Algorithms and Bureaucrats: Evidence from Tax Audit Selection in Senegal

Pierre Bachas, Anne Brockmeyer, Alipio Ferreira, Bassirou Sarr*

December 10, 2024

Click here for the latest version of the paper

Abstract

Can algorithms enhance the work of bureaucrats in developing countries? Developing economies are often data-poor environments, where individual bureaucrats have substantial discretion over key decisions, such as selecting taxpayers for audits. Exploiting a trove of newly digitized micro-data, we conduct a field experiment across tax offices in Senegal whereby half of the annual audit program is selected by tax inspectors and the other half is selected by a risk-scoring algorithm. We find that inspector-selected audits are 18 percentage points more likely to be conducted and detect 89% more evasion. Algorithm-selected audits are less cost-effective and do not generate less corruption. Even an ex-post optimized algorithm trained on audit outcomes would have increased aggregate detected evasion only moderately compared to the inspector selection. This is consistent with the inspectors' high skills, the complexity of the task and the imperfection of the available data.

*Pierre Bachas: World Bank Research and EU Tax Observatory, pbachas@worldbank.org; Anne Brockmeyer: World Bank, IFS, UCL and CEPR, abrockmeyer@worldbank.org; Alipio Ferreira: Southern Methodist University, alipioferreira@smu.edu; Bassirou Sarr: Senegal Ministry of Finance. We thank Denis Cogneau, Laurent Corthay, Lucie Gadenne, Janet Jiang, Nicola Limodio, Jan Loeprick, Markus Kitzmuller, Imran Rasul, Dan Rogger, Eduardo Souza-Rodrigues, Gabriel Zucman for helpful comments and discussions, and seminar/conference audiences at Berkeley Haas, IFC, IFS, IIPF, CESifo Public Economics Week, INSPER, TARC Exeter, WB Tax Conference, CMI TaxCapDev Conference, Paris School of Economics, Oxford Centre for Business Taxation, Oxford University Economics Department, PUC Rio, University of Muenster, University of New Mexico, Norwegian School of Economics, NTA Conference, North Texas Economics Conference, and Universidad del Pacífico. We also thank the Senegal Tax Administration (DGID), in particular, Bassirou S. Niasse, Amadou A. Badiane, Oumar D. Diagne, Hady Dieye, Mor Fall, Serigne M. Fall, and Mathiam Thioub. We thank Samba Mbaye, Assane Sylla, and Medoune Sall from the CRDES for their collaboration, Oumy Thiandoum for excellent research assistance, the Paris School of Economics and CEPREMAP for administrative support, and the World Bank, UKAID via EDI, the Centre for Tax Analysis in Developing Countries (TaxDev), and the UKRI through Brockmeyer's Future Leaders Fellowship (grant reference MR/V025058/1) for financial support. The findings, interpretations, and conclusions do not necessarily represent the views of the World Bank, its affiliated organizations, its Executive Directors, or the governments they represent, nor the Government of Senegal. All errors are our own.

1 Introduction

Lower-income countries are often data-poor environments, where many policy decision have traditionally been taken in a discretionary rather than data-driven manner. For instance, individual government bureaucrats may decide which taxpayers to audit, which water treatment facilities to inspect, and for which manufacturing plants to monitor pollution (Khwaja et al., 2011; OECD, 2023). In high-income countries, such decisions are often data-driven, with limited input from individual agents. Data-driven decision models leverage all available information in a systematic manner and are non-arbitrary in theory, but they require high-quality data. Discretionary decisions, on the other hand, can leverage bureaucrats' private information and experience, but could open the door to biases and corruption.[1]

This paper studies whether a data-driven algorithm can improve tax audit selection, compared to discretionary case selection by tax inspectors. Our setting is the tax audit program in Senegal, which aims to target firms with high amounts or high rates of tax evasion. Three features make Senegal a uniquely suitable context for our study. First, tax inspectors have traditionally enjoyed discretion in selecting cases for audits, often without having to provide a rationale for their selection. Second, tax collections in Senegal are only around 15% of GDP and evasion is high, estimated to be at least 30% of true tax liabilities. The potential to improve tax compliance through enforcement actions such as audits is thus large. Third, Senegal recently made important investments in its data infrastructure, e.g. mandating electronic tax returns and digitizing third-party data sources. The information basis for a data-driven audit selection mechanism has hence improved substantially. Senegal's experience in exploiting the new data is broadly relevant for other developing countries where donors are pushing for similar tax digitization efforts.

In collaboration with the Senegalese tax authority, we designed a *risk-scoring* algorithm, aiming to identify firms with high evasion amounts or high evasion rates. This collaboration is the first effort to systematically leverage all available data for tax enforcement. Our algorithm draws on taxpayers' self-assessment declarations for different taxes and on third-party records from customs, public procurement, and firm-transaction reports. For each taxpayer (firm), the algorithm calculates a risk score, composed of multiple indicators that aim to proxy for tax evasion. One class of indicators flags inconsistencies, for instance, when a firm's self-declared sales are lower than its third-party reported sales. Another class of indicators flags anomalies, such as abnormally low profit rates compared to other firms of the same size group and sector. The choice of risk indicators is based on best practices in audit selection and discussions with experts from the World Bank and the International Monetary Fund (IMF).

[1] International organizations hence tend to recommend reforms that limit the amount of discretion given to bureaucrats (Cordova-Novion and Sahovic, 2010; World Bank, 2005).

The audit selection algorithm was implemented at scale in the main tax audit offices in Senegal in the years 2018-2020. Each January, the tax administration sets an annual audit program that in theory should be fully conducted by the audit offices during the year. We refer to the audit program for a tax office and year as an audit *list*. To allow for an experimental evaluation of the algorithm, half of the cases on each list were selected by tax inspectors at discretion, and half were selected by the risk-scoring algorithm. Within each list, the ordering of cases was randomized to minimize the possibility that cases were treated differently because of how they were selected. By design, inspectors knew how a case was selected. Overlap between the inspector selection and algorithm selection was possible but limited in practice.

To quantify the relationship between the case selection method and audit outcomes, we run a horse race between algorithm-selected audits and inspector-selected audits within each list. Concretely, we use an OLS regression of audit outcomes on an indicator variable for algorithm selection and audit list (i.e. office $\times$ year) fixed effects. The outcome measures come from three different data sources. First, we digitized the universe of audit results records which report the audit start and end date, inspectors working on the case, infractions uncovered, and additional tax to pay. Second, we surveyed a subset of audited firms to capture firms' perceptions of the audit process and corruption outcomes. Finally, we use survey data on inspector characteristics, skills and experience to examine how human resources are deployed, as part of the audits are endogenously assigned to tax inspector teams. Importantly, historical audit outcome data was not available in digital format at the audit selection stage. In lower-income countries, audit outcome data is usually stored outside of the tax administration's main IT system, often in paper format. Our risk-scoring algorithm is hence not a trained machine-learning algorithm, but rather one built on economic concepts and logic, exploiting the available data as thoroughly as possible.

We document four sets of results. First, we examine audit outcomes. Contrary to the protocol requiring full program implementation, we find that the overall execution rates of the annual audit program are low. Only 53% of selected firms are actually audited. Inspectors have a preference for auditing cases which they (or other inspectors) selected: the audit execution rate is 18 percentage points lower for algorithm-selected audits. This is partly due to the fact that algorithm-selected firms and inspector-selected firms differ on several dimensions: algorithm-selected firms are smaller, older and less profitable. Yet even after we control for firm characteristics, an audit execution gap of 14 percentage points between the two audit case types remains. Conditional on implementation, almost 90% of audits uncover tax evasion and the detection rate does not vary significantly between inspector-selected and algorithm-selected cases. However, conditional on detection, inspector-selected cases yield 89% higher assessments of taxes evaded on average. Our results are robust across subsamples and time periods.[2]

[2]A silver-lining for the use of an algorithm is the fact that algorithm-selected firms are 13 percentage points less likely to dispute the audit result, compared to a mean share of 84% of firms that dispute and obtain a reduction in the amount

The second part of our analysis puts the audit process under the microscope. Using our uniquely detailed administrative data, we find that algorithm-selected audits use less manpower, are implemented faster by various measures, and are conducted by more junior teams. The implementation of algorithm-selected audits is evaluated less positively by firms, as captured by firms' perceptions of auditors' professionalism, competency and efficiency. We detect no statistically significant difference in the incidence of corruption as reported by taxpayers, and can reject that corruption is more than 0.2 standard deviations lower in algorithm-selected cases. Combining the amount of evasion uncovered and measures of audit costs into a measure of productivity, we find that algorithm-selected audits are on average significantly less productive than inspector-selected audits.

Third, we examine the reasons for the execution and detection gap between inspector-selected and algorithm-selected audits. We can rule out two pre-specified hypotheses for why inspectors might prefer their own selection: that the risk score is not predictive of evasion, and that inspectors lack information on what risks to look for in algorithm-selected audits. In particular, we document that the risk score predicts audit implementation and detected evasion, which confirms that it contains information relevant for audit outcomes. We also show through a cross-randomized information treatment in the desk audit program that providing inspectors with additional information, i.e. risk flags and excel sheets with the microdata, does not increase their likelihood of implementing an audit. We then train a random forest model to predict audit execution, conditional on selection. We find that the model inspectors use to decide whether to audit inspector-selected cases is very different from the model they use for algorithm-selected cases. To show that inspectors' choice of which firms to audit is strategic, we train a random forest to predict evasion, using the sample of all executed audits as training data. Applying to model to obtain predicted evasion for all programmed audits (including those that were not implemented), we find that the realized execution rate is strongly increasing in predicted evasion. Conditional on predicted evasion, execution rates do not differ between algorithm-selected and inspector-selected cases.

Finally, we use the random forest model predicting evasion to examine whether an ex-post optimized algorithm, trained on the outcome data, could have increased the aggregate revenue collected through audits. Concretely, we predict evasion for all firms in the audit program based on pre-existing observables, and rank firms by predicted evasion within each tax office. We then select the top cases, such that the number of cases implemented by each office equals the realized number of audits. This machine-learning algorithm could have increased aggregate revenue by up to 20%. These potential revenue gains are much smaller than the returns from successful machine-learning applications in other contexts, as we discuss below.

Overall, our results highlight the difficulty of improving bureaucrat decision-making through algo-

of confirmed evasion. The lower dispute rate reduces but does not close the gap in confirmed evasion amounts between inspector-selected and algorithm-selected firms.

rithms in lower-income countries, when bureaucrats are highly skilled and engage in a complex task. Our simple risk-scoring algorithm, designed without recourse to audit outcome data, falls far behind the performance of inspector-selected tax audits, and even the machine-learning algorithm could improve audit performance only moderately.

The poor performance of our risk-scoring algorithm is despite our best efforts to draw on all available digital data and to fine-tune the algorithm in three subsequent years based on our lessons from the context, data and results; and despite the fact that inspectors chose which algorithm-selected cases to audit, hence adding their skill to the algorithm selection. Given our close collaboration with and endorsement from senior management at the tax administration, we believe that our results are not driven by inspectors' desire to derail the algorithm. In fact, inspectors were highly motivated to use an algorithm for audit selection before the start of the program. In our inspector survey, 80% of the respondents agreed with the statement that audit case selection should be automated and data-driven. Inspectors also have strong financial incentives to uncover evasion, as they receive bonuses proportional to the uncovered evasion.

Instead, we interpret our results as showing that inspectors are indeed better than the risk-score at identifying high-evasion cases. Inspectors involved in audits of firms are among the top civil servants, highly trained, experienced and incentivized. The data that the algorithm runs on, however, is imperfect. For instance, match rates between different datasets are incomplete, key variables are likely measured with noise, and outcome data to train the algorithm was not available at the onset.

Data limitations are also a key explanation for the limited revenue increase that the machine-learning algorithm achieves. Although our intervention is implemented at scale is a medium-sized country, only 500 full audits were implemented during 2018-2020. This means that the size of the training data for a machine-learning algorithm is limited, while the number of predictors is huge, given then large amount of information available in administrative datasets. In addition, with the exception of a small share of desk audit cases, the training data are not randomly sampled. It is hence unclear how well the predictions of a model trained on selected cases extrapolate out of sample.

In summary, our results highlight the competence of tax inspectors while also pointing towards potential benefits from improved data quality and an extended series of audit results data.

As in any experiment, the specific design choices we made imply some limits to the breadth of our analysis. Our empirical design was intended to capture the difference in audit outcomes between algorithm-selected and inspector-selected cases. The baseline hypothesis was that the likelihood that an audit detects evasion and the detected evasion amount could be increased. As such, our design does not provide direct evidence on (welfare-)*optimal* audit selection, which would need to take into account the evasion uncovered from audits, the real and deterrence effect of audits on audited taxpayers in the medium term, and the potential real and compliance spillovers of audits on non-audited

firms.

This paper is organized as follows. After a review of the relevant literature in Section 1.1, Section 2 presents the context and data. Section 3 presents the design of our intervention. Section 4 presents results on audit outcomes, Section 5 presents results on the audit process, and Section 6 evaluates potential mechanisms. Section 7 tests an ex-post optimized algorithm and Section 8 discusses policy implications and concludes.

1.1 Related Literature

Our work contributes to a recent but rapidly growing literature on how algorithms can enhance or substitute for human decision-making (see Table A.1 for an overview). Machine-learning algorithms have been shown to improve decision-making and welfare in a variety of settings, including judicial bail decisions (Kleinberg et al., 2017), the targeting of health, safety and water quality inspections (Johnson et al., 2023; Glaeser et al., 2016; Hino et al., 2018), hiring decisions for teachers and police officers (Chalfin et al., 2016), and the targeting of tax rebates or public credit guarantees to firms (Andini et al., 2018, 2022).[3] These examples are exclusively from the US and other high-income countries. Our contribution is to test the performance of an algorithm in a low-income country context. As in previous studies, the central agents in our context (tax inspectors) are white-collar workers. However, in our setting, the skill difference between these agents and the average worker in the economy is much larger than in previous studies. This can help explain why the agents in our setting perform so well compared to the algorithm.[4]

In the context of tax audit selection specifically, machine-learning algorithms have shown promise in Italy and the US (Battaglini et al., 2024; Black et al., 2022). In high-income countries, audit outcomes and potential predictors of these outcomes are usually available in digital format. High-income countries can also afford to have a share of their audit program randomized. In lower-income settings, in constrast, audit outcomes are rarely available in digital format, the number of audits is smaller and audits are usually not randomized.[5] In contrast to applications focused on machine learning, we hence test the performance of a risk-scoring algorithm as a potentially more suitable tool for lower-income

[3] The evidence is not unequivocally positive. For instance, algorithms can exhibit biases (Obermeyer et al., 2019), and some tasks, such as providing employee feedback, are more successfully done by humans than by algorithms (Margalit and Raviv, 2024).

[4] Our results also link with Stevenson and Doleac (2024), who study a risk assessment algorithm in the context of sentencing. They find that strict implementation of the algorithm recommendations would have reduced incarceration, but individual judges pursued different objectives and hence did not always adhere to the algorithm.

[5] An exception is Pakistan, where a legal vacuum lead to the randomization of audits, which Best et al. (2021) exploit to show that audits had no causal effect on medium-term compliance. This finding contrasts with Advani et al. (2023) who documents a positive compliance impact of audits in the UK, and Kotsogiannis et al. (2024) who find positive medium-term effects for full audits in Rwanda. For audits of individuals, the welfare impacts have been found to be large, especially at the top of the income distribution (Boning et al., 2023). This last finding connects with our evidence from Senegal, where inspectors have a revealed preference for auditing large firms and maximizing detected evasion amounts rather than detected evasion rates.

contexts. Our study connects with another experiment in Senegal by Knebelmann et al. (2024) who show, in the context of property taxation, that a simple rule to assess tax liabilities leads to more accurate assessments than bureaucrat discretion. A likely explanation for why our results differ is that the tax inspectors we study are elite government agents engaging in a complex task. In contrast, property tax assessors are relatively low-skilled bureaucrats with limited performance incentives, delivering a more mechanical task which can be more easily automated. Our findings also differ from Haseeb and Vyborny (2022), who show that using a proxy means test instead of discretion in allocating cash transfers to poor households in Pakistan improved targeting and welfare. The decision-making agents in their setting are lower-level elected officials, subject to political pressures and with no particular training in identifying the poor.

