returns of income produced in the years 2007-2012 by 4,721,593 sole-proprietorship taxpayers. Among these tax returns, 257,701 returns filed by 199,259 different taxpayers receive an audit. The probability of receiving an audit over the five years after filing tax returns is 2.1% and similar across sectors, except for businesses operating in agriculture that have a much lower audit rate (0.7%).[30] Among audited tax returns, the delinquency rate is 44%, and varies across sectors. About 56% of audited taxpayers operating in construction and manufacturing (e.g., small construction firms,

[30]In comparison, the U.S. Government Accountability Office (2009) reports that the IRS in 2008 audited about 1% of estimated noncompliant sole proprietors.

plumbers, artisans, bakers) do not pay back the detected evasion. Businesses providing services in trade and private services (e.g., lawyers, hairstylists, coffee shop owners, architects) have a delinquency rate of 43% and 40%, respectively. Finally, businesses in agriculture are delinquent 34% of the time, and those providing health, education, and recreational services (e.g., physicians, dentists) are delinquent 24% of the time. The probability that the taxpayer appeals the audit is on average 7%, with low variation across sectors: the appeal rate ranges between 6% for construction and manufacturing and 7.3% for agriculture.

Audits detect evasion in 77% of the cases. The average detected tax evasion is €20,053, 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 67% of the taxable income. The average recovered tax evasion is €4,451.[31]

The average audited taxpayer has been in operation for 13 years, only 24% of the businesses have employees and those with employees on average employ 3.4 workers. The average reported turnover is €84,560, with relevant heterogeneity (standard deviation: €33,918; 90th percentile €167,648) partly reflecting differences across industries.

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

B Model Performance: Further Details

Alternative sets of predictors. A key decision that characterizes the prediction model is the choice of predictors. We use 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) as our baseline set of predictors. In addition, we exploit the granularity of the geographical and

[31] We can only measure the evasion detected by an audit. In principle, this could miss sophisticated evasion approaches. In the US context, Guyton et al. (2021) show that detected evasion falls at the extreme top of the income distribution, which is evidence for this type of sophisticated evasion. In contrast, looking at more granular percentiles within the top 1% of our data, we do not find drops in the detected share of evasion. This suggests that the tax evaders in our context are not pursuing these sophisticated strategies.

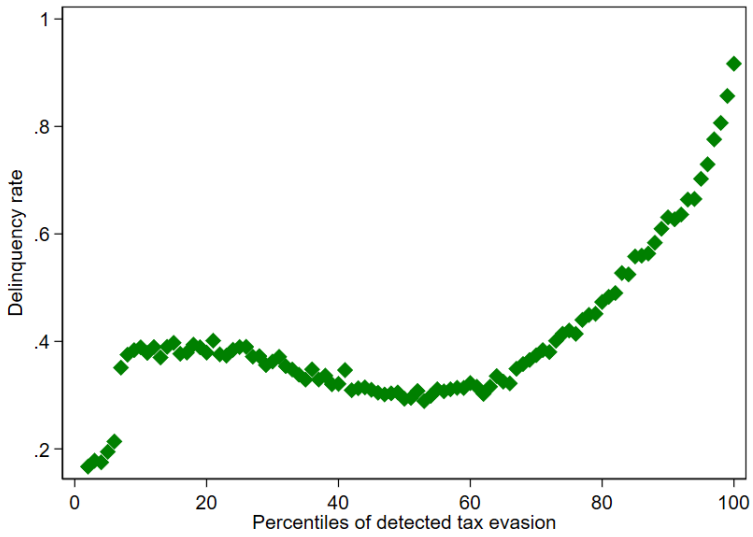


Fig. A1. Distribution of the delinquency rate by percentiles of detected tax evasion. Notes: This figure reports the delinquency rate for the percentiles of positive detected tax evasion.

sectoral information contained in the data using Mundlak-type predictors (Mundlak, 1978), defined as the average at the municipality and at the 5-digit sector level of two key financial accounts, namely taxable income and turnover. In Table A2, Panel A, we test the sensitivity of our algorithm to alternative sets of predictors. First, we show that by substituting the Mundlak controls with the 5-digit sector fixed effects (with or without the Mundlak 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 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 in predicting delinquency.[32] On the other hand, both $TaxEva$ and $TaxGot$ show a strong sectoral and geographical dimension: when removing sector and province fixed effects (last row), the performance of the model for both variables is dramatically reduced.[33]

[32] 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, and VAT transactions.

[33] We also experimented with the use of lagged variables as additional controls. Results show that the performance remains roughly unchanged. However, the use of lagged variables however reduces

Table A2
Comparison across random forest specifications.