In addition, our analysis is relevant for the environmental literature that has studied how the targeting of inspections and the allocation of inspectors to cases affect inspection outcomes. Duflo et al. (2018) show that discretionary targeting lowers pollution by more than random targeting of inspections, as random inspections found fewer extreme polluters than discretionary inspections. Building on this finding, we examine whether data-driven targeting can further improve upon discretionary targeting. Furthermore, Duflo et al. (2013) show that, if inspected firms pick their inspectors (or vice versa, as in our context), a conflict of interest emerges which can significantly bias audit outcomes. Our algorithm may hence influence audit outcomes by removing this potential conflict of interest.[6]

Finally, our work contributes to two strands of the literature on state capacity and development. One strand has documented the value of information for improving tax compliance (Kleven et al., 2011; Naritomi, 2019; Pomeranz, 2015) and has studied the potential for new technologies to generate and process information for tax enforcement purposes, with sometimes mixed results (Okunogbe and Santoro, 2022; Okunogbe and Tourek, 2024; Brockmeyer and Saénz Somarriba, 2025).[7] Another strand of this literature has focused on the role of bureaucrats in building state capacity (Besley et al., 2022; Finan et al., 2017). This strand has documented the value of discretion or autonomy for bureaucrat performance (Bandiera et al., 2021; Rasul and Rogger, 2018), but it has also highlighted the risks associated with discretion, e.g. favoritism (Szucs, 2023). For tax collectors specifically, the literature has quantified the impact of performance incentives (Khan et al., 2015, 2019) and the value of local information on taxpayers (Balán et al., 2022; Dzansi et al., 2022). Our study pushes the frontier by comparing the systematic use of observable and digitized data (in the form of the algorithm) with the value of the tax inspectors' own knowledge and experience. Unlike the lower-level tax collectors studied in previous papers, whose performance could be enhanced through the provision of information, the inspectors in our setting outperform even the systematic use of all available information,

[6] We also connect with another strand of the environmental literature which studies the adoption of new technologies such as satellite imagery to detect infractions, e.g. illegal mining (Saavedra, 2023) and deforestation (Assunção et al., 2023).

[7] Other studies have shown how technology can help governments manage expenditure and prevent leakages and waste (Muralidharan et al., 2016; Banerjee et al., 2020).

consistent with their high qualifications and experience.

2 Context and Data

2.1 Tax Policy

Senegal has low tax revenue and experiences widespread tax evasion. According to the World Bank data, tax revenue as a share of GDP is around 15%, as in most other low and middle-income countries. Evasion is substantial, suggesting that improved enforcement could have high returns. The World Bank estimates a tax gap of 5-6% of GDP for Senegal, which is relatively stable since 2017 (World Bank 2024). Senegal's Ministry of the Economy estimates evasion rates of 35% for the personal income tax, 30% for VAT and 26% of the corporate income tax (Faye et al. 2022). Importantly, these estimates are based on the share of evasion that is realistically detectable through enforcement, not on a full-compliance scenario, which would yield much higher evasion estimates.

Senegal's tax structure is typical for a lower middle-income country (see Table B.1). The Value Added Tax (VAT) is the largest source of revenue, representing slightly more than 29% of total tax revenue in 2022, followed by income taxes (29% of total tax revenue) and customs duties (15%). The VAT applies monthly at a standard rate of 18%, with a reduced rate for selected activities (e.g. tourism). The Corporate Income Tax (CIT) is paid annually at a rate of 30% of profits or 0.5% of turnover, whichever is larger. Small firms with a yearly turnover of less than 50 million CFA Francs (about 100,000 USD) are eligible for a simplified tax (*Contribution Globale Unique*, CGU) on turnover, which replaces the previously mentioned taxes and has rates varying from 1% to 8% depending on economic sectors and turnover. As in most other countries, firms withhold the personal income tax (PIT) at source for their employees, also referred to as Paye-As-You-Earn (PAYE).

2.2 Tax Enforcement

The tax administration (*Direction Générale des Impôts et des Domaines*, or DGID) is tasked with enforcing the tax code. Its main enforcement tool is the annual audit program. Audits are implemented by 26 different tax offices, with taxpayers segmented by size, region and sector (Figure B.1). Audits are either full audits, carried out by a team of inspectors at the taxpayer's premises, or desk audits, conducted remotely using the firm's tax returns and third-party data. Almost all audits are targeted at firms.

At the beginning of each calendar year, audit offices are provided with a target number of audits to complete, and prepare a list of taxpayers to audit. This list is reviewed and approved by tax office management and by senior management, usually with very minor changes.

Before our intervention, the selection of cases for tax audits was discretionary in that it did not follow

an explicit rule. No rationale was required for desk audit selection. For full audit selection, some units required a justification, e.g. a review of the firm's audit history or a summary of relevant indicators such as total sales and the profit margin. However, the justification could take different forms and was not systematically required for all cases. Therefore, the criteria used for case selection varied across units and inspectors.

Figure B.2 illustrates the steps in the audit process. After examining a case, inspectors list the detected irregularities and associated penalties and communicate them to the taxpayer in an initial "notification". They can also request additional information from the taxpayer. Upon receiving the notification, taxpayers have 30 days to respond.[8] The inspector then examines the response and has 60 days to prepare and send a "confirmation" with the confirmed irregularities and penalties and the final amount to pay. The inspector then generates a revenue order for the tax collection unit, which requires the taxpayer to make a payment within ten business days.

2.3 Data

During the last decade, DGID invested in digitizing its tax information and required taxpayers to file electronically. As a result, data availability has expanded substantially. Our study draws on data from three administrative sources and two surveys. The administrative data include the self-assessment declarations filed by taxpayers, third-party reports, and audit reports. The tax declarations and third-party data were used in the calculation of the risk scores and had just been digitized at the onset of the project. The audit reports were hand-collected and digitized by our research team after the end of the intervention, in 2021 and 2022. To study how audits were implemented, we complement the administrative data with a tax inspector survey and a taxpayer survey, designed by the research team. Only aggregated results from these surveys were shared with the tax administration.

Tax Declarations. Table 1, Panel A, provides an overview of the available tax declarations. Our primary sources of information are the declarations for the CIT, VAT and PAYE declarations, covering the period of 2014-2019. The CIT data covers about 5,000-7,000 firms per year, and the VAT data about twice as many.[9] The PAYE data allows us to calculate the number of employees and the aggregate wage bill for each firm. Around 2,000 firms file tax under the simplified regime CGU. Less than 150 financial institutions pay the *Taxe sur les Activités Financières* (TAF), a VAT-substitute for the financial sector.

Third-Party Reports. Table 1, Panel B, describes the information about taxable transactions and activities that we obtain from third parties. Imports and exports are recorded by the customs authority, procurement from state institutions is recorded by the treasury, and firm-to-firm transactions are

[8] Failure to respond is interpreted as the taxpayer agreeing with the inspector's findings.

[9] The number of VAT filers is higher than the number of CIT filers because self-employed individuals and unincorporated firms may file VAT but not CIT.

recorded in VAT annexes that firms file since 2017. While these data are provided at the transaction level, we aggregate them at the firm-year level to merge with the tax declarations.

Audit Reports. We collect audit process and results data in two ways (Table 1, Panel C). First, we digitized all audit result reports for 2017-2020. The reports cover all process steps from audit announcement to notification, confirmation, and payment request. They contain the name(s) of the inspector(s) who conducted the audit, the taxes verified in the audit, infractions detected, evaded amounts, applicable penalties, and the dates of each step in the audit process. In addition, we asked inspectors to report audit information in an excel sheet pre-filled with their list of audit cases. These excel files contain information on audit cases that is not directly observed in the administrative audit reports, such as the number of days that an inspector spent working on a case.

Tax Inspector Survey. In 2017, prior to our intervention, we conducted a detailed survey among 97 tax inspectors involved in conducting audits, capturing information about their demographics, employment history, perceptions of the audit function, methods for audit selection, and use of different sources of information. The sample includes almost all tax inspectors involved in tax audits at the time, except for those who were unavailable throughout December 2017 and January 2018. Descriptive statistics are presented in Section 2.4.

Taxpayer Survey. After the completion of the 2018 and 2019 audit programs, we surveyed 742 firms in the Dakar region, most of which had been selected for audit as part of the program. We conducted the survey in two waves, from October to December 2020 and from March to May 2021. The survey allowed us to elicit taxpayers' experience with the audit process, including perceptions of corruption, the audit risk, and their general perception of the tax administration. The first survey wave focused on the 2018 audit program. We sampled all full audit cases of our program; a matching number of desk-audit cases, randomly sampled and stratified across offices to preserve the relative distribution of planned audits across offices; and a matching number of non-audited cases, similarly stratified. This yielded a total list of 1226 targeted firms. For the second survey wave, we targeted an additional 702 firms, which are again equally divided into the full audit program cases, desk audit cases and non-audited cases, all selected by stratified random sampling.[10] We asked surveyors to prioritize interviews of full audit cases (27% of the sample), and the response rate for these cases was slightly higher (42% vs 37% for other cases).[11]

[10] In the second wave, the total sample was 937 firms, which includes 702 new firms plus 235 firms that had not been contacted in the first wave.

[11] The random sample was restricted to firms that filed CIT. In our current analysis, we do not use responses from firms that were not part of our audit program.

2.4 Tax Inspectors

The bureaucrats central to the tax enforcement process are tax inspectors and their managers, who are usually former tax inspectors themselves. Tax inspectors form an elite corps in the public administration and often proceed to hold influential political roles. At the time of writing, both Senegal's president and the prime minister are former tax inspectors. Selection into the tax administration function is highly competitive. Most inspectors have completed a Master's degree from the *Ecole Nationale d'Administration* (ENA), and over 60% hold a PhD. According to our survey, the mean and median inspector has ten years of professional experience. Over 70% of inspectors report being satisfied with their work at DGID, and a similar fraction report being motivated by their work.

Tax inspectors are very well-paid. Annual base salaries were typically 5-6 Million FCFA in the period we study, and stand at 7-8 Million FCFA (approximately 12,000 to 13,500 USD) in 2024. In addition, inspectors receive a bonus that depends on the amount of evasion uncovered in audits and that is usually higher than the base salary. The median (mean) bonus for inspectors in our data was 7 million FCFA (10.6 million FCFA). The size of the bonus varies across inspectors (e.g. with seniority), but quantifying the exact strength of incentives is difficult, as it depends on the endogenous assignment of inspectors to cases. There is no institutional reason for incentives to be correlated with the case selection mechanism. Appendix B.3 discusses the bonuses and performance evaluation in more detail.

When selecting cases for audits, inspectors typically start from a long list of cases to consider, roughly half of which are randomly chosen cases. Inspectors then examine the long-listed cases to derive a short list. Figure 1, Panel A, shows the objectives that inspectors pursue in audit selection. The most commonly named objective was 1) a diversity of audits (i.e. a desire to create the perception of a non-zero audit probability for most taxpayers, which we interpret as an interest in the deterrence function of audits), 2) the detected evasion amount, and 3) the detected evasion rate.[12] Panel C shows that there is a clear ranking in the types of information used in audit selection. The most commonly used data sources (over 90% of all cases) are the self-assessment declarations, closely followed by third-party data (60-80% of cases), with soft information being the distant third data source (around 40% of cases). Over 60% of inspectors report finding quantitative information more useful than soft information, while 22% prefer soft information, and the remaining respondents consider the two sources equally important. Most inspectors use excel or other softwares to analyze taxpayers' quantitative information. In response to the survey's open-ended question about proxies of evasion, frequently-mentioned indicators include low turnover compared to other variables, frequent losses, and high VAT credits. Respondents almost unanimously agree that turnover is the single most important variable they consider.

[12] Panel B illustrates the trade-off inspectors face between maximizing the detectable evasion amount and the evasion rate. Roughly 45% of inspectors prefer to audit larger firms, even if they yield lower detected evasion amounts and evasion rates, but a similar fraction of inspectors trade off these two objectives.

Inspectors are critical of their status quo audit selection method and keen to improve. In our survey, 55% of respondents consider that they do not have sufficient data for audit selection, and almost all want more advice on how to select cases. Concretely, 85% of respondents agree that a more systematic analysis of data would be beneficial to audit selection and 70% think that an automated selection based on risk indicators would be a good idea.[13]

To summarize, tax inspectors are highly-trained elite civil servants, strongly incentivized via bonus payments to detect evasion through audits, and motivated to use new tools to improve audit selection and hence audit performance.

3 Experimental Design

3.1 Intervention: Risk-Based Audit Selection

Given demand for a more systematic audit selection method from tax inspectors and DGID management, we collaborated with DGID to design and introduce a risk-scoring algorithm for audit selection. The objectives of the algorithm were a) to ensure that audit selection followed objective and quantifiable criteria, and b) to increase the detected evasion amounts (plus associated penalties and fines) and detected evasion rates.

We faced two types of constraints in designing the algorithm. First, the algorithm needed to be intuitive, transparent, and easy to communicate to policymakers and tax inspectors. Second, the algorithm could not be trained on past audit data, as outcome data was only available for a small and selected subset of audits.[14] Given these constraints, we designed an algorithm based on intuitive indicators discussed with and validated by the tax administration. The choice of indicators drew on technical assistance work done by the World Bank in Pakistan and Turkey, best practices shared by the tax administration in Denmark, and feedback from experts at the World Bank and the IMF. Our strategy is broadly applicable in other lower-income countries where similar constraints on audit selection are likely to bind.

Our algorithm generates a risk score for each firm that allows us to rank firms within a tax office. The risk score combines two classes of risk indicators at the firm level: inconsistencies and anomalies. Inconsistencies are *within-firm indicators* that flag taxpayers with inconsistent information across different datasets. For example, an inconsistency arises if the self-reported turnover is lower than the third-party reported turnover constructed as the sum of exports, procurement contracts, and purchases

[13] This is consistent also with the fact that inspectors spend a non-trivial share of their time, 27%, to decide which cases to audit, while they spend 52% of their time on the actual audits, and the remainder on administrative tasks and complaints. An "effortless" selection via the algorithm would free more time to conduct audits.

[14] Accessing and systematically digitizing all paper records of audit results, which came in a variety of formats and were considered sensitive information by the administration, took several years and a substantial investment of time and resources by the research team.

declared by other firms. In contrast, anomaly indicators are *across-firm indicators* that flag outlying behaviors potentially associated with tax evasion relative to the firm's peers. For example, one anomaly indicator flags firms with a low profit margin relative to firms of the same economic sector and similar size. Each inconsistency and anomaly is captured in the form of a ratio. We assign "points" from 1 to 10 based on the deciles of the ratio, to take into account the severity of the irregularity. We then aggregate the points using importance weights which we assigned. We weighted inconsistencies higher than anomalies as we were more confident that inconsistencies reflect noncompliance, while anomalies could also reflect temporary economic difficulties or poor management. Appendix C describes the risk scoring algorithm in detail.

3.2 Study Design

To evaluate the performance of the risk-scoring algorithm, we introduced algorithm-selected audits in all audit lists. We asked each office to report the total number of planned audits for the year. Inspectors were then tasked to select half this number of cases following their discretionary method, while the algorithm selected the other half of cases. To obtain the algorithm selection, we ranked firms by risk score within a tax office and selected the top cases on the list until the required number of cases was reached. We had access to the inspector-selected cases before running the algorithm selection, allowing us to tag overlapping cases selected by both methods.

An alternative research design would have been to randomize the use of the algorithm across tax offices. This was infeasible in our context, given political constraints and the small number of tax offices conducting audits (26), most of which joined our program only in the third year. The combination of both selection methods is also the more realistic policy, as fully replacing discretionary selection risks reducing revenue and can breed resistance among inspectors. Instead, with our design, both inspectors and managers were initially enthusiastic to use the algorithm, as evidenced in workshops we held with all tax offices.[15] Our design effectively allows us to study the combination of algorithm selection with inspector selection. Indeed, although the annual audit program was supposed to be fully implemented, in practice inspectors enjoyed discretion to pick which firms on the annual program list to audit. We discuss this in more detail in Section 4.

We randomized the order of case types displayed on each list and asked inspectors to adhere to the proposed ordering. We intended to ensure that both case types were treated similarly in terms of timing and effort exerted to complete the audit.[16] Inspectors were provided with a protocol emphasizing

[15] Indeed, although our project ran only from 2018 to 2020, DGID still asked us to generate the algorithm selection in 2021 and 2022.

[16] Specifically, the order of cases was randomized across all selected cases for an office/inspector in 2018. In 2019 and 2020, the first case on each list was inspector-selected, with subsequent cases alternating between the two selection methods. Cases were randomly allocated to slots on the list. For desk audits, where each inspector selects their own list, the algorithm-selected cases were randomly distributed across inspectors in a tax office.

the importance of conducting all audit cases with the same rigor.