A. 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	40,553	0.131	0.082	90,021	13,220
5 dgt sector FE + Mundlak's municipality	1,469	40,553	0.131	0.083	90,029	13,218
Mundlak's municipality & 5-dgt sector†	255	40,553	0.131	0.083	90,032	13,213
no Mundlak's controls	251	40,553	0.127	0.082	90,251	13,225
no detailed financial accounts	150	40,553	0.093	0.081	91,994	13,232
no province and 2-dgt sector FE	20	40,553	0.072	0.065	93,048	13,347

B. Alternative Training and Testing Samples

Training sample		Testing sample			Out-of-sample R-squared		Out-of-sample RMSE	
Return year	Audit year	Return year	Audit year	N. obs.	*TaxEva*	*TaxGot*	*TaxEva*	*TaxGot*
2007–2009†	2009–2014†	2007–2009†	2009–2014†	40,553	0.131	0.083	90,032	13,213
2007–2009†	2009–2014†	2010–2012	2011–2014	11,078	0.128	0.078	72,020	15,247
2007–2009†	2009–2011	2007–2009†	2012-2014	23,290	0.088	0.072	89,458	13,881
2007–2009†	2009–2014†	2007–2009†	2012-2014	23,290	0.106	0.085	88,573	13,781
2007–2009†	2009–2014*	2007–2009†	2012-2014	23,290	0.095	0.081	89,115	13,813

Notes: This table reports measures of fit computed on the testing sample. "Return year" is the year when the income declared in the tax return is generated. † indicates the baseline specification. The training sample includes 161,940 tax returns audited between 2009 and 2014 (baseline) and 67,924 tax returns audited between audit years 2009 and 2011. * indicates that the training set is a random extraction of tax returns audited between 2009 and 2014 of size equal to the number of audits in the years 2009–2011. We report the average goodness of fit over 10 such random extractions.

Model stability over time. Our baseline model is trained using a random 80% sample of tax returns of income produced in years 2007-2009, and it is tested on the remaining 20% of that set of tax returns. In Panel B, to assess stability over longer periods of time, we test the performance of our model trained on tax returns of years 2007-2009 to predict evasion in tax returns of income produced in later years (2010-2013). The measures of fit of the latter model (second row) are similar to those of the baseline model (first row of the panel).

Alternative training and testing samples. To maximize the number of observed audits in our baseline model, we use the 2007-2009 tax returns audited over the entire period (years 2009 - 2014). A potential concern could arise if later calendar year audits' outcomes incorporate critical information from the revealed outcomes of early audits. In principle, this could inflate the performance of our model, because some audits at time t could be informed by information not available at time t. To assess the scope of this issue, we recompute the performance of our model, enforcing a clean separation between training and testing data. We train the model from 2007-2009 tax returns audited in years 2009-2011, and test it on 2007-2009 tax returns

the number of years of audits that can be included and substantially reduces the sample size.

audited in years 2012-2014.[34] In the third row of Panel B, we report the fit measures for this different setting. Results show a mild decline in performance. The R-squared for *TaxEva* falls to 0.088, whereas the one for *TaxGot* goes to 0.072. This reduction might be due to a change in the testing sample, the use of a smaller training sample, the relevance of future information for prediction, or sampling luck. We perform the following exercises to understand how much of the reduction in fit could be accounted for by a change in testing data or a reduction in the training sample size. First, we repeat our row-3 exercise when using our baseline training sample (from row 1) as training sample. Second, we extract a random number of observations from the baseline training sample (from row 1) equal to the number of observations from the earlier audits training sample (from row 3). We repeat the random extraction ten times and then measure the average change in fit when considering the testing sample of late audits. The last two rows of Panel B show that a large part of the decline in performance is due to these factors rather than to relevant future information. We conclude that there is no significant concern about using the full range of audit years in our baseline sample. To maximize the sample size for training our model, we choose the specification that uses a training sample with tax returns audited at any time as our baseline specification.

Early versus late audits. Finally, we test whether the model's performance is similar when tax returns are audited at different times. This check is important since in our "discard and replace" exercise the discarded audits are those audited soon after filing tax returns (e.g., within three years, which we label "early") and the replacements come from those that would have been audited later (e.g., four or five years, labeled "late"). For our exercise to give gains from replacement, the prediction model needs to consistently rank of tax returns regardless of audit timing. Even better, its predictions should be equally accurate in levels regardless of whether an audit is early or late. We examine whether the model's predictions change in accuracy between early and late audits. In Figure A2 we reproduce Figure 2 by plotting separately the average actual outcome (*TaxEva* and *TaxGot*) of early and late audits for each percentile of predicted outcome. The deviations from the predicted evasion detected and recovered are similar between early and late audits. This suggests that there are no systematic differences in the selection on unobservables between early and late audits. This in turn supports our "discard and replace" exercise.