The intervention started with the four large and three medium taxpayer offices in 2018 and included four regional tax offices in 2019 and 2020. The remaining regional tax offices joined the program in 2020, but we were not able to collect outcome data for these offices. Our evaluation is hence focused on the eleven offices with complete data. Each year, the algorithm's indicators and weights were slightly updated. Tables C.2 and C.3 displays the number of cases selected by year, tax office and selection method. By design, the number of inspector-selected and algorithm-selected cases was approximately the same for most years and tax offices, with some deviations. We discuss these deviations in Section 4.4 and show that our results are not sensitive to them.

3.3 Additional Treatments for Desk Audits

Desk audits are more numerous than full audits, and are conducted by individual inspectors (or pairs of inspectors), so that there is no endogenous assignment of inspectors to cases. We took advantage of these features to introduce two additional treatments in desk audit lists. First, in the years 2018 and 2019 we selected some cases at random (within the pool of firms of the respective tax office) and included them in the audit lists. These random cases serve as a benchmark to assess the quality of the algorithm because inspectors were unaware that these cases were randomly selected. Instead, both randomly and algorithm-selected cases were presented as having been selected by the "new method". This approach avoids the possibility that inspectors ignore randomly selected cases, deeming them low-return. At the same time, it may lower inspectors' enthusiasm for algorithm cases if the random cases were consistently worse in terms of audit yields.

Second, the desk audit program is accompanied by an "information treatment" cross-randomized across all case types (inspector, algorithm, and random). The treatment consisted in providing inspectors with case-specific information for two-thirds of desk audits: for one-third of the cases, inspectors received a summary report containing the three main risks flagged by the algorithm (e.g. abnormally low profit rate, turnover lower than third-party reported turnover); for another one-third of cases, they received the same information plus an excel spreadsheet with the firm's tax declarations and third-party data for the last four years (i.e. the data used by the algorithm); and for the remaining third of cases they received no additional information. With this intervention, we aimed to test whether providing high-quality, readable information improves audit implementation and performance.

3.4 Empirical Strategy

To assess the implementation and performance of algorithm-selected cases, we compare the outcomes of these cases to those of inspector-selected cases within the same audit list. Concretely, we estimate

the following model via Ordinary Least Squares:

$$y_{i\ell} = \beta_0 + \beta_1 Algorithm_{i\ell} + \beta_2 Overlap_{i\ell} + \beta_3 Random_{i\ell} + \gamma_\ell + \varepsilon_{i\ell}, \quad (1)$$

where $y_{i\ell}$ is the outcome of an audit for firm i selected in audit list ℓ, the dummies $Algorithm_{i\ell}$, $Overlap_{i\ell}$ and $Random_{i\ell}$ indicate algorithm-selected, overlap and randomly-selected cases, γ_ℓ denotes the list fixed effects and $\varepsilon_{i\ell}$ is a conditional mean zero error term. The audit lists are tax office-year specific for full audits, and can be inspector-year specific for desk audits. The dummy for random selection is applicable only for desk audit cases. Note that overlap cases appeared on the lists sent to inspectors as inspector-selected cases. The list fixed effects allow us to control for any office and year-specific factors, such as the identity of the office manager, the fact that more experienced inspectors are more likely to be assigned to the large and medium taxpayers offices, or that inspectors in regional offices spend a larger share of their time on non-audit administrative tasks, such as taxpayer service and information campaigns. We use robust standard errors to perform inference.

The main coefficient of interest is β_1, which captures the difference in audit outcomes between algorithm-selected cases and inspector-selected cases, the omitted category. A fair comparison between the two selection methods requires that a) the number of cases selected with both methods is equal and b) inspectors' performance incentives were the same for both case types. We have no reason to believe that condition b) was violated in our context. Our study design attempted to ensure that condition a) is met. While there are some deviations from this condition, we demonstrate below the robustness of our results to additional tests for audit lists that did not meet the condition.

Nonetheless, any difference in audit outcomes between algorithm-selected and inspector-selected cases we observe can be due to a combination of a change in the type of taxpayer selected and a possible change in inspector type or skill set. This is because tax office managers assign full audit cases to teams of inspectors. Although we emphasized that all audit cases should be treated in the same way, we cannot preclude the possibility that managers take into account the case type when assigning inspectors with different skill sets to cases. We investigate this further below.[17]

3.5 Descriptive Statistics on Selected Firms

Table 3 examines the characteristics of firms selected for audit. Panel A compares the characteristics of firms selected by inspectors to the general population of firms. Inspector-selected firms are orders of magnitude larger than the average firm appearing in the administrative tax data (column 1), more likely to engage in import or export (column 4), and slightly though not significantly more profitable (column 3). Inspector-selected firms do not significantly differ from other firms in their location or

[17] Taxpayers were not systematically informed how they were selected for audit, and we have no evidence that inspectors shared this information with them. We thus do not expect a change in taxpayer behavior.

age (columns 5-6).

Panel B compares firms selected by the algorithm to those that were selected by inspectors, using equation 1. Algorithm-selected firms are smaller than firms selected by inspectors, though still substantially larger than the average formal firm (column 1).[18] They also exhibit a lower profit rate than inspector-selected firms, but are more likely to engage in international trade (columns 3 and 4). The difference in profitability appears by construction, as one of the indicators for our risk score is low profitability compared to other firms in the same sector and size group. Our survey data also indicate that algorithm-selected firms have a higher share of sales in cash than inspector-selected firms (column 8). The algorithm is blind to this firm characteristic unobserved in the administrative data, but inspectors may be able to asses or proxy for it. A high perceived share of cash sales may lead them to avoid auditing a firm, as it would be difficult to detect and document tax evasion on cash sales.

4 Audit Outcomes

We now examineine the association between algorithm selection and audit outco. First, we consider the association between the selection mechanism and audit execution, given that only a part of the programmed audits were implemented. Second, we analyzeevasion detection, evasion amount, and taxpayer dispute. Third, we examine results for the resource cost of audits.

4.1 Audit Execution

Based on the inspectors' audit reports, we can observe which cases on the list were started and which ones were not. We consider an audit as executed if the case was started, even if it did not lead to a fine. Overall, only 53% of cases selected for full audits and 33% of cases selected for desk audits are actually audited. The implementation rates vary slightly across years but are not lower in later years. Hence, the incomplete implementation is not due to the fact that some audits are implemented in later years than they were scheduled. We consider all audit data reported until the end of 2021, and merge this information with audit program lists for 2018, 2019 and 2020.

Table 4 summarizes our main results. Full audit cases selected by the algorithm are 18 percentage points less likely to be implemented than inspector-selected cases, corresponding to a 34% reduction compared to the mean implementation rate (column 1). "Overlap" case, which were selected by both the inspectors and the algorithm (but which appeared as "Inspector selected" on the lists) were 15 percentage points more likely to be implemented than cases selected only by inspectors.[19]

[18] Figure C.1 provides a graphical illustration of the gradient between firm size and audit selection within tax offices.

[19] On the other hand, replacement cases, which appeared at the bottom of the list sent to inspectors and which were explicitly marked as replacements for algorithm-selected cases that turned out to be void (e.g. taxpayers that had become inactive or were not reachable) were 47% less likely to be implemented than other algorithm-selected cases (coefficients not shown). These cases were labeled as replacement cases (rather than algorithm x replacement) on the list sent to

The results for the desk audit program are qualitatively similar, but the differences are much smaller (columns 2 and 3). Algorithm-selected desk audits were about 4 percentage points less likely to be implemented, which corresponds to 13% of the mean. Our preferred specification is the one controlling for inspector fixed effects (column 3). Overlap cases are again more likely to be implemented, though the coefficient is not statistically significant. Random cases are no less likely to be implemented than algorithm-selected cases, likely because the two case types appeared in the same way to inspectors (as selected by "new methods") and the expected return of a case might not be clear to inspectors.[20]

4.2 Audit Yield

Table 4, columns 4-6, show the relationship between the selection method and the probability of evasion detection, conditional on implementation.[21] Overall, 89% of full audits and 73% of desk audits detect some evasion. For full audits (column 4), algorithm-selected cases are 4 percentage points more likely to detect evasion, and overlap cases are 8 percentage points more likely. Only the point estimate on overlap cases is economically meaningful and marginally statistically significant.

For desk audits (column 6), algorithm-selected cases are 4 percentage points less likely to detect evasion, and overlap cases are 8 percentage points less likely to do so, although neither of these differences is statistically significant. The point estimates on random and replacement cases are relatively precisely estimated zeros, suggesting that inspector selection does not perform better than random selection. Overall, these results indicate that there is no meaningful relationship between the case selection method and the likelihood that an audit detects evasion. However, conditional on positive detection, we find that algorithm-selected full audits detect significantly smaller amounts of tax evasion plus fines (columns 7-9), though this result is not significant for desk audits once we control for inspector fixed effects (column 9).[22]

As the analysis in columns 4-6 and 7-9 is conditional on audit execution and detection of evasion, respectively, it is unclear what the results would look like in the full sample if all planned audits had been conducted. To examine whether our results are affected by sample selection, we conduct a Lee (2009) bounds analysis. When the outcome is the binary detection dummy, we randomly trim a share of cases among the algorithm-selected firms with detection (without detection) to obtain a lower

inspectors, but DGID audit lists did not previously include replacement cases. So it is reasonable to assume that inspectors perceived these cases as algorithm-selected cases of lower priority ranking, given their role as a replacement.

[20] Replacement cases are significantly less likely to be implemented but the magnitude of this effect is smaller for desk audits than for full audits. This is likely due to the fact that the share of replacement cases in total cases is larger for desk audits than for full audits.

[21] Intent-to-treat analysis that do not condition on audit implementation are shown in Table D.1.

[22] Similarly, we do not detect statistically significant differences in the detected evasion rate, measured either as a share of liability or as a share of the firm's mean turnover across several years, by selection method. We show these results in Table D.3.

(upper) bound. When the outcome is the log(evasion), we rank cases by the outcome variable among algorithm-selected cases, and trim log(evasion) at the top (bottom) cases to obtain the lower (upper) bound estimates. The share of cases to trim is list-specific, calculated based on the the list-specific attrition rates. The bounds are noted in square brackets below the main point estimate on algorithm selection and are consistent with our main results. We cannot reject the null of no correlation between the selection method and detection of evasion, but we find a negative and statistically significant association between algorithm-selection and the amount of evasion for full audits.

To summarize, we find that algorithm-selected audits are less likely to be implemented, but similarly likely to yield a detection of evasion when implemented. Conditional on detection, algorithm audits are associated with slightly smaller amounts of evasion. In general, results are starker for full audits compared to desk audits. This is consistent with the fact that full audits are more costly and involve higher stakes for both the administration and the taxpayer.

4.3 Dispute of Audit Outcomes

Once an audit concludes with a detection of tax evasion, the taxpayer receives a notification and the possibility to dispute the audit results. After dispute and negotiation with the tax administration, a confirmation of the audit results is issued. The confirmed amount of evasion is usually lower than the notified amount. This may indicate errors in how the audit was conducted, uncertainty over the exact amount evaded, or collusive behavior between taxpayers and inspectors to alleviate the penalties.

Algorithmic case selection might reduce uncertainty and collusive behavior. To test this hypothesis, we express the confirmation amount as a share of the notification amount, and plot its distribution for inspector- and for algorithm-selected cases, separately for full and desk audits (Figure 3).[23] For most audits, the confirmation amount is substantially lower than the notification amount. The confirmation amount matches the notification amount in only 22% of completed full audits and 29% of desk audits. For full audits, the mean and median of the confirmation/notification share are 40% and 18% respectively. For desk audits, these figures are 62% and 58%, indicating slightly less pushback against the notification amount. For full audits, the figure suggests that inspector-selected cases are subject to more pushback against the notification amount than algorithm-selected cases.

Table D.4 shows these results in regression format. Among full audits for which both a confirmation and a notification is present in our data, algorithm audits are 13 percentage points more likely to have matching confirmation and notification amounts.[24] This difference in dispute also has an impact on the final difference in the detected evasion amount between the algorithm and inspector-selected audits. For full audits, algorithm audits have a 35% lower notified amount and only a 19% lower confirmed

[23] The figure excludes cases with no confirmation, which represent about a third of these cases.
[24] The result is very similar when we control for firm size.

evasion amount than inspector-selected audits (columns 3 and 5).[25] These results suggest that the lower likelihood of disputes for algorithm audits might improve audit outcomes from a revenue raising and human resource perspective.

4.4 Robustness Tests

We now examine the robustness of our results across subsamples. This allows us to show that our results are not driven by shortcomings in the implementation of our intervention, and that there is little heterogeneity in the results across tax offices and time periods.

First, we tackle the issue that the number of inspector-selected and algorithm-selected cases in a list sometimes deviated from the intended 50-50 split. The most important deviation occurred in the Large Taxpayers Office, where in 2019 and 2020 the tax administration decided that 30% of the full audit cases would be selected by the algorithm, instead of 50%. Other deviations occurred when managers adjusted the number of cases selected at discretion slightly upwards or downwards. If both inspectors and the algorithm had ranked cases by perceived risk and had selected cases following the rank order, the average riskiness of selected cases would be higher when a smaller number of cases was selected. We hence rerun our results for the following sub-samples: a) lists with an identical number of inspector-selected and algorithm-selected cases (as they appear to the researcher), b) lists with an identical number of inspector-selected and algorithm-selected cases (as they appear to the tax inspectors), and c) lists in which the number of algorithm cases is weakly smaller than the number of inspector-selected cases. The latter subsample is less restrictive than the first two, but also less clean from an identification perspective, as it may give the algorithm cases an advantage.[26]

Second, we recognize that there is wide heterogeneity across tax offices and years. The Large Taxpayer Unit has the best inspectors and the lowest number of firms per inspector. Although large firms are more complex than smaller firms, inspectors are more familiar with them. Large firms are also expecting an audit every four to five years. The association between audit selection method and outcomes may thus be different in the LTU than in other tax offices which concentrate many more firms per inspector. We hence rerun our analysis excluding the Large Taxpayers Unit.

Third, the experiment's implementation differed in each of the three years. The year 2018 was the first one, and inspectors may have felt the cost of the novelty of the algorithm. In 2020, both firms and tax inspectors were affected by the pandemic. The 2019 implementation is the cleanest from a

[25] For desk audits, there are no statistically significant differences in dispute of outcomes between inspector and algorithm-selected audits.

[26] The difference between sub-samples a) and b) comes from a methodological change over time in how overlap cases (chosen by both the inspector and the algorithm) were dealt with. In 2018, we ran the algorithm to select a number of cases equal to the number of inspector-selected cases, allowing for overlap cases. The overlap cases appear to inspectors as inspector-selected cases. Thus, the number of "pure" algorithm cases appearing on their list is weakly lower than the number of inspector-selected cases. In 2019 and 2020, we corrected for this by adding an additional algorithm case for each overlap case. Overlap cases still appear on the inspectors' lists as inspector-selected cases.

research perspective, as we carefully prepared and tightly monitored the execution of the intervention, including by asking inspectors to report audit execution steps in pre-filled excel files. We thus re-estimate the results for the 2019 program only.

Figure 2 depicts the coefficients on the algorithm selection indicator from Equation 1 for the three main outcomes: the probability of starting the audit, the probability of detecting positive evasion, and the amount of detected evasion (plus fines, in log points). Panel A is for full audits, and Panel B for desk audits. The two panels show the baseline coefficient with two types of standard error computation, and the various subsamples as discussed above and in the legend.

The figure shows that the point estimates are stable across subsamples, and that the findings are qualitatively and quantitatively robust. For full audits, algorithm cases are less likely to be conducted, no more likely to detect evasion (though point estimates are always positive), and conditional on detection, algorithm-selected cases uncover significantly smaller evasion amounts. For desk audits, however, the point estimates are almost always small and statistically indistinguishable from zero. The negative association between algorithm selection and audit execution is only present in the full sample, but disappears when we limit the analysis to lists with more comparable numbers of algorithm and inspector-selected observations. Our inability to detect statistically significant differences for desk audits is not because of a lack of power. In fact, the number of desk audits is higher and the variation of outcomes among these cases smaller than for full audits. Our experimental design also allows a within-inspector comparison.

5 The Audit Process

5.1 Time and Human Resource Costs

We next examine how human resource allocation and the length of audits varies conditional on the audit selection method. Table 5, Panel A show the results for the number of inspectors and proxies for audit length, using the sample of implemented audits. The most accurately captured measure of audit cost is the number of inspectors working on an audit, which is reported in the administrative data. Column 1 shows that algorithm-selected full audits are composed of teams that are about 10% smaller than the average team of 2.9 members. In addition, teams for algorithm-selected audits have lower seniority, as measured by their age and length of service (Table E.1).[27]

Table 5, Columns 2-6, show the association between audit selection and three different measures of audit length. Overall, all measures suggests that algorithm-selected full audits are shorter. In column

[27]These differences in team composition are consistent with the fact that, for desk audits, we find that inspectors with above-median experience have lower execution rates and are less likely to favor the use of an algorithm for audit selection (Table E.2). We do not detect any correlation between overall list-level execution rates and algorithm execution rates (Figure E.1).