[34] One may be concerned that audits in 2009-2011 may be systematically different from later audits if funding to the IRA changed significantly over time. However, funds to the IRA from the Italian Treasury were pretty stable over our whole sample period, fluctuating only slightly around a non-trending level of €3.2 billion per year. Data is available upon request.

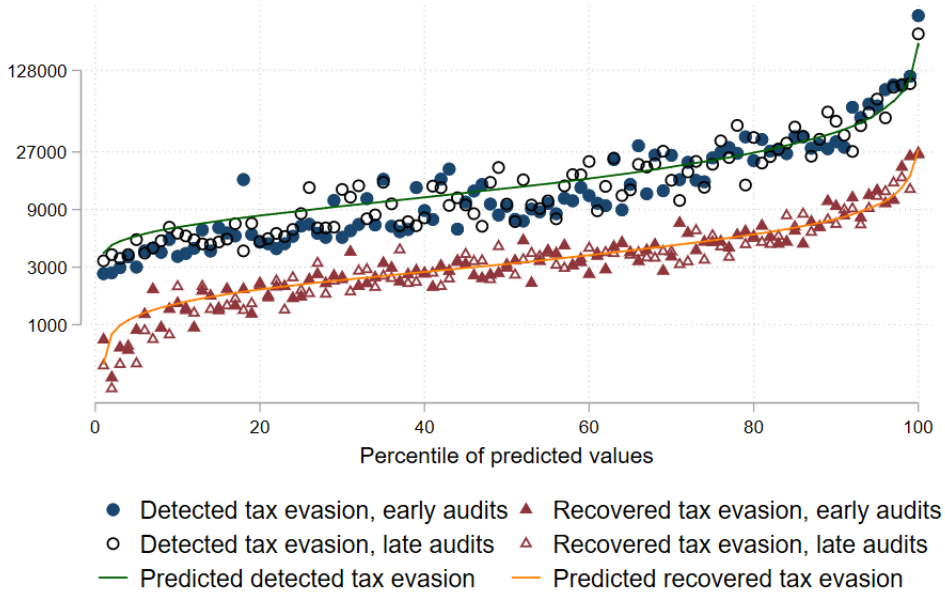


Fig. A2. Model fit by audit promptness.
Notes: This figure reports the detected tax evasion (dots) and the recovered tax evasion (triangles) of realized audits in the testing sample by percentiles of predicted values. Full and empty geometric figures indicate audits occurred by the third year after filing the tax return (early audits) and after the third year after filing the tax return (late audits), respectively. The green line displays the predicted detected tax evasion, and the orange line displays the recovered tax evasion. The sample includes tax returns of income produced in the years 2007-2012 that are audited by the IRA. The y-axis is represented on a logarithmic scale.

C Relevance of Predictors

In this section, we explore which variables are the main drivers of our prediction model. We investigate the importance of each predictor by computing the Shapley values, following Štrumbelj and Kononenko (2010). Their idea is to consider the predictors of a model as players in a cooperative game. Inferring how much a variable contributes to a prediction should consider the correlation of that variable with the other predictors. This is analogous to computing how much a player contributes to the coalition payoff of a cooperative game, where the other players might be complements or substitutes. The Shapley value is the weighted sum of the marginal contribution of a predictor value across all possible coalitions. Using the solution in Shapley (1953), the Shapley value of predictor p in a set of predictors P for observation i is defined as

$$\phi_p^P(x_{i;}\varphi) = \sum_{x' \subseteq \mathcal{C}(\mathrm{x}) \setminus \{\mathrm{p}\}} \frac{|x'|!(n-|x'|-1)!}{n!} \left[\varphi(x_i|x' \cup \{p\}) - \varphi(x_i|x')\right] \tag{A.1}$$

where x_i is the vector of predictors for observation i and $\varphi(\cdot|x')$ is the predictive model trained on the features x'. $C(\mathrm{x}) \setminus \{\mathrm{p}\}$ is the set of possible coalitions of n variables when excluding the p^{th} variable, and $|x'|$ denotes the number of included variables. We compute the Shapley values using the SHAP package by Lundberg and Lee (2017) to deal with computational complexity. To facilitate the interpretation, we follow Joseph (2020) and compute the Shapley share coefficients as a summary statistics for the contribution of each predictor x_p. The Shapley share coefficients are defined in the interval [-1,1] as

$$\Gamma_p^P(\hat{f}) \equiv \left[sign\left(\hat{\beta}_p^{lin}\right) \left\langle \frac{|\phi_p^P(x_{i;}\varphi)|}{\sum_{l=1}^n |\phi_l^P(x_{i;}\varphi)|} \right\rangle_{\Delta} \right], \tag{A.2}$$

where $\langle\cdot\rangle$ computes the average over Δ, for computational tractability, a random sample of audited returns. $sign\left(\hat{\beta}_p^{lin}\right)$ is the sign of the coefficient for predictor p estimated in a linear regression of the outcome k on the vector of predictors.