2, the outcome is the audit duration as reported by firms subject to a full audit. The mean duration is 28 days, and algorithm-selected audits are nine days faster on average, which is an economically large difference. In columns 3-4, we use as outcome the audit duration as measured by the difference between the date of notification and the audit start date. This measure captures the length of an audit but not the time inspectors actively spend working on the case, as they typically work on several cases simultaneously.[28] The average full audit takes 166 days until notification. Algorithm-selected audits are completed 25 days faster. The point estimate is not statistically significant at conventional levels, but the lower Lee bound is significant. For desk audits there is no statistically significant difference in speed between algorithm and inspector-selected audits.

Finally, in columns 5-6, the outcome is the self-reported number of days that inspectors work on a case. For full audits, we multiply the reported number of days by the number of inspectors working on a case. Full audits require 188 inspector-days while desk audits require only 9 days on average. Although the point estimates on the algorithm dummy are negative, the difference between algorithm and inspector cases is not statistically significant, likely due to a reduced sample size and to noise in the outcome. For full audits, the outcome is only reported for 51 cases, and it is unclear whether inspectors were referring to their own time investment or to that of the full team.

The audit duration measures in columns 2-6 are not consistently available for all executed cases (for example, the self-reported number of days working on case was only filed by a subset of inspectors). However, we investigate whether the probability of reporting the data differs across selection method and found no difference, as shown in panel B of Table 5.[29]

In addition to the amount of detected evasion, audit reports indicate the specific year of the infraction and a short description.[30] To test if the scope of inspections varied by selection method, we run regressions using as outcomes the number of years covered in the audit report, the number of detected infractions, and the share of fines to evasion amount, which indicates the severity of detected infractions. The results show that in algorithm cases, inspectors systematically reported fewer years with any infraction and fewer infractions, both for full and desk audits (Table E.3, columns 1-4). All differences are highly statistically significant. There is little evidence for systematic differences in the fine-evasion ratios, as indicated both by the small coefficients and the fact that the point estimate for full audits is not statistically significantly different from zero (Table E.3, columns 5-6).[31]

[28] As discussed above, a dispute of audit results can extend the duration between notification and confirmation.

[29] The results in Table 5 are almost unchanged when we control for turnover.

[30] By Senegalese law, an audit can investigate tax declarations up to four years prior to the audit start date.

[31] Table E.4 shows that there are no statistically significant differences between algorithm and inspector-selected audits on a range of other outcomes that capture the severity of the detected infractions.

5.2 Audit Productivity

Given the large differences in audit outcomes and audit process measures, the question of whether algorithm-selected audits may be more cost-effective than inspector-selected audits arises. Table 6 shows the results for our main specification in Equation 1, but uses audit productivity measures as outcomes, as per the column titles. We divide the log uncovered evasion by the different audit cost measures presented in the previous section: the number of inspectors working on a case (columns 1 and 2), the number of days from opening to closing a case from the administrative data (3 and 4), the audit length in number of days from the survey data (5 and 6) and the self-reported number of days the inspectors worked on the case (7 and 8). In all estimations except in column 8, the association between algorithm selection and audit productivity is negative. The coefficient on algorithm selection is statistically significant when using the number of inspectors as a cost measure, or the number of days from the administrative data (for full audits). This means that, although algorithm audits use fewer resources, the difference in evasion is larger than the difference in audit cost, leading to lower productivity for algorithm audits.

However, the averages captured by the regression coefficients mask substantial heterogeneity across cases. Figure E.2 shows distributions of the three main audit productivity measures, for algorithm-selected cases and for inspector-selected cases. The productivity distribution for inspector-selected cases is shifted slightly to the right for all measures. For our preferred measure of productivity — the evasion amount divided by the number of agents working on a case — the two distributions are also statistically significantly different as per the Kolmogorov-Smirnov test statistic, consistent with the results in Table 6. Yet there is substantial overlap between the two distributions, suggesting that a large share of algorithm-selected cases have higher productivity than the marginal inspector-selected case. This rationalizes the fact that implementation of the inspector-selected set of audits is incomplete and some algorithm audits are implemented.

Audit productivity is positively correlated with firm size, consistent with the preference given to larger firms in audit selection and execution (Appendix Figure E.3).

5.3 Taxpayers' Perceptions

An alternative measure of inspectors' effort and performance comes from the taxpayer's point of view. We surveyed around 600 firms that had been selected for audits, and asked them about their perceptions of interactions with the tax authority. The objective was to consider the taxpayer's perspective of the auditing process, in terms of efficiency and incidence of corruption, as well as their beliefs concerning what behaviors might trigger an audit and the capabilities of the tax administration.

We construct two indices to capture respectively the perceived efficiency of audits and the perceived incidence of corruption of audits. Each index is built by combining three questions following the

procedure in Anderson (2008). To measure efficiency, we combine the questions on taxpayers' assessment of the auditor's (i) technical knowledge, (ii) efficiency during the audit, (iii) capacity to uncover all evaded taxes. For corruption, we combine the questions on the perception of (i) the auditor's dishonesty during the audit, (ii) the frequency of bribes paid to inspectors (among firms similar to the respondent's), (iii) the existence of a preferential tax audit treatment for firms connected with the tax administration.

Obtaining information on the incidence of corruption is challenging (Sequeira, 2012). Yet, taxpayers report that bribes are paid in 16% of audits (64% response rate), and 27% of respondents think that connected firms receive preferential treatment (89% response rate). Inspectors obtain an average score of 6.4/10 on honesty, lower than their score of 7.3/10 on knowledge, but equal to their score on efficiency.

Table 7 shows the regression results for the taxpayers' responses, conditional on the selection method that led to their audit. Panel A uses the sample of all interviewed firms that were also in the audit selection and reported a recent audit, while Panel B conditions on firms that were audited as per our audit reports.[32] Columns (1) to (3) show results of taxpayers' evaluation concerning the efficiency of the audit. For full audits (column 1) firms selected by the algorithm report a significantly lower audit efficiency score, compared to firms selected by inspectors. The magnitude is large: the efficiency index score is 0.38 standard deviation lower for algorithm selected cases. The coefficient is less negative when we condition on firms that were actually audited as part of the program, and no longer significant (Panel B). Firms selected by both inspectors and the algorithm reported higher evaluation grades of inspectors. The lower efficiency of audits does not apply to desk audits, maybe due to the more limited interactions (column 2).

Columns (4) to (6) display the association between audit selection and the perceived incidence of corruption. We do not find evidence of differential corruption perception across selection methods (Column 4). For full audits, we can reject that the algorithm improved corruption incidence perception by more than 0.2 standard deviations, both for firms reporting an audit and for firms receiving an audit based on the administrative data.

The taxpayer survey results are consistent with the notion that inspectors were less invested in algorithm cases. Given that the questions asked about interactions in "full audits", it is not surprising that the coefficients among firms selected for desk audits are mostly small and statistically insignificant.[33] Yet, they do not lend support to algorithmic selected cases facing a different incidence of corruption.

[32] Firms that were actually audited as per the administrative audit data account for about a half of firms reporting having been audited. This could mean either that firms perceive routine interactions with the tax administration as audits, or that the administrative data is incomplete, maybe because inspectors break protocol and fail to report cases with no infraction.

[33] For the same reason, it is difficult to explain the significant negative coefficient found for corruption among conducted desk audit cases selected by the algorithm (Panel B, Column 5).

6 Mechanisms

The previous sections show that algorithm-selected audits were less likely to be implemented, and, when implemented, detected lower evasion. At the same time, algorithm-selected audits consumed less human resources, took less time, and their results were less often disputed by taxpayers. We now examine the reasons for why algorithm-selected cases display a lower execution rate. We hypothesize that inspectors are strategic in their choice of cases to implement, prioritizing cases that they expect, based on observable characteristics, to yield a high amount of evasion.

First, we show that the risk score is predictive of tax evasion, thus ruling out that inspectors would not want to use it due to its low performance. Second, we show that providing additional information (risk flags and micro data) for algorithm audits does not increase their execution rates, suggesting that inspectors' hesitance to implement algorithm cases is not due to a lack of information. Third, we predict execution based on observable characteristics. We find that only a third of the execution gap between algorithm and inspector-selected audits can be explained by firms' observable characteristics. Instead, we show in the final subsection that inspectors use different criteria when deciding whether to implement an inspector-selected audit versus an algorithm-selected audit, but are highly strategic: for both case types, the execution rate is strongly increasing in the predicted detectable evasion. We conclude that algorithm-selected cases are executed at lower rates because they have a lower expected return on average than inspector-selected cases.

6.1 The Risk Score Predicts Audit Performance

A necessary though not sufficient condition for inspectors to conduct algorithm-selected audits is that the risk score predicts audit outcomes, i.e. the detection of evasion, the evasion amount or evasion rate. To test whether this is the case, Table 8 reruns our main estimation (Table 4, column 1), adding the risk score as an additional regressor. We allow the association between the risk score and audit outcomes to differ between inspector-selected and algorithm-selected cases, as by construction the variation in risk scores is high for inspector-selected cases, but limited for algorithm-selected cases, which all have a high risk score. We find that inspector-selected cases with a higher risk score are more likely to be audited (column 1), and are slightly but not statistically significantly more likely to yield a detection of evasion (columns 3 and 4), and that they exhibit a much higher recovered evasion amount (columns 5 and 6). For both full and desk audits, a one standard deviation increase in the risk score is associated with 38% more tax evasion detected. For algorithm-selected cases, there is no detectable association between the risk score and audit outcomes. This is not surprising given that all cases have high risk scores in this sample and there is thus limited variation.

The results for inspector-selected cases, especially the association with evasion amounts, could be driven by the weighting of risk scores with turnover, which mechanically increases the risk score for

larger firms. However, the results are almost unchanged when we control for baseline turnover (Table F.1) or use the unweighted risk score as the regressor (Table F.2).[34]

6.2 Providing Information does not Increase Audit Performance

Another factor that may determine whether or not inspectors implement an audit is the information they have about the case. Inspectors may have better information about the cases they selected themselves. They may have limited information about the compliance risks among algorithm-selected cases, and hence struggle to start and implement algorithm-selected audits. To test this hypothesis, we conducted a randomized information treatment among desk audit cases. For about a third of cases, inspectors were provided with the risk flags from the algorithm. For another third of cases, inspectors were provided with both the risk flags and with a user-friendly spreadsheet containing the full information available from the firms' tax declarations and third-party data. For the remaining cases, inspectors were not provided with any additional information by the research team.[35]

To test the effect of the information treatment, we add an indicator variable for the treatment to our main regression specification (Equation 1). We also add an interaction between the information treatment and the algorithm selection dummy, to consider the possibility that the additional information is more valuable for algorithm cases, which inspectors are less familiar with. Table F.4 shows that the information treatment did not affect the probability of starting a case, nor subsequent audit outcomes. Columns 1-3 bundle the two treatments into one treatment indicator, while columns 4-6 show the effects of the two sub-treatments "risk indicators only" and "risk indicators and data spreadsheet". Columns 1-3 show that the negative association of algorithm selection with the main audit outcomes remains significant, whereas the interaction with the information treatment has a small and statistically insignificant coefficient. Columns 4, 5, and 6 show that neither of the treatments had any statistically or economically meaningful effect on audit outcomes. This is true for both algorithm and inspector-selected cases.[36] We conclude that a lack of information about algorithm-selected cases does not contribute to explaining low execution rates among these cases.

6.3 Observable Characteristics do not Explain the Execution Gap

We now examine whether the observable characteristics of firms selected by the algorithm help explain why these taxpayers were less likely to be audited. First, we rerun our main analysis of the association between algorithm selection and audit execution rates (Table 4, column 1) after adding

[34] Table F.3 shows that the association between individual risk score components and audit outcomes largely have the expected sign, but given the small number of observations and noise in the data, we cannot detect with statistical confidence which components of the risk score are most predictive of audit outcomes.

[35] Table C.5 shows that taxpayer characteristics were balanced across treatment groups.

[36] The null result also holds in a simplified estimation where we only add the information treatment dummy to Equation 1, without interacting it with the mode of audit selection.

firm characteristics as regressors. We show the results in Table 9, Panel A. In the first column, we reproduce our main result on audit execution as a benchmark: algorithm-selected audits are 18 percentage points less likely to be executed. In column 2, we control for firm size (log turnover), which reduces the execution gap between algorithm and inspector-selected cases, but the gap still remains highly statistically significant and large at 13 percentage points.[37] In columns 3-5, we add the profit rate, firm productivity and other firm characteristics, including the distance to the tax office in travel time (as per google maps, weekday average), and the Euclidian distance to the tax office in meters as additional regressors. Adding these regressors does not substantially reduce the remaining audit execution gap, nor does it increase the R2 of the estimation.[38]

While the results suggest that a simple OLS model cannot rationalize the execution gap based on observable characteristics, it is possible that there are non-linearities and interactions in the association between firm characteristics and audit execution. We hence turn to more flexible prediction models. As average execution rates vary substantially across tax offices and years, we first demean both the execution rate and all potential predictors by the list-level average, where a list is a tax office $\times$ year. This allows us to use data from all tax offices but focus on the role of firm characteristics in predicting within-office variation in execution rates. As predictors of execution, we consider three lags of turnover, productivity (turnover/wage bill), the profit rate, firm age and distance to the tax office. We estimate (train) a model to predict audit execution in the sample of inspector-selected cases using three different approaches: a flexible OLS, a double-lasso and a random forest (Breiman, 2001). When using the OLS or double-lasso approach, we discretize all predictors into deciles to allow for non-linearity.

Concretely, we estimate (train) the following model on inspector-selected cases:

$$\dot{y}_{i\ell} = \beta \dot{X}_i + \epsilon_{i\ell} \tag{2}$$

where $\dot{y}_{i\ell} \equiv y_{i\ell} - \bar{y}_\ell$ and ℓ indicates a tax office by year list.

The results are shown in Panel A of Figure 4, for full audits and in Panel A of Figure F.1 for desk audits. Focusing on the left part of the panel, the black solid line marks the realized execution rate for inspector-selected cases, compared to the tax-office-year mean. The black markers show the average of the prediction using different models, and the box plot illustrates the range of these predictions, when using 100 randomly selected 70% subsamples to re-estimate the model. All estimates within the p5 to p95 range suggest that the execution rate among inspector-selected cases is significantly higher

[37] Similarly, when we use only firms' turnover ranking to predict execution, the simulated choice of case to execute is still far from the realized execution choice, as shown in Table F.6.

[38] Panels B and C in Table 9 show similar results when the outcome variable is the detection of evasion conditional on audit execution (Panel B) or the evasion amount (Panel C). In both cases, the association between algorithm selection and the audit outcome changes little when we add firm characteristics as controls in the estimation.

than the office-year mean, and the p5-p95 range is tightly centered around the realized execution rate. This shows that the models perform well when predicting audit execution out-of-sample among the inspector-selected cases. The mean squared prediction errors printed below each box plot also suggest that, at the individual level, the random forest yields more accurate predictions than the OLS and Lasso, consistent with the fact that it allows for interactions among the predictors.

However, when we use the same models (trained on the sample of inspector-selected cases) to predict execution in the sample of algorithm-selected cases, we cannot match the realization. The right part of Panel A shows that predicted execution rates (grey markers) are far higher than realized execution rates (grey solid line). The predicted execution rates are similar to those for the inspector-selected cases, and around 18 percentage points higher than the realized execution rates for algorithm-selected cases. While turnover is an important predictor of execution, and would suggest lower execution rates for algorithm-selected firms which are smaller on average than inspector-selected firms, other firm characteristics work in the opposite direction.[39]

Hence, even a flexible modeling approach based on observable characteristics cannot help explain the audit execution gap between inspector and algorithm-selected cases. Instead, the evidence suggests that inspectors use different (mental) models when deciding whether to audit inspector-selected cases versus algorithm-selected cases. To see this more directly, consider the importance score attached to different predictors in the random forest model. When running the model on inspector-selected cases, the five most important predictors for execution are lagged log turnover, followed by log distance in meters (93% relative importance), lagged profit rate (93%), firm age (89%) and lagged productivity (86%). In contrast, when estimating the same model on algorithm-selected cases, log distance in meters ranks first among the predictors, followed by lagged log turnover (with only 86% relative importance), lagged productivity (84%), duration of a trip in minutes (84%) and firm age (81%).[40]

6.4 Inspectors' Audit Execution Decisions are Strategic

Finally, we test the hypothesis that inspectors are strategic and maximize detectable evasion when deciding on which audits to implement. To do this, we train a random forest model to predict the amount of detected evasion, in the sample of all cases that were audited. In addition to the predictors discussed in the previous section, we also control for a dummy indicating whether the case was selected by the algorithm. This means we allow for observable characteristics to map into detectable

[39] We have also reversed the exercise and used the algorithm sample to train the model, applying it then to the sample of inspector-selected cases. In this case, the predicted execution rates in the sample of inspector-selected audits are similar to the realized execution rate for algorithm-selected audits and much lower than the realized execution rate for inspector-selected audits.