Figure A3 reports for each outcome the fifteen most important predictors, according to Shapley shares, which are averaged over a random sample of 1,000 observations. The figure additionally reports the predictors at the 25th, 50th, and 75th percentile of estimated Shapley shares.

Figure A3 shows that many features have predictive power: the 75th percentile of predictive power is still meaningfully away from zero. However, by the 50th percentile, the predictive power is very small. The main predictors for both variables are different types of reported income (total, gross, taxable) and income source (real estate, professional, compensations), turnover, operating costs , and taxes (on purchases and imported intermediate goods, credit, withholdings). Moreover, Mundlak controls turn out to be among the most important predictors. Interestingly, there is meaningful variation in the predictive power of some predictors across the two outcome variables. While some accounts have high Shapley shares for both outcomes, their sign is the opposite. For example, real estate income is associated simultaneously with lower predicted detected evasion and higher recovered evasion. This may be because it is hard to hide that income, so the evasion of taxpayers with high values along this dimension is generally low, but whenever it occurs, the tax authority is more likely to recover the evaded amount of taxes. This underlines the tradeoff

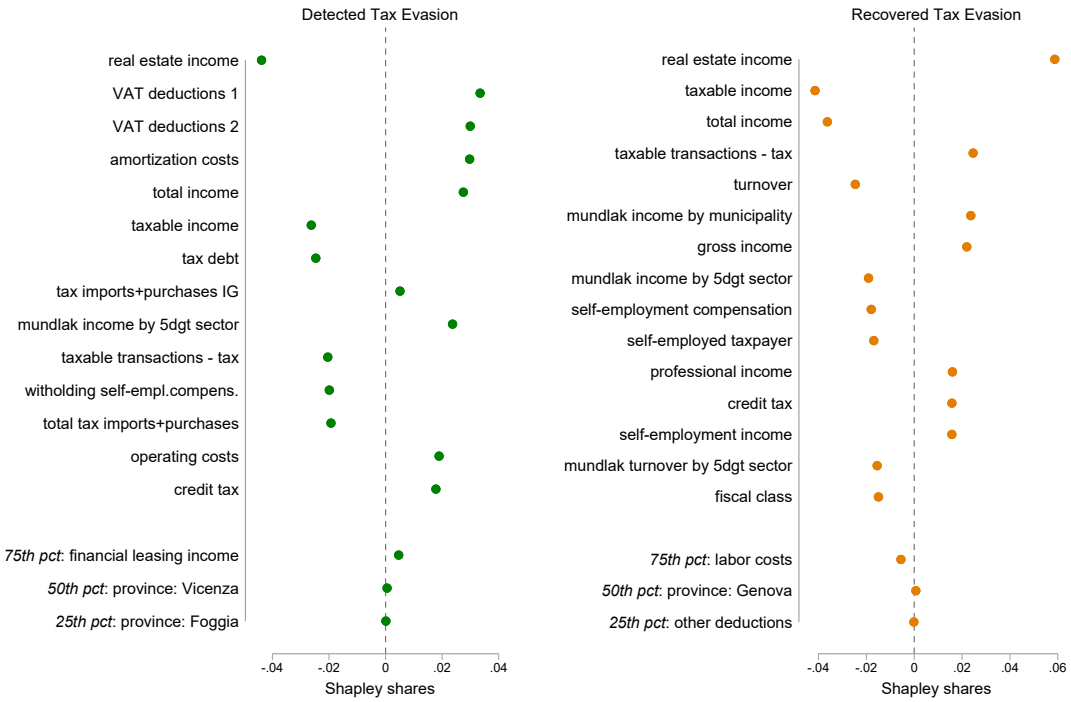


Fig. A3. Shapley share coefficients.

Notes: This figure reports the importance of different predictors according to the Shapley share coefficients. The left figure illustrates the predictors of detected tax evasion, and the right figure shows those of recovered tax evasion. Each figure shows the predictors ordered by the absolute value of their Shapley share. The fifteen most important predictors and the predictors at the 25th, 50th, and 75th percentile according to the Shapley shares are displayed. The reported Shapley shares are based on Shapley values computed on a random subsample of 1,000 observations. "IG" stands for intermediate goods.

between multiple outcomes when building a prediction model and complements our discussion in Section 6.

D Discard and Replace Exercise: Further Details

Expanding the geography of the pool of available replacements. In this section, we repeat our baseline "discard and replace" exercise (Policy B), allowing discarded and replacement tax returns 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 tax returns 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 tax returns) within the province, and replacing them with an equal number of audits on the tax returns with the highest predicted outcome among later audits (i.e., among those that were audited 4 or 5 years after filing tax returns) within the province. Results from these exercises are presented in Figure A4. 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 within an office and gives the same results as in Figure 4, 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.

Montecarlo simulations. In Section 5, we consider replacing with the average tax return from all tax returns audited 4-5 years after filing (Policy C). Because the "average tax return" is not a feasible direct choice, in this section we consider replacement with a random subset of actual tax returns. We conduct 100 simulations, and in each one, we draw a random subset of tax returns as replacements. The 95% confidence intervals for this exercise are reported in Figure A5. 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.

E Algorithmic Fairness and Vertical Equity: Further Details

The adoption of an ML-guided selection in the policy exercises discussed might indirectly result in the disproportionate reallocation of audits towards tax returns in specific income groups and sectors, generating concerns around the fairness of the algorithm as highlighted by Black et al. (2022) analyzing audits selected with ML classification techniques by the IRS in the United States. At the same time, this could imply lower deterrence effects in those income groups or sectors that are less targeted by the algorithm. To address these concerns, in this section, we consider a "constrained" discard and replace exercise. We focus on Policy B, replacing discarded audits from the pool of tax returns audited 4 or 5 years after filing. The constraints

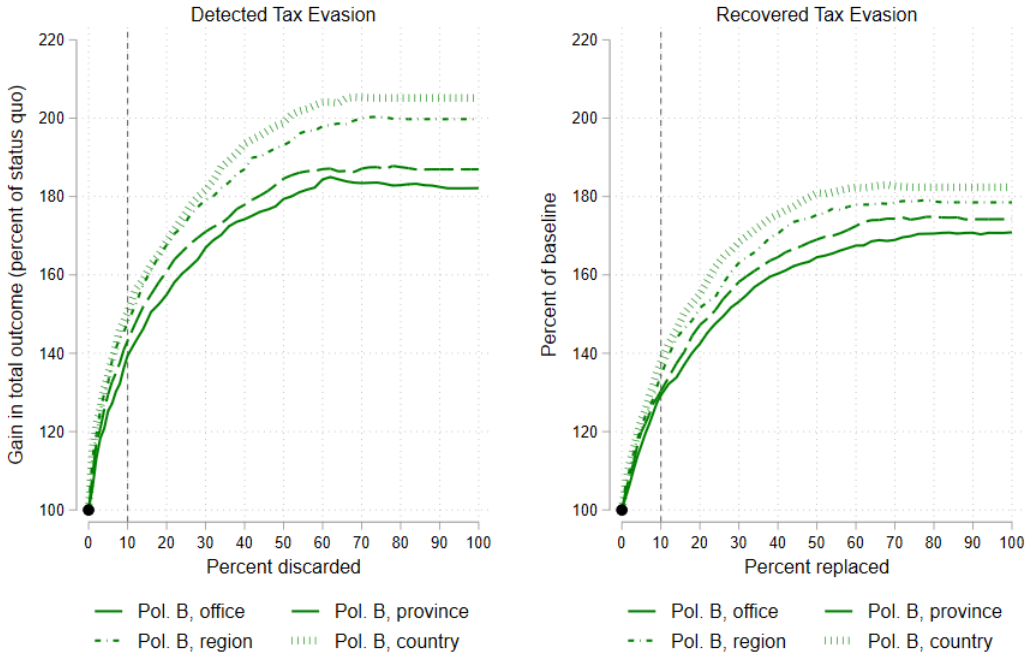


Fig. A4. Gain from Policy B discard and replace by organizational level.
Notes: This figure reports on the y-axis the gain in total outcome of discarding a given percentage of early tax audits with the lowest predicted outcome and replacing them with an equal number of tax returns with the highest predicted outcome among those audited later (Policy B) by considering replacements across different organizational levels of the tax authority. The gain reported is computed by replacing the discarded audits with late audits with the highest predicted value at the local office level in the solid line, provincial office level in the dashed line, regional office level in the dash-dotted line, and centralized-country level in the dotted line. The x-axis reports the discarded percentage. The vertical dashed line indicates the 10% level of discarding. The outcome variable is detected tax evasion in the left panel and recovered tax evasion in the right panel. All values are reported relatively to the status quo total outcome of early audits set at 100 and represented by a dot.

we explore are either (1) income-decile constraints or (2) business sector constraints. To implement these constraints, we treat each return-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 year-office-decile pool among those tax returns that were audited 1-3 years after filing. We replace within the same year-office-decile pool, using tax returns audited 4-5 years after filing.