[40] Table F.5 shows the point estimates on the predictors in a linear prediction model, similarly evidencing the fact that the size and size-ranking of point estimates are different for inspector-selected cases and algorithm-selected cases. Table F.6 shows that an execution strategy that simply relies on ranking firms by turnover also yields a better match with the realized execution pattern.

evasion in different ways for algorithm-selected and inspector-selected cases. Given the continuous nature of the outcome, we follow a two-stage selection procedure, where we first predict whether evasion is detected, and then predict the

We then use the model to predict the hypothetically detectable evasion amount for cases that inspectors chose not to audit. Figure 4, Panel B, plots the results for full audits. For cases that were audited, the distribution of the predicted and realized evasion amounts are very similar, suggesting that the model yields reasonable predictions. The distribution of the expected detectable evasion for non-audited cases, however, is shifted leftwards, indicating that these cases have much lower potential.[41] The figure also shows that the realized execution rate is clearly upward-sloping in predicted evasion amounts.

Hence, regardless of the mental model that inspectors apply to decide on audit execution (which is different for inspector and algorithm cases), inspectors are strategic in their choice of audits to implement, prioritizing cases based on expected detectable evasion. The results also imply that, although the risk score has some power to predict audit outcomes as we showed in Section 6.1, the inspectors are still more skilled than the algorithm in identifying high-evasion cases.

7 Performance of an Alternative Machine-Learning Algorithm

The analysis has shown that our risk-scoring algorithm cannot outperform inspectors' case selection. An open question is whether a better algorithm, trained on audit outcome data to predict detectable evasion, could have improved audit performance.

We hence leverage our random forest prediction of evasion and simulate the revenue potential of audits selected according to this machine-learning algorithm. Concretely, we take the following steps. First, we train our two-stage random forest to predict evasion on the sample of implemented audits, with tenfold increased weights for cases in the top quartile of the evasion distribution. These weights allow us to match aggregate revenue (detected evasion and fines) more accurately than an unweighted prediction.[42] Second, we apply the resulting model to predict detectable evasion for all firms in the annual audit programs, i.e. those that could have potentially faced an audit. Third, we calculate predicted aggregate revenue from several case selection scenarios, holding the total number of audits per tax office constant.

Table 10 presents the results for our preferred simulations. Panel A, column 1 shows the realized aggregate revenue from full audits, across all tax offices in our study, for 2018-2020. Column 2

[41] This pattern holds also when considering only inspector-selected audits or only algorithm-selected audits (Figure F.2). This figure also shows that, conditional on predicted evasion, the execution rates for algorithm and inspector-selected cases are not very different.

[42] Alternative ways of specifying the weights do not achieve as close a match between the realized and predicted aggregate revenue. Table G.2 and Figure G.1 document the performance of our prediction models.

shows the predicted aggregate revenue from the same audit program. This means that we sum up the predicted revenue from all audit cases that were in reality implemented, using the prediction of our random forest model. The prediction is close to but still slightly below the realized aggregate revenue. We hence compare the results of our simulations to the predicted revenue in column 2. Column 3 shows the potential revenue gain from simulation 1, in which we re-optimize across all cases that were selected for the annual audit program, either by inspectors or by the risk-scoring algorithm. Concretely, we rank program cases based on their predicted evasion within each tax office and select the top-ranked N cases, where N is the realized number of audits. Aggregate revenue is predicted to increase by only about 20%. Note that this is likely an overly optimistic estimate, as it would be achieved only if inspectors fully adhered to the proposed case selection. The experience of our intervention, however, suggests that this is unlikely.[43]

Panel B shows that the revenue gains in both simulation scenarios are higher for desk audits than for full audits. This is consistent with the fact that inspectors invest less in desk audit selection and select part of their share of cases at random. For the risk-scoring algorithm, too, desk audit selection might be further away from the optimum, as data quality and hence evasion detection potential decreases with firm size. Table G.1 shows the results separately by tax office. Consistent with the fact that the lowest skilled inspectors work in the small taxpayer offices, the revenue gains from optimization are highest in these offices (panel D).

While the revenue gains from optimizing with a machine-learning algorithm are not negligible, they are far smaller in our context than in most other machine learning applications (see Table A.1). Our limited optimization gains are consistent both with the inspectors' skills and incentives and with the challenges to building a machine-learning algorithm in our context. Inspectors are highly trained and experienced civil servants, and receive strong financial incentives to maximize detected evasion. The algorithm, however, has to be trained on just a few hundred audit cases per year, while the list of potential predictors includes over a hundred variables from the tax returns and third-party data.[44] This is contrary to an ideal machine-learning application, which would rely on a large training dataset with a limited number of predictors (Athey and Imbens, 2019).

In addition, Senegal's newly digitized tax data still faces quality issues, such as missing identifiers or missing firm names, which prevent matching across datasets (or lead to erroneous matches). Some data sources, such as the contract registry as well as business-to-business payment and withholding records were digitized only after our intervention. Czajka (2023) has shown the usefulness of these data for enforcement purposes. It is hence possible that the inclusion of these data could have

[43] We do not conduct the optimization exercise across all firms, as it would require us to predict evasion for the population of all firms, which is very different in characteristics from the audit program sample. In addition, data missingness would be a bigger problem in the population, and we would not be able to use the selection year as a predictor for evasion.

[44] As the list of potential predictors is very large compared to the size of the training dataset, we manually selected predictors from the list for our random forest. Figure G.2 shows the importance ranking of predictors.

strengthened the algorithm.

8 Conclusion

This paper has studied whether an algorithm run on newly digitized micro data can help improve tax audit selection in a developing country context. We find that it is difficult to build an algorithm able to outperform tax inspectors in selecting firms with the highest amounts of evasion. Inspectors prefer to conduct the audits they select themselves, rather than those selected by a risk-scoring algorithm, and inspector-selected audits perform better on most dimensions, especially in terms of evasion uncovered and cost-effectiveness. We rule out a number of pre-specified explanations for inspectors' differential execution decisions, and show that inspectors behave strategically, prioritizing audits of firms with higher amounts of predicted evasion. Ultimately, even an ex-post optimized algorithm trained on outcome data could only have increased aggregate evasion detection moderately, compared to the set of cases the inspectors chose to audit.

Our findings contrast with a growing literature that demonstrates a huge potential for machine-learning algorithms to improve upon or replace the work of humans. Our analysis points to the importance of contextual features in shaping the potential for machine-learning-based enhancement. The corporate tax audits we study are complex and multi-dimensional tasks that require knowledge of law, accounting, data science, and context-specific information. The agents whose performance we attempted to enhance are elite civil servants, highly trained and strongly incentivized to perform. The data which our algorithm is fed with, on the other hand, face quality limitations, as is the case for most administrative datasets in lower-income countries. In addition, even in a relatively sizable country such as Senegal and working at scale, audit outcome data would have to be collected over an extended period to build a sufficiently large training dataset, especially considering the large number of potential predictors of evasion.

Hence, although our findings cannot directly quantify the value of improved data quality, they nonetheless suggest that data infrastructure investments can broaden the applicability of modern machine learning techniques for public policy design in lower-income countries. Our results also highlight the benefits of bureaucrat discretion and autonomy. In other contexts, agent discretion has limited the performance of algorithms, as agents may pursue different objectives than the algorithm (Kim et al., 2024; Stevenson and Doleac, 2024). In our setting however, agent discretion has likely limited the damage caused by an under-performing algorithm.

References

Advani, Arun, William Elming, and Jonathan Shaw (2023) "The Dynamic Effects of Tax Audits," *The Review of Economics and Statistics*, 105 (3), 545–561, 10.1162/rest_a_01101. 6

Anderson, Michael L. (2008) "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 103 (484), 1481–1495, https://ideas.repec.org/a/bes/jnlasa/v103i484y2008p1481-1495.html. 23, 39

Andini, Monica, Michela Boldrini, Emanuele Ciani, Guido de Blasio, Alessio D'Ignazio, and Andrea Paladini (2022) "Machine learning in the service of policy targeting: The case of public credit guarantees," *Journal of Economic Behavior Organization*, 198, 434–475, https://doi.org/10.1016/j.jebo.2022.04.004. 6, 48

Andini, Monica, Emanuele Ciani, Guido de Blasio, Alessio D'Ignazio, and Viola Salvestrini (2018) "Targeting with machine learning: An application to a tax rebate program in Italy," *Journal of Economic Behavior Organization*, 156 (C), 86–102, https://EconPapers.repec.org/RePEc:eee:jeborg:v:156:y:2018:i:c:p:86-102. 6, 48

Assunção, Juliano, Clarissa Gandour, and Romero Rocha (2023) "DETER-ing Deforestation in the Amazon: Environmental Monitoring and Law Enforcement," *American Economic Journal: Applied Economics*, 15 (2), 125–156. 7

Athey, Susan and Guido W. Imbens (2019) "Machine Learning Methods That Economists Should Know About," *Annual Review of Economics*, 11 (Volume 11, 2019), 685–725, https://doi.org/10.1146/annurev-economics-080217-053433. 29

Balán, Pablo, Augustin Bergeron, Gabriel Tourek, and Jonathan L. Weigel (2022) "Local Elites as State Capacity: How City Chiefs Use Local Information to Increase Tax Compliance in the Democratic Republic of the Congo," *American Economic Review*, 112 (3), 762–97, 10.1257/aer.20201159. 7

Bandiera, Oriana, Michael Best, Adnan Qadir Khan, and Andrea Prat (2021) "The Allocation of Authority in Organizations: A Field Experiment with Bureaucrats*," *The Quarterly Journal of Economics*, 136 (4), 2195–2242, https://EconPapers.repec.org/RePEc:oup:qjecon:v:136:y:2021:i:4:p:2195-2242. 7

Banerjee, Abhijit, Esther Duflo, Clement Imbert, Santhosh Mathew, and Rohini Pande (2020) "E-governance, accountability, and leakage in public programs: Experimental evidence from a financial management reform in india," *American Economic Journal: Applied Economics*, 12 (4), 39–72. 7

Battaglini, Marco, Luigi Guiso, Chiara Lacava, Douglas L. Miller, and Eleonora Patacchini (2024) "Refining public policies with machine learning: The case of tax auditing," *Journal of Econometrics*, 105847, https://doi.org/10.1016/j.jeconom.2024.105847. 6, 48

Besley, Timothy, Robin Burgess, Adnan Khan, and Guo Xu (2022) "Bureaucracy and development," *Annual Review of Economics*, 14, 397–424. 7

Best, Michael, Jawad Shah, and Mazhar Waseem (2021) "Detection Without Deterrence: Long-Run Effects of Tax Audit on Firm Behavior," *WP*. 6

Black, Emily, Hadi Elzayn, Alexandra Chouldechova, Jacob Goldin, and Daniel Ho (2022) "Algorithmic fairness and vertical equity: Income fairness with IRS tax audit models," *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1479–1503. 6

Boning, William C, Nathaniel Hendren, Ben Sprung-Keyser, and Ellen Stuart (2023) "A Welfare Analysis of Tax Audits Across the Income Distribution," Working Paper 31376, National Bureau of Economic Research, 10.3386/w31376. 6

Breiman, Leo (2001) "Random Forests," *Machine Learning*, 45 (1), 5–32. 26

Brockmeyer, Anne and Magaly Saénz Somarriba (2025) "Electronic Payment Technology and Tax Compliance: Evidence from Uruguay's Financial Inclusion Reforms," *American Economic Journal: Economic Policy*. 7

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan (2016) "Productivity and Selection of Human Capital with Machine Learning," *American Economic Review*, 106 (5), 124–27, 10.1257/aer.p20161029. 6, 48

Cordova-Novion, Cesar and Tari Sahovic (2010) *Inspections reforms : do models exist*: Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/216181471587685722/Inspections-reforms-do-models-exist. 2

Czajka, Leo (2023) "Fraud Detection Under Limited State Capacity: Experimental Evidence from Senegal,"Technical report, Working Paper. 29

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan (2013) "Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from India," *The Quarterly Journal of Economics*, 128 (4), 1499–1545. 7

——— (2018) "The value of regulatory discretion: Estimates from environmental inspections in India," *Econometrica*, 86 (6), 2123–2160. 7

Dzansi, James, Anders Jensen, David Lagakos, and Henry Telli (2022) "Technology and Tax Capacity: Evidence from Local Governments in Ghana," Working Paper 29923, National Bureau of Economic Research, 10.3386/w29923. 7

Faye, Issa, Abdoulaye Sow, and Sambane Yade (2022) "Evaluation du Potential Fiscal du Secteur Informel au Sénégal,"Technical report, Ministere de l'Economie, du Plan et de la Coopération'. 8

Finan, Frederico, Benjamin A Olken, and Rohini Pande (2017) "The personnel economics of the developing state," in *Handbook of Economic Field Experiments*, 2, 467–514: Elsevier. 7

Glaeser, Edward L, Andrew Hillis, Scott Duke Kominers, and Michael Luca (2016) "Crowdsourcing city government: Using tournaments to improve inspection accuracy," *American Economic Review*, 106 (5), 114–18. 6, 48

Haseeb, Muhammad and Kate Vyborny (2022) "Data, discretion and institutional capacity: Evidence from cash transfers in Pakistan," *Journal of Public Economics*, 206 (C), S0047272721001717, https://EconPapers.repec.org/RePEc:eee:pubeco:v:206:y:2022:i:c:s0047272721001717. 7

Hino, M., E. Benami, and N. Brooks (2018) "Machine Learning for Environmental Monitoring," *NAture Sustainability*, 1, 583–588. 6, 48

Johnson, Matthew S., David I. Levine, and Michael W. Toffel (2023) "Improving Regulatory Effectiveness through Better Targeting: Evidence from OSHA," *American Economic Journal: Applied Economics*, 15 (4), 30–67, 10.1257/app.20200659. 6, 48

Khan, Adnan Q, Asim I Khwaja, and Benjamin A Olken (2015) "Tax farming redux: Experimental evidence on performance pay for tax collectors," *The Quarterly Journal of Economics*, 131 (1), 219–271. 7

Khan, Adnan Q., Asim Ijaz Khwaja, and Benjamin A. Olken (2019) "Making Moves Matter: Experimental Evidence on Incentivizing Bureaucrats through Performance-Based Postings," *American Economic Review*, 109 (1), 237–70, 10.1257/aer.20180277. 7

Khwaja, Muwer Sultan, Rajul Awasthi, and Jan Loeprick (2011) *Risk-based tax audits: approaches and country experiences*: The World Bank. 2, 52

Kim, Hyunjin, Edward L Glaeser, Andrew Hillis, Scott Duke Kominers, and Michael Luca (2024) "Decision authority and the returns to algorithms," *Strategic Management Journal*, 45 (4), 619–648. 30

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2017) "Human Decisions and Machine Predictions*," *The Quarterly Journal of Economics*, 133 (1), 237–293, 10.1093/qje/qjx032. 6

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein (2018) "Discrimination in the Age of Algorithms," *Journal of Legal Analysis*, 10, 113–174. 48

Kleven, Henrik Jacobsen, Martin B Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez (2011) "Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark," *Econometrica*, 79 (3), 651–692. 7

Knebelmann, Justine, Victor Pouliquen, and Bassirou Sarr (2024) "TDiscretion versus Algorithms: Bureaucrats and Tax Equity in Senegal," *Working Paper*. 7

Kotsogiannis, Christos, Luca Salvadori, John Karangwa, and Theonille Mukamana (2024) "Do tax audits have a dynamic impact? Evidence from corporate income tax administrative data," *Journal of Development Economics*, 170, 103292. 6

Lee, David S. (2009) "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 76 (3), 1071–1102, 10.1111/j.1467-937X.2009.00536.x. 17

Margalit, Yotam and Shir Raviv (2024) "When Your Boss is an Algorithm: The Effect of Algorithmic Management on Worker Performance," *Available at SSRN*. 6

Muralidharan, Karthik, Paul Niehaus, and Sandip Sukhtankar (2016) "Building State Capacity: Evidence from Biometric Smartcards in India," *American Economic Review*, 106 (10), 2895–2929, 10.1257/aer.20141346. 7