Figure A6 shows the result of this exercise. The top green line is the unconstrained exercise Policy B 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

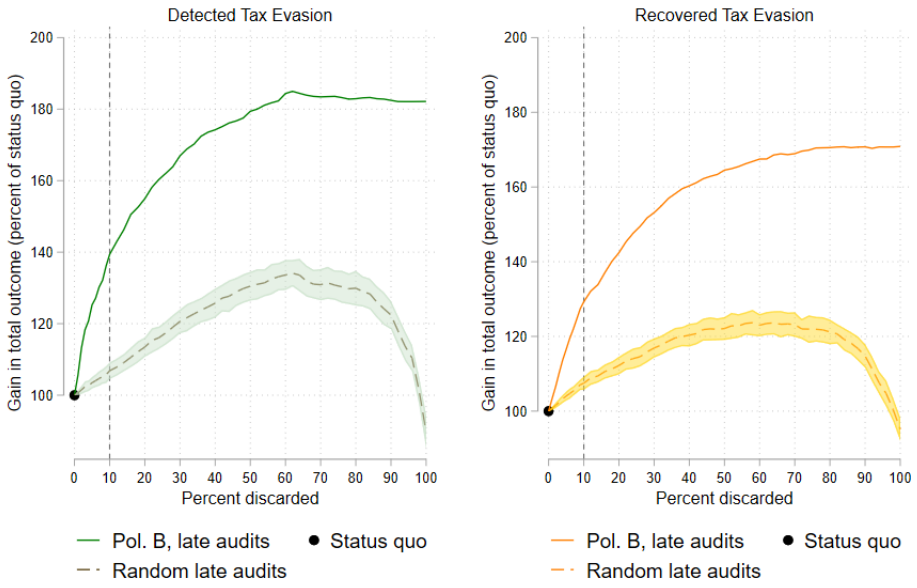


Fig. A5. Gain from discard and replace with random late audits.
Notes: This figure reports on the y-axis the gain in total outcome of discarding a given percentage of early tax audits with the lowest predicted outcome and replacing them with an equal number of tax returns with the highest predicted value using 100 samples randomly drawn among those audited later. The dashed line and the shaded areas report the average gain of a draw and its 95% confidence intervals, respectively. The x-axis reports the discarded percentage. The vertical dashed line indicates the 10% level of discarding. The outcome variable is detected tax evasion in the left panel and recovered tax evasion in the right panel. All values are reported relatively to the status quo total outcome of early audits set at 100 and represented by a dot.

shows results when 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 the status quo.

Table A3 shows that the proposed policies are not associated with a selection of tax returns that differs 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 tax returns under the alternative versions of Policy B replacement schemes mentioned above (i.e., the baseline replacements within office, replacements within office and sector, and within office and income decile). Results for different outcome variables are reported in different panels (*TaxEva* in panel A,