Naritomi, Joana (2019) "Consumers as tax auditors," *American Economic Review*, 109 (9), 3031–72. 7

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan (2019) "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, 366 (6464), 447–453. 6

OECD (2023) *Tax Administration 2023: Comparative Information on OECD and other Advanced and Emerging Economies*: OECD Publishing, Paris. 2

Okunogbe, Oyebola and Fabrizio Santoro (2022) "The Promise and Limitations of Information Technology for Tax Mobilization," *The World Bank Research Observer*, 38 (2), 295–324, 10.1093/wbro/lkac008. 7

Okunogbe, Oyebola and Gabriel Tourek (2024) "How Can Lower-Income Countries Collect More Taxes? The Role of Technology, Tax Agents, and Politics," *Journal of Economic Perspectives*, 38 (1), 81–106, 10.1257/jep.38.1.81. 7

Pomeranz, Dina (2015) "No Taxation Without Information: Deterrence and Self-Enforcement in the Value Added Tax," *American Economic Review*, 105 (8). 7

Rasul, Imran and Daniel Rogger (2018) "Management of bureaucrats and public service delivery: Evidence from the nigerian civil service," *The Economic Journal*, 128 (608), 413–446. 7

Saavedra, Santiago (2023) "Technology and State Capacity: Experimental Evidence from Illegal Mining in Colombia," *Available at SSRN 3933128*. 7

Sequeira, Sandra (2012) "Chapter 6: Advances in Measuring Corruption in the Field," in *New Advances in Experimental Research on Corruption*, 145–175: Emerald Group Publishing Limited, https://EconPapers.repec.org/RePEc:eme:rexezz:s0193-2306(2012)0000015008. 23

Stevenson, Megan T and Jennifer L Doleac (2024) "Algorithmic risk assessment in the hands of humans," *American Economic Journal: Economic Policy*. 6, 30

Szucs, Ferenc (2023) "Discretion and favoritism in public procurement," *Journal of the European Economic Association*, jvad017. 7

World Bank (2024) "Public Finance Review Tool," knowledge tool. 8

World Bank, The (2005) *Good practices for business inspections : guidelines for reformers*: Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/286811468329950178/Good-practices-for-business-inspections-guidelines-for-reformers. 2

Tables

Table 1: Number of Firms by Data Source (Digitized Data)

		2014	2015	2016	2017	2018	2019	2020
A Self reported	CIT	5136	5969	6218	6594	6720	7233	NA
	VAT	11181	11901	12699	13352	13969	14213	13538
	PAYE	7061	7518	7870	8513	8782	9005	8621
	CGU	1581	1827	2026	2203	2650	2671	2801
	TAF	86	105	112	122	121	111	118
B Third party	Imports	8963	12427	13068	11859	13551	13677	10591
	Exports	1398	1724	1881	1824	1697	1659	1558
	Procurement	809	735	1380	1340	1903	1897	1684
	VAT annexes	6	9	21	805	3606	3209	NA
C Audits data	Digitized	NA	NA	1	3294	2753	2946	3714
	Self-reported (Excel)	NA	NA	NA	102	561	664	51

Notes: This table shows the number of unique taxpayers (firms) by year of available data in digital format, in the main datasets used to construct the risk scoring algorithm and analyze its performance. The available data covers the years 2014 to 2020, and the experiment was conducted in the years 2018, 2019, and 2020. The self-reported data (Panel A) include the self-assessment declarations for the Corporate Income Tax (CIT), Value-Added Tax (VAT), Pay-As-You-Earn withheld personal income tax (PAYE), the simplified Senegalese tax for small enterprises called CGU, and the Senegalese tax for financial services enterprises called TAF. The third-party datasets (Panel B) include transactions data from customs (imports and exports), procurement records, and VAT annexes. The audit data (Panel C) come from a) an effort to digitize audit reports from 2017 onwards, covering all audits conducted under our intervention, and b) an excel spreadsheet which the research team requested tax inspectors to fill out to report audit outcomes and audit process information. This table is discussed in Section 2.3.

Table 2: Number of Audit Cases by Selection Method and Tax Office (2018-2020)

	Full Audits			Desk Audits			
Tax Office	Inspectors	Algorithm	Total	Inspectors	Algorithm	Random	Total
Large Taxpayer Unit	265	131	396	320	157	60	477
Medium Taxpayer Unit	155	153	308	320	290	135	610
Liberal Professions	46	50	96	235	222	85	457
SME(Regional)	71	73	144	422	401	84	823
Total	537	407	944	1297	1070	364	2367

Notes: Number of selected cases by tax office, audit type, and selection method. The sum of the rows is larger than the total because there are overlapping cases between algorithm and discretion. Random cases and replacement cases are exclusive to desk audits. This table is discussed in Section 2.3.

Table 3: Firm Characteristics of Algorithm vs Discretionary Selection

Panel A: Characteristics of Inspector-Selected Cases Relative to the Full Population of Firms

Data Source:	Tax Declarations				Admin. Audit Data		Taxpayer Survey		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	log(Turnover)	log(Payroll)	Profit Rate	P(Trade)	Duration Trip	Firm Age	Employees	% Sales in Cash	Audit Frequency
Inspectors' Selection	3.30***	1.35***	0.01	0.17***	-0.00	0.97*	-5.22	-6.46	-0.02
	(0.87)	(0.28)	(0.01)	(0.02)	(0.05)	(0.44)	(6.40)	(6.27)	(0.20)
N	22576	7433	5925	61238	60608	51992	696	702	640
R2	0.23	0.13	0.05	0.02	0.06	0.10	0.09	0.03	0.03
Mean outcome	10.82	14.81	-0.15	0.41	1.95	5.91	31.24	50.32	2.77

Panel B: Comparison of Inspector versus Algorithm Selected Cases

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Algorithm	-1.78***	-1.67***	-0.07**	0.05*	0.13	0.17***	15.97	12.31**	-0.05
	(0.42)	(0.41)	(0.03)	(0.03)	(0.09)	(0.06)	(18.28)	(5.00)	(0.25)
Inspectors x Overlap	0.04	1.02**	0.00	0.06	0.35	0.09	-98.48	9.95	0.37
	(1.07)	(0.48)	(0.04)	(0.08)	(0.23)	(0.19)	(60.08)	(22.06)	(0.77)
Algorithm x Random									
N	1186	1021	906	1323	1300	1323	268	277	268
R2	0.12	0.12	0.09	0.14	0.06	0.06	0.28	0.13	0.11
Mean outcome	18.39	16.51	-0.02	0.51	2.05	2.13	44.72	45.70	2.79

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows the OLS regression coefficients of how inspector selected firms characteristics of the firms depending on the selection methods for full audits (Table C.6 shows the same exercise for desk audits). Panel A compares the characteristics of firms selected for full audits by inspectors, to the general population of firms, pulling the 2018-2020 audit selection years. Panel B compares the characteristics of firms selected by the algorithm relative to those that were selected by the inspectors (all selected full audit cases 2018-2020). In panel A, columns 1-6, the regressions show the difference in the firms' characteristics using a dummy variable that indicates that the firm was selected for full audits by inspectors, at some point during the years 2018-2020. The sample is the entire set of firms with tax declarations in Senegal. The regressions control for fixed effects at the tax office level. Robust standard errors are shown in parentheses (Huber-White formula). Data on the characteristics of the firms stem from three sources. From the tax declarations, we use the log of the yearly declared turnover, the log of the yearly declared payroll, the profit rate, and the probability that the firm has exports or imports. We use the value for the year before the firm was selected for audit. We use data from the firm registry on the firm's age and the distance between its location and DGID, in minutes of travel time (computed using GoogleMaps for a Monday at 3 PM). Finally, we use the taxpayer survey to compute the (self-reported) number of full-time employees, the share of total sales done in cash, and the perceived yearly frequency of full audits. This table is discussed in Section 3.5.

Table 4: Algorithm Selection and Audit Outcomes

	P(Execution)			P(Detection | Execution)			log(Evasion) | Detection		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	-0.18***	-0.05**	-0.04**	0.04	-0.05*	-0.04	-0.64***	-0.25*	-0.16
	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.18)	(0.13)	(0.14)
				[-.03*,.06**]		[-.07**,-.03]	[-.78***,-.46***]		[-.35***,-.06]
Inspectors x Overlap	0.16**	0.05	0.04	0.08*	-0.07	-0.08	0.29	-0.58*	-0.40
	(0.07)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.46)	(0.31)	(0.37)
Algorithm x Random		-0.00	-0.00		0.00	0.01		-0.22	-0.21
		(0.03)	(0.03)		(0.04)	(0.04)		(0.18)	(0.19)
Tax Office x Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector x Year	No	No	Yes	No	No	Yes	No	No	Yes
N	944	2731	2731	507	1016	997	453	751	732
R2	0.26	0.23	0.32	0.15	0.32	0.43	0.31	0.29	0.43
Mean outcome	0.53	0.37	0.37	0.89	0.73	0.73	19.29	17.74	17.71
N (Lee bounds)				455		895	417		671

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of the main audit outcomes on the selection method of the case, for full and desk audits separately. Specifications for full audits control for tax office x year fixed effects, which corresponds to the selection list for full audits. For desk audits, we show specifications with tax office x year fixed effects, and inspector x year fixed effects, the latter corresponding to the list for desk audits. The first outcome (Columns 1 to 3) is the probability that the case on the list was actually executed by the inspectors. The second outcome (Columns 4 to 6) is the probability that the audit resulted in a penalty (initial or final notice), conditional on the case being conducted. The third outcome (Columns 7 to 9) is the log amount of the evasion plus penalties uncovered, conditional on the case ending in positive penalty. The evasion amount is obtained from the final notice, which we complement with the initial notice when the final notice is not available. The sample includes all cases selected in the 2018, 2019, and 2020 audit programs. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Robust standard errors (Huber-White) are shown in parentheses, and inspector x year levels when inspector fixed effects are included. Lee bounds are shown in brackets for the coefficient on the algorithm-selection dummy, computed based on the attrition between algorithm and inspector cases for variables after audit execution. Table D.1 shows intent-to-treat effects, Table D.2 shows robustness to controlling for audit slot fixed effects, and Table D.3 shows the results when using the evasion rate (instead of the evasion amount) as the outcome. This table is discussed in Section 4.1.

Table 5: Resources Allocated to Audit Execution

Panel A: Resource Outcomes

	Number of Agents	Duration in Days (Firm Survey)	Days From Start to Conf. (Admin Data)		Days Working on Case (Self-reported)	
	(1) Full audits	(2) Full audits	(3) Full audits	(4) Desk audits	(5) Full audits	(6) Desk audits
Algorithm	-0.22***	-8.66**	-25.29	6.26	-7.23	-3.50
	(0.08)	(3.72)	(15.92)	(22.12)	(30.22)	(2.44)
	[-.36***,-08]	[-9.48**,-7.61*]	[-26.05*,-12.82]	[5.04,26.07]	[-17.86,15.92]	[-3.17,-1.86]
Inspectors x Overlap	0.02	-26.52	-20.15	25.96		-22.76*
	(0.16)	(22.04)	(38.38)	(42.54)		(13.51)
Algorithm x Random				2.13		-2.04
				(21.28)		(1.74)
N	507	214	285	241	51	108
R2	0.23	0.25	0.24	0.37	0.26	0.57
Mean outcome	2.87	28.16	166.11	123.04	188.03	9.06
N (Lee bounds)	449	91	258	195	45	90

Panel B: Data Availability

	(1)	(2)	(3)	(4)	(5)	(6)
Algorithm		-0.01	0.00	-0.00	-0.04	-0.03
		(0.04)	(0.00)	(0.00)	(0.02)	(0.02)
Inspectors x Overlap		-0.03	0.00	-0.00	-0.01	0.01
		(0.06)	(0.00)	(0.00)	(0.04)	(0.04)
Algorithm x Random				0.00		0.05*
				(0.00)		(0.03)
N		507	507	997	507	997
R2		0.21	0.99	0.99	0.49	0.44
Mean outcome	1.00	0.21	0.56	0.26	0.10	0.11

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of resources used in an audit on the selection method of the case. The measures of resources are the number of agents working on the case (for full audits only, column 1) and the audit duration measured in three different ways. The sample includes all cases selected and conducted from the 2018, 2019, and 2020 audit programs, for which we had data on the outcome variables. Panel A shows the results for the outcomes, and Panel B shows the results of regressions for dummies of data availability. Column 1 shows the results for the number of agents involved in the audit, which is obtained from the audit records and is only relevant for full audits because desk audits are typically conducted individually. Column 2 shows the results for audit duration as reported by taxpayers in the taxpayer survey. The survey asked them about their last full audit experience. Columns 3 and 4 use the difference between the date of initial notice and the date of the start of the audit (either verification announcement or information request or start date of audit) as a proxy for the duration of the audit. This measure is available only for cases that had an initial notice. We drop negative values, winsorize the data at the 99th percentile, and control for dummies indicating availability of date of initial notice and final notice. These controls are helpful because the definition of "start date" may vary depending on the availability of information. Columns 5 and 6 measure the duration of the audit using the self-reported number of days worked on a case, reported on a spreadsheet that inspectors needed to fill out during the 2019 audit program. We multiplied the answer by the number of inspectors that worked in the case. Robust standard errors (Huber-White formula) are shown in parentheses, and inspector x year levels when inspector-fixed effects are included. Lee bounds are shown in brackets. This table is discussed in Section 5.1.

Table 6: Algorithm Selection and Audit Productivity

	Evasion/Number of Agents		Evasion/Duration in Days (Survey)		Evasion/Days Start to Conf. (Admin.)		Evasion/Days Working (Self reported)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	-0.54***	-0.25*	-0.13	-0.38	-0.55**	-0.52	-0.40	0.36
	(0.18)	(0.14)	(0.40)	(0.48)	(0.27)	(0.37)	(0.60)	(0.50)
Inspectors x Overlap	0.31	-0.36	1.00	0.53	0.17	-2.68***		0.48
	(0.44)	(0.36)	(2.21)	(0.81)	(0.58)	(0.60)		(1.32)
Algorithm x Random		-0.11		-0.45		-0.10		0.18
		(0.19)		(0.64)		(0.34)		(0.50)
N	453	732	95	93	251	154	39	60
R2	0.31	0.44	0.21	0.44	0.22	0.59	0.26	0.41
Mean outcome	18.22	17.47	15.95	14.82	15.04	13.58	13.61	15.91

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of four measures of audit productivity on the selection method of the case for full and desk audits separately. We show specifications for full audits controlling for tax office x year fixed effects, which corresponds to the selection list for full audits. For desk audits, we show specifications with inspector x year fixed effects. The first outcome (Columns 1 and 2) measures productivity as the log of evasion (plus fines) divided by the number of agents involved in the case; the second outcome (Columns 3 and 4) uses log of evasion divided by the number of days the last full audit lasted as declared by the taxpayer in the taxpayer survey (notice that the result for desk audits may not be relevant here); the third outcome (Columns 5 and 6) uses log of evasion divided by the number of days the audit lasted as defined by the date of start and date of notification; and the fourth outcome (Columns 7 and 8) uses log of evasion divided by the number of days the audit lasted as self-reported by the inspector (only available for 2019). The sample includes all cases selected in the 2018, 2019, and 2020 audit programs. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Robust standard errors (Huber-White) are shown in parentheses, and inspector x year levels when inspector-fixed effects are included. This table is discussed in Section 5.2.