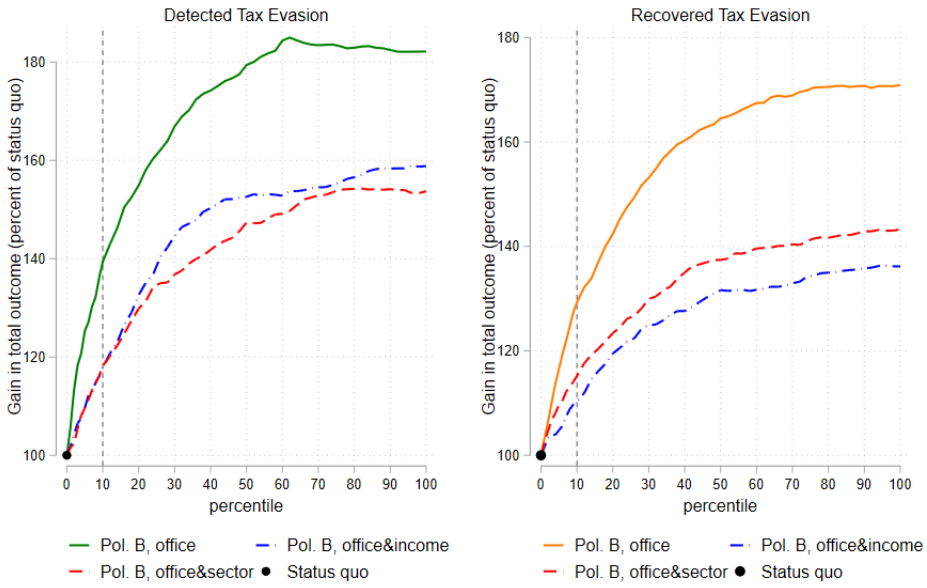


Fig. A6. Gain from Policy B discard and replace within sector or income
This figure reports on the y-axis the gain in total outcome of discarding a given percentage of early tax audits with the lowest predicted outcome and replacing them with an equal number of tax returns with the highest predicted outcome among those audited later (Policy B) with additional restrictions on possible replacements. The gain reported is computed by using the late audits of the same office (as in the baseline Policy B in Figure 4) in the solid line, of the same office and 2-digit sector in the long dashed line, of the same office and income decile in the short dashed line. The x-axis reports the discarded percentage. The outcome variable is detected tax evasion in the left panel and recovered tax evasion in the right panel. All values are reported relatively to the status quo total outcome of early audits, set at 100 and represented by a dot.

TaxGot in panel B). We find that the replacement tax returns selected following the ML predictions are filed by taxpayers with similar demographic characteristics to the ones who filed the discarded tax returns irrespective of the replacement scheme (e.g., gender, age and marital status). Business characteristics are balanced too: the discarded and replacement pools involve businesses that are comparable in terms of family business status, presence of employees, and years of activity. Perhaps interestingly, the sectoral 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

Table A3
Characteristics of tax returns when discarding and replacing 10% audits.

		Panel A Target: TaxEva			*Panel B* Target: TaxGot		
	Status quo	Office	Office/Sector	Office/Income	Office	Office/Sector	Office/Income
Woman	0.226	0.210	0.221	0.218	0.215	0.221	0.226
Age	2.695	2.876	2.839	2.843	2.876	2.842	2.844
Married	0.681	0.680	0.677	0.680	0.705	0.695	0.689
Family business	0.102	0.104	0.103	0.103	0.118	0.118	0.116
Has employees	0.425	0.464	0.449	0.447	0.472	0.460	0.452
N. employees	1.907	2.495	2.181	2.165	2.371	2.224	2.157
Years of activity	13.000	12.882	12.957	12.917	13.510	13.222	13.289
Sectors:							
Agriculture	0.036	0.035	0.036	0.037	0.041	0.036	0.042
Trade	0.334	0.313	0.334	0.321	0.314	0.334	0.336
Construction and manufacturing	0.236	0.291	0.236	0.276	0.209	0.236	0.217
Private services	0.367	0.336	0.367	0.342	0.398	0.367	0.376
Other services	0.027	0.025	0.027	0.023	0.039	0.027	0.029
Taxable income	26,369.872	32,645.647	29,076.491	28,034.211	41,322.108	31,738.924	28,850.210
Turnover	211,931.373	284,864.367	250,800.450	240,193.120	288,590.007	257,956.994	241,327.015

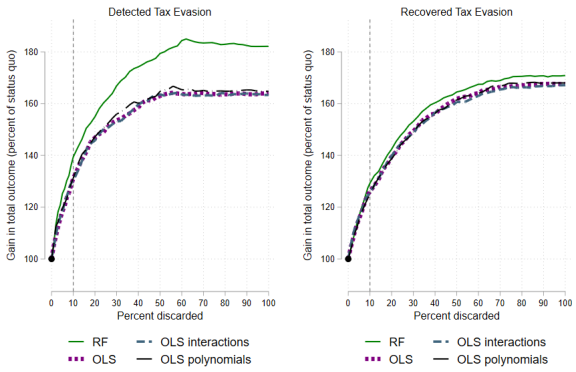
Notes: This table reports the mean characteristics of tax returns (as indicated in the first column) after discarding and replacing 10% audits under Policy B. Columns represent alternative sample selection, depending on the target variable and on the replacement pool. Column 2 ("Status quo") reports the mean for the actual selection. "Office" indicates replacement with a later tax return in the same IRA office; "Office/Sector" and "Office/Income" allow replacement only with tax returns of the same sector of activity and income decile, respectively. Financial accounts are expressed in euro. Other services include health, education and recreational services.

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.[35]

[35]For both outcomes, no income decile remains ever with zero coverage for any replacement level. For tax evasion, we find only one 2-digit sector without coverage (electricity, gas, steam and air conditioning supply) for replacement rates above 85%.