Table 7: Algorithm Selection, Perceived Audit Efficiency and Corruption

Panel A: Surveyed Firms (Self-Reporting a Recent Audit)

Outcome:	Efficiency Index			Corruption Index		
	Full Audits	Desk Audits	All Audits	Full Audits	Desk Audits	All Audits
	(1)	(2)	(3)	(4)	(5)	(6)
Algorithm	-0.42***	0.01	-0.17	0.03	-0.06	-0.03
	(0.16)	(0.14)	(0.11)	(0.15)	(0.14)	(0.10)
Inspectors x Overlap	0.47***	0.03	0.17	-0.56*	-0.12	-0.17
	(0.18)	(0.26)	(0.21)	(0.29)	(0.27)	(0.20)
Algorithm x Random		-0.08	-0.04		-0.08	-0.04
		(0.19)	(0.18)		(0.18)	(0.16)
N	197	272	469	198	271	469
R2	0.10	0.02	0.03	0.09	0.10	0.05
Mean outcome	0.00	0.00	0.00	-0.06	0.04	0.00

Panel B: Only Audited Firms (as per Administrative Audit Data)

	(1)	(2)	(3)	(4)	(5)	(6)
Algorithm	-0.37	0.35	-0.03	0.25	-0.26	-0.02
	(0.27)	(0.23)	(0.17)	(0.25)	(0.22)	(0.17)
Inspectors x Overlap	0.25	0.42	-0.03	-0.30	-0.09	0.11
	(0.24)	(0.34)	(0.29)	(0.48)	(0.33)	(0.29)
Algorithm x Random		0.01	0.09		-0.03	-0.06
		(0.27)	(0.24)		(0.26)	(0.23)
N	86	120	209	86	122	211
R2	0.10	0.09	0.03	0.12	0.12	0.04
Mean outcome	0.00	-0.06	-0.03	-0.12	-0.04	-0.08

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results with fixed effects at the year and tax office level for outcomes from the taxpayer survey. Given the lower number of observations in the survey, we do not run the desk audit regression with year $*$ inspector fixed effects, as is done with outcomes from the administrative data. Robust standard errors are in parenthesis. The table shows the association between audit selection method and taxpayers' perception of the efficiency of audits (columns 1-3), and of the incidence of corruption during audits (columns 4-6). The sample of surveyed taxpayers is a subsample of the audit cases selected by the yearly audit programs between 2018 and 2020. Panel A shows the results for all firms interviewed in the survey, conditional on them self-reporting an audit in recent years. Panel B is restricted to firms for which we can confirm that an audit part of the yearly audit program took place, according to the administrative audit reports. The two outcomes are (1) an index of perceived efficiency of audits, and (2) an index of perceived corruption of audits. Each measure is constructed by combining three questions into a normalized index following Anderson (2008). To measure efficiency, we combine the questions on taxpayers' assessment of the auditor's (i) technical knowledge, (ii) efficiency during the audit, (iii) capacity to uncover all evaded taxes. For corruption, we combine the questions on the perception of (i) the auditor's dishonesty during the audit, (ii) the frequency of bribes paid to inspectors (among firms similar to the respondent's), (iii) the existence of a preferential tax audit treatment for firms connected with the tax administration. The coefficients are measured in terms of standard deviation of the respective indexes. Evidence that respondents' likelihood of answering the questions is not associated with the audit selection method is presented in Table E.5. This table and the survey questions the indices are based on are discussed in Section 5.3.

Table 8: Association of Risk Score with Audit Outcomes

	P(Execution)		P(Detection|Execution)		log(Evasion)|Detection	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	-0.30***	-0.07***	-0.00	-0.08*	-1.08***	-0.32
	(0.05)	(0.03)	(0.04)	(0.04)	(0.29)	(0.21)
Risk score	0.05**	-0.00	0.00	0.00	0.38***	0.38***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.13)	(0.13)
Alg. x Risk score	-0.01	0.02	0.01	0.01	-0.32**	-0.41***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.15)	(0.13)
N	944	2731	507	997	453	732
R2	0.27	0.32	0.15	0.43	0.32	0.44
Mean outcome	0.53	0.37	0.89	0.73	19.29	17.71

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows the correlation of the algorithm risk score with the three main audit outcomes: the probability of starting the selected audit, the probability of detecting evasion among started cases, and the log of detected evasion amount plus fines. The risk score is a continuous variable computed at within "clusters" of firms of similar economic activity and size. We standardize the risk score values at the cluster-year level, i.e., we subtract the mean and divide by the cluster's standard deviation. We do that because the risk score scale changed in the three iterations of the experiment. The coefficients show how the risk score correlates with the outcomes for discretionary cases (non-interacted term) and for algorithm cases (interacted term). The results stem from OLS regressions with fixed effects at the list level (year X tax office for full audits, and year x inspector for desk audits). Robust standard errors (Huber-White) are shown in parentheses. This table is discussed in Section 6.1. Tables F.1 to F.3 show robustness tests for these results.

Table 9: Explaining the Execution Gap for Full Audits: The Role of Observable Characteristics

Panel A: Outcome P(Execution)					
	(1)	(2)	(3)	(4)	(5)
Algorithm	-0.18***	-0.13***	-0.13***	-0.13***	-0.12***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
N	944	944	944	944	944
R2	0.26	0.34	0.34	0.34	0.34
Mean outcome	0.53	0.53	0.53	0.53	0.53

Panel B: OutcomeP(Detection|Execution)					
	(1)	(2)	(3)	(4)	(5)
Algorithm	0.04	0.04	0.04	0.04	0.04
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
N	507	507	507	507	507
R2	0.15	0.16	0.16	0.16	0.17
Mean outcome	0.89	0.89	0.89	0.89	0.89

Panel C: Outcome log(evasion)|Detection					
	(1)	(2)	(3)	(4)	(5)
Algorithm	-0.64***	-0.49***	-0.50***	-0.48***	-0.48***
	(0.18)	(0.19)	(0.18)	(0.18)	(0.19)
Turnover	No	Yes	Yes	Yes	Yes
Profit rate	No	No	Yes	Yes	Yes
Productivity	No	No	No	Yes	Yes
Firm char.	No	No	No	No	Yes
N	453	453	453	453	453
R2	0.31	0.37	0.39	0.40	0.41
Mean outcome	19.29	19.29	19.29	19.29	19.29

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of the main audit outcomes on the selection method of the case for full audits. We show specifications for full audits controlling for tax office x year fixed effects, which corresponds to the selection list for full audits. Each column progressively adds control variables at the firm level: Column 1 has no controls, Column 2 controls for turnover (log), Column 3 controls in addition for the profit rate (profit over turnover), Column 4 controls in addition for the firm productivity (turnover divided by the wage bill), and Column 5 controls in addition for further firm characteristics. The first outcome (Panel A) is the probability that the case on the list was conducted by the inspectors. The second outcome (Panel B) is the probability that an audit results in a penalty (initial or final notice), conditional on the case being conducted. The third outcome (Panel C) is the log amount of the evasion plus penalties, conditional on the case ending in positive penalty. The evasion amount is obtained from the final notice and complemented with the initial notice when the final notice is not available. The sample includes all full audt cases selected in the 2018, 2019, and 2020 audit programs. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Robust standard errors (Huber-White) are shown in parentheses., and inspector x year levels when inspector fixed effects are included. This table is discussed in Section 6.3.

Table 10: Optimization Gains from Machine-Learning Algorithm Trained on Audit Outcome Data

Panel A: Full Audits			
(1)	(2)	(3)	(4)
Realized Revenue Log(mean)	Predicted Revenue Log(mean)	Δ Revenue vs Predicted w/ RF Selection Among Program Cases	Δ Revenue vs Predicted w/ Size-Ranking Among Program Cases
21.1	21.03	21.69	7.25
Panel B: Desk Audits			
(1)	(2)	(3)	(4)
Realized Revenue Log(mean)	Predicted Revenue Log(mean)	Δ Revenue vs Predicted w/ RF Selection Among Program Cases	Δ Revenue vs Predicted w/ Size-Ranking Among Program Cases
19.15	18.77	43.12	24.03

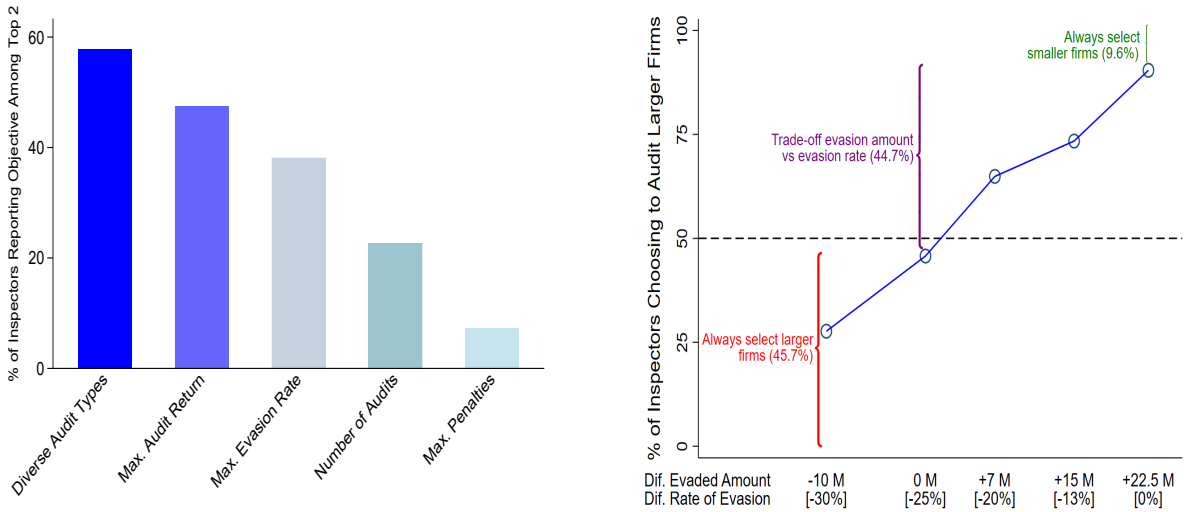
Notes: This table shows the results from our optimization exercise in which we use the audit results from the realized set of audits during 2018-2020 to train a machine-learning (random forest) algorithm to predict evasion. Column 1 shows the realized revenue (evasion and fines) from audits across all tax offices. Column 2 shows the predicted revenue, for firms that experienced an audit. Column 3 shows the revenue increase (in percent, compared to the predicted revenue in column 2) when using the random forest for audit selection. Concretely, we use our estimated random forest to predict evasion for all cases in the annual audit program list, then rank cases by predicted evasion, and pick the top cases such that the number of audits by tax office and year is the same as the realized audit implementation. Column 4 shows the revenue gain (compared to predicted revenue in column 2) from ranking cases by turnover. [The numbers in column 4 are not correct and are currently being reviewed.] This table is discussed in Section 7. Table G.1 shows the same results disaggregated by tax office.

Figures

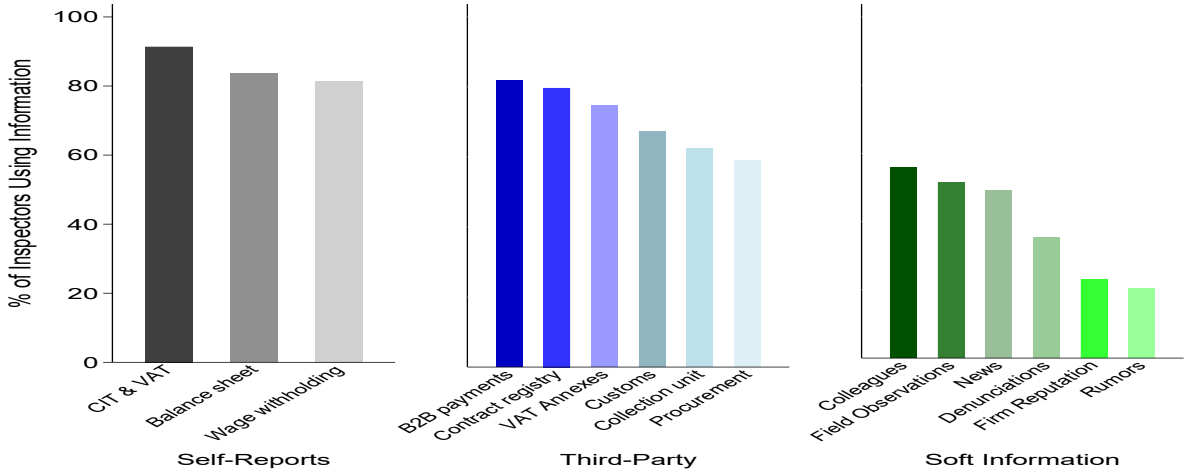
Figure 1: Tax Inspector Objectives and Practices in Audit Selection