F Additional Figures

Panel A: OLS augmented models



Panel B: LASSO augmented models

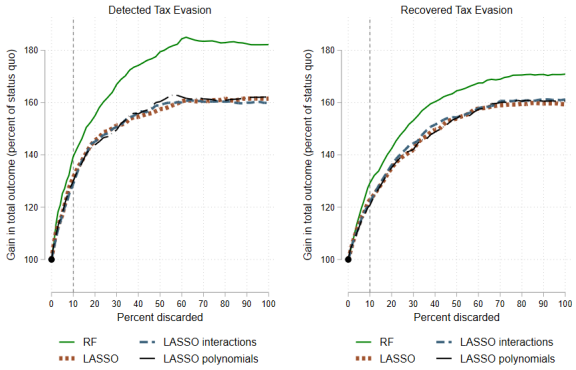


Fig. A7. Gain from Policy B discard and replace by augmented prediction model.
This figure reports on the y-axis the gain in total outcome of discarding a given percentage of early tax audits with the lowest predicted outcome variable and replacing them with an equal number of tax returns with the highest predicted outcome among those audited later (Policy B) by prediction model. The gain reported is computed by using the predictions of OLS models in panel A, and LASSO models in panel B. The gains from OLS and LASSO models considered in Figure 6 are depicted in dotted lines. Gains from augmented specifications including interactions or cubic polynomials of the ten OLS predictors with the highest t-statistic are indicated with dashed and dashed-dotted lines, respectively. The x-axis reports the discarded percentage. The vertical dashed line indicates the 10% level of discarding. The outcome variable is detected tax evasion in left panels and recovered tax evasion in right panels. All values are reported relatively to the status quo total outcome of early audits set at 100 and represented by a dot.

Additional Tables

Table A4
Summary statistics by macro-sector.

Variable	Macro-sector	Mean	Share zeros	SD	p5	p10	p25	p50	p75	p90	p95
Declared income	Services	11269.68	0.14	9145.80	0	0	3725	10336	17681	24870	28113
	Manufacturing	8293.61	0.16	6258.57	0	0	2918	8859	13291	16491	17841
	Wholesale	10796.35	0.13	7806.61	0	0	4470	10883	16845	21632	23615
	Retail	5692.62	0.22	5115.22	0	0	707	5885	9920	12473	13501
	Agriculture	3751.67	0.23	4590.21	-1575	0	152	3221	6928	10176	11562
Detected evasion	Services	15493.50	0.40	133297.77	0	0	0	3570	12520	30177	52737
	Manufacturing	29207.31	0.37	130735.92	0	0	0	5669.5	19699.5	63187.5	123916
	Wholesale	21655.07	0.45	135916.23	0	0	0	2716	12511	31710	61022
	Retail	11030.50	0.42	69969.03	0	0	0	3060	10925	23534	37742
	Agriculture	11879.61	0.45	107575.27	0	0	0	1760	8324	18100	27354
Predicted revenues	Services	44428.90		105558.59	1000	4399	12713	26117	46260	90930	139301
	Manufacturing	84646.54		179129	2650	8116	19828	41631	91291	186639	289925
	Wholesale	78847.28		191862.78	2226	7448	18665	34277	64019	166274	291070
	Retail	105732.84		263592.66	3569	8603	21858	48627	99088	191967	319688
	Agriculture	23667.85		54751.84	0	0	2242	8920	22698	58552	99696

Notes: This table reports summary statistics on the distribution of three relevant variables in different macro-sectors. The variables are: declared income, detected evasion, and predicted revenues by the Sector Studies' model. The sample includes all sole proprietorships that are part of Sector Studies and for each macro-sector excludes the largest firms (measured as the top quartile of declared income).

Table A5
Summary statistics on audit risk.

Macro-sector	Below Threshold	Above Threshold	Share audited	Income share by quintile of predicted revenues				
				1	2	3	4	5
Agriculture	0.021	0.013	0.018	0.891	0.629	0.414	0.263	0.050
Retail	0.023	0.010	0.018	0.156	0.163	0.133	0.097	0.043
Wholesale	0.026	0.014	0.019	0.443	0.432	0.407	0.248	0.062
Manufacturing	0.025	0.015	0.020	0.444	0.335	0.227	0.132	0.045
Services	0.023	0.015	0.018	0.628	0.581	0.522	0.375	0.123

Notes: This table reports summary statistics on the audited share of taxpayers in each macro-sector, distinguishing between those declaring below and above their predicted revenues. In addition, it shows the median declared income share of revenues in different quintiles of the predicted revenues distributions. The sample includes all sole proprietorships that are part of Sector Studies and for each macro-sector excludes the largest firms (measured as the top quartile of declared income).