(a) Objectives in Audit Selection

(b) Trade-Off: Evasion Amount vs Rate

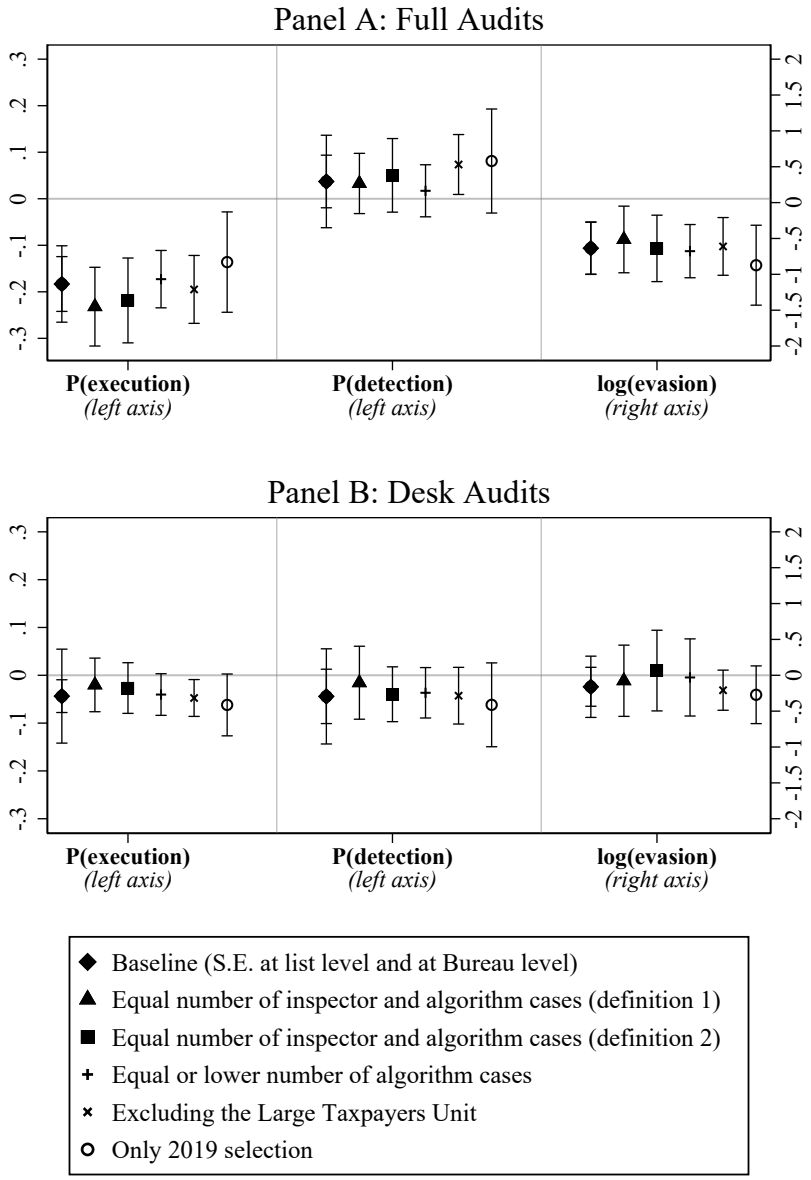


(c) Data Used in Audit Selection



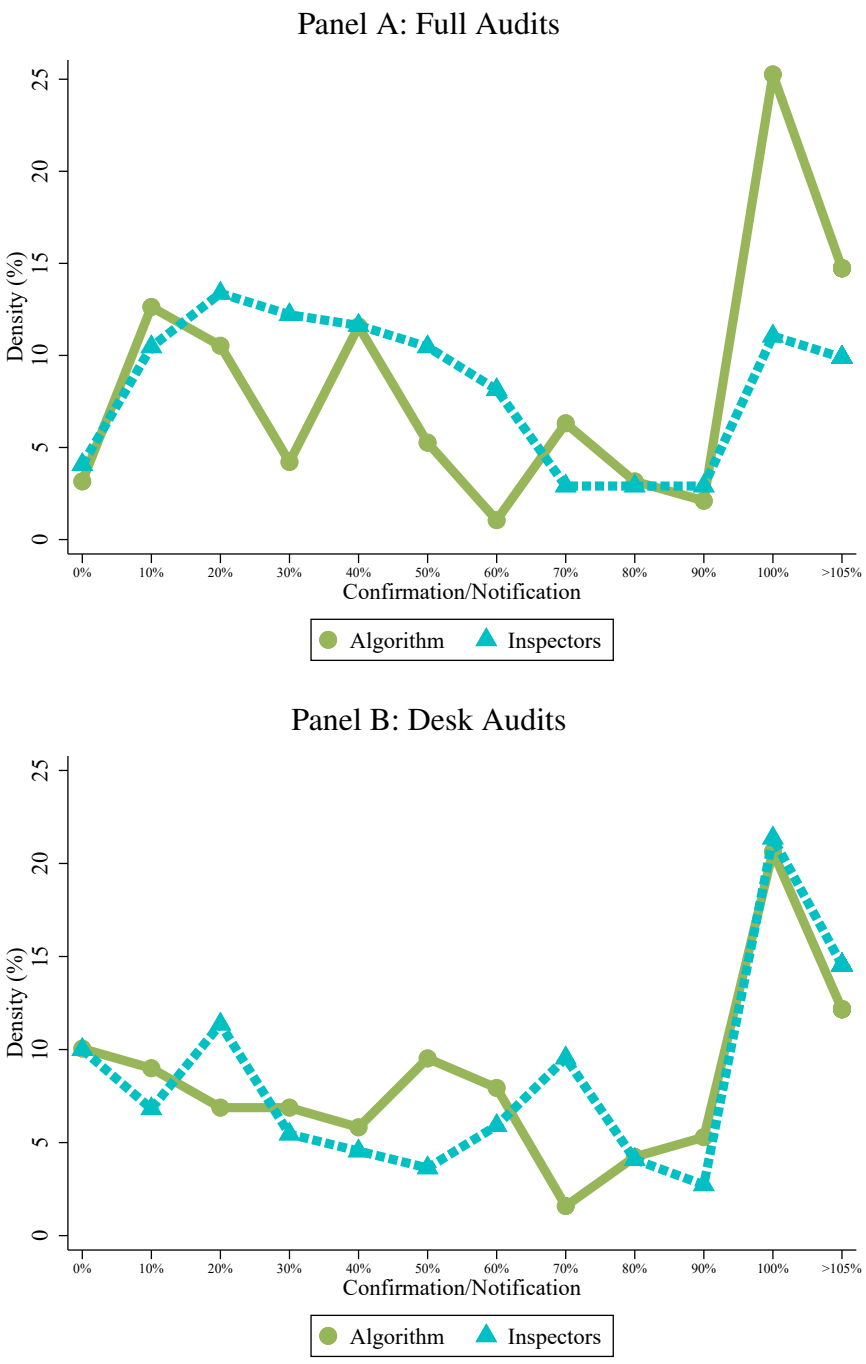
Notes: This figure shows inspectors' objectives, trade-offs and the information they use for audit selection. The data is collected at baseline (December 2017-January 2018) from our survey of 97 tax inspectors actively involved in conducting audits. Panel (a) shows the top two reported personal objectives of audits, from the following choice set: diversity of audit types, maximize audit returns, maximize the evasion rate, conduct the largest number of audits, and maximize the assessed penalties. Results are similar when reporting the top objective. Panel (b) shows the trade-off inspectors make between aiming for larger audit returns, in absolute amounts, versus higher detected evasion rates. This is based on a set of scenarios in which inspectors were provided with the choice of auditing one of two firms, one large and one small, with different evasion amounts and hence different evasion rates. The figure plots the share of inspectors who prefer to audit the larger of the two firms (y-axis), as a function of the degree of tax evasion differential in amounts between the firms (x-axis). Each amount of tax evasion differential translates to a specific difference in evasion rates. Each dot corresponds to a different scenario. In the first scenario (bottom left dot), auditing the large firm is worse both for the expected evasion amount recovered (-10M FCFA) and for the evasion rate differential (-30%); by the last question (top right dot) the expected evasion amount is much larger when auditing the large firm (+22.5 M FCFA) and the evasion rate is equal across the large and small firm. Panel (c) shows the frequency of using different data types for audit selection, classifying data into three categories: firms' self-reports, third-party reported data, and soft information which is hard to codify. This figure is discussed in Section 2.4.

Figure 2: Algorithm Selection and Audit Outcomes: Robustness of Estimates Across Subsamples



Notes: This figure shows the robustness of our results on the probability of execution, detection and evasion amounts uncovered by the algorithm selection relative to the inspector selection, in different subsamples. Panel A shows results for Full audits, and panel B shows results for desk audits. In each sub-panel, the first coefficient reproduces our main result from Table 4, which we call the baseline. We show standard errors when clustering at the list level, as in our main analysis, and at the bureau level. The remaining coefficients using the same empirical specifications as Table 4, employing inspector fixed-effects for the desk audit estimations, and limit the sample in the following way: 2) lists that feature an equal number of inspector-selected and algorithm-selected cases (as observed by the researchers), 3) lists that feature an equal number of inspector-selected and algorithm-selected cases (as observed by the inspectors), 4) office-year observations where the number of algorithm-selected cases is equal to or lower than the number of inspector-selected cases, 5) small and medium taxpayer offices only (i.e. excluding the large taxpayer office), 6) audit lists only for the year 2019 (which we consider as the cleanest execution year for the program). For the distinction between methods 1 and 2 for equalizing the numbers of algorithm and inspector-selected cases, see footnote 26. Tables D.5 to D.10 show the details of all these estimations with numbers of observations in table format. This figure is discussed in Section 4.4.

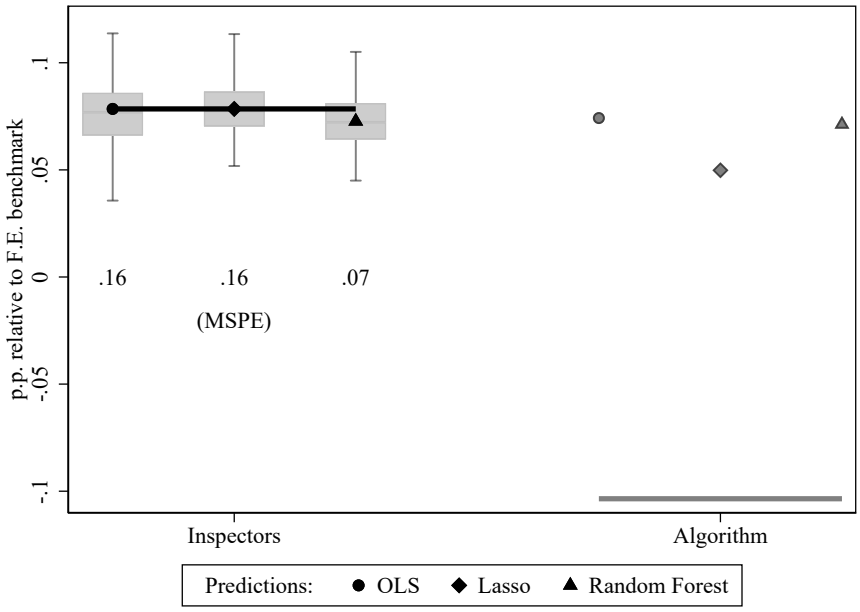
Figure 3: Dispute of Audit Results
Distribution of Confirmed Amount/Notified Amount of Evasion



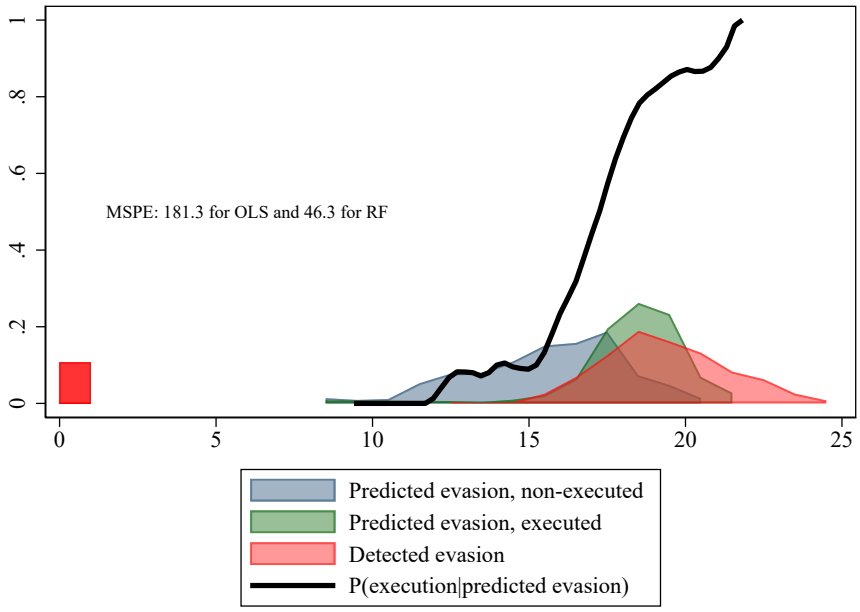
Notes: This figure shows the distribution of the confirmed amount over the notified amount of evasion, for each selection method and audit type. The share is calculated for audit cases were both a non-zero notified amount and a non-missing confirmed amount are reported. This reduces the sample of cases we can consider, for full audits from 507 to 264, and for desk audits from 1,016 to 372. The share of confirmation to notification measures potentially both the extent to which taxpayers bargain to lower the detected evasion and fines due and the quality of the initial assessment on taxes due by the inspectors. This figure is discussed in Section 4.3.

Figure 4: Predicting Inspector Choices and Audit Outcomes Using Observable Characteristics

Panel A: Predicted and Realized Execution of Casess



Panel B: Predicted and Realized Detected Evasion Amount



Notes: Panel A of this figure shows the predicted mean execution of audits based on observable characteristics of the firms. Several models are estimated using only inspector-selected firms, and the model is used to predict the execution rate of algorithm cases based on the firms' characteristics. The difference between the realized execution rate and the predicted execution rate represents the part of the algorithm's effect that cannot be explained by firm characteristics. This figure is discussed in Section 6.3. Panel B plots the distribution of detected evasion for all full audits (log of FCFA amounts), as well as the results from a random forest model predicting detected evasion based on firms' observable characteristics. The black solid line plots the distribution of the realized audit execution rate by level of predicted evasion. This figure is for full audits only. Figure F.1 shows that same analysis for desk audits. This figure is discussed in Section 6.4.

Online Appendix: Not for Publication

This appendix contains additional information and analyses. Appendix A provides an overview of the relevant machine learning literature. Appendix B provides further information on the context of our study. Appendix C presents the design of the risk-scoring algorithm. Appendix D presents additional results on audit outcomes. Appendix E presents additional results on the audit process. Appendix F presents additional results on mechanisms. Appendix G presents additional results on the machine-learning optimization exercise.

A Related Literature

Table A.1: Machine Learning Applications to Enhance the Work of Government Officials

Authors, Year	Paper	Journal	Policy, Country	Tool	Result
Andini et al. (2018)	Targeting with Machine Learning: Tax Rebates	JEBO	Tax rebates to boost individual consumption, Italy	Decision Tree, k-NN, Random Forest	ML-based targeting of the tax rebate to consumption-constrainted households, compared to means-tested targeting, could have increased food consumption by 41.8% or, alternatively, 29.5% of program funds could have been saved, holding food consumption constant.
Andini et al. (2022)	Machine Learning for Public Credit Guarantees	JEBO	Credit guarantees for firms, Italy	Decision Tree, Random Forest, Lasso	ML-based targeting of guarantees to firms that are both credit-constrained and credit-worthy could increase firms' likelihood of obtaining a bank loan by approximately 100%, compared to a simple rule-based targeting.
Battaglini et al. (2024)	Refining Public Policies with Machine Learning: The Case of Tax Auditing	JoEctrics	Tax audit selection, Italy	Random Forest	The ML algorithms can rank audits based on expected tax evasion and recovered evasion. Replacing the 10% worst-performing audits with audit cases selected by the algorithm can increase detected evasion by 38% and recovered evasion by 29%.
Chalfin et al. (2016)	Productivity and Selection of Human Capital	AER P&P	Police hiring, US	Stochastic Gradient Boosting	ML-based police hiring decisions could reduce officer miscondoct (involvement in shootings or verbal abuse) by 4.8%.
Chalfin et al. (2016)	Productivity and Selection of Human Capital	AER P&P	Teacher tenure decisions, US	Regressions with lasso penalty for model complexity	ML-based tenure decisions for teachers, compared to relying on principal ratings, could increase student test score gains by 75% for maths and 105% for English. Using ML for teacher tenure decisions is 2-3 times as effective as reducing elementary school classroom size by one third.
Glaeser et al. (2016)	Crowdsourcing City Government	AER P&P	Restaurant inspections, US	Random Forest, Grad Boosted Trees	The study uses a prediction tournament to explore how tournaments can improve upon traditional consultancy. Implementing the top-performing algorithm from the tournament could increase restaurant inspection productivity by 30-50%.
Hino et al. (2018)	Machine Learning for Environmental Monitoring	Nature	Water pollution inspections, US	Regression Forest	ML-based water quality inspections can detect 600% more violations of pollution limits compared to discretionary targeting. When accounting for local budgetary constraints and maintaining a minimum probability of inspection for all facilities, the algorithm could still increase the number of violations detected by 100%.
Johnson et al. (2023)	Improving Regulatory Effectiveness through Better Targeting	AEJ App	Workplace safety inspections, US	Super Learner	ML-based targeting of inspections could avert 120% more injuries than the current random assignment. This approach could generate up to 876 million in social value by optimizing inspection allocations.
Kleinberg et al. (2018)	Human Decisions and Machine Predictions	QJE	Bail decision, US	Gradient Boosted Trees	ML-based bail decisions can reduce crime by up to 24.7% without increasing jail rates, or alternatively, decrease the number of people jailed by up to 41.9% without raising crime rates.

Notes: This table summarizes the literature on how machine learning algorithms can enhance or substitute for the work of government officials. This table is mentioned in Section 1.1.

Table B.1: Senegal's Tax Revenue Composition in Comparison

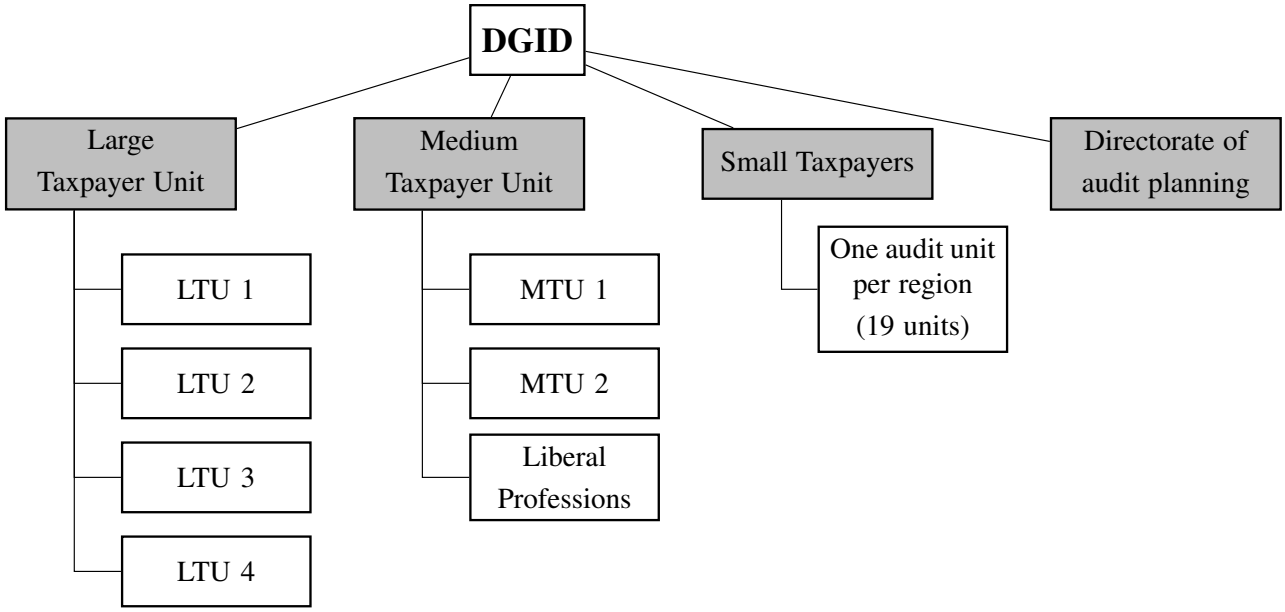
	(1) Senegal	(2) Sub-Saharan Africa	(3) Lower-Middle Income Countries	(4) High Income Countries
Total tax revenues (% of GDP):	19.63	15.32	18.32	32.31
Sources of revenue (% of total):				
VAT	29.36	26.59	28.49	19.98
Excise taxes	9.27	10.12	10.76	5.34
Personal income tax	15.80	18.76	14.75	21.80
Corporate income tax	13.28	19.09	21.57	12.81
Taxes on international trade	14.78	15.73	12.07	0.64
Other taxes	12.80	8.23	7.13	13.49
Social security contributions	4.70	1.48	5.21	25.94

Notes: This table shows total tax revenues, including social contributions, as a percentage of GDP, along with the breakdown of total revenue by category (as a percentage of total revenue) for Senegal, Sub-Saharan Africa, lower-middle-income countries, and high-income countries. The data includes observations from 40 out of 47 Sub-Saharan African countries, 46 out of 53 lower-middle-income countries, and 57 out of 83 high-income countries. We use data from the latest available year (2021 or 2022) from the Government Finance Statistics. The category "Other taxes" includes taxes on payroll and workforce, taxes on property, and any other taxes. This table is discussed in Section 2.1.

B Context Appendix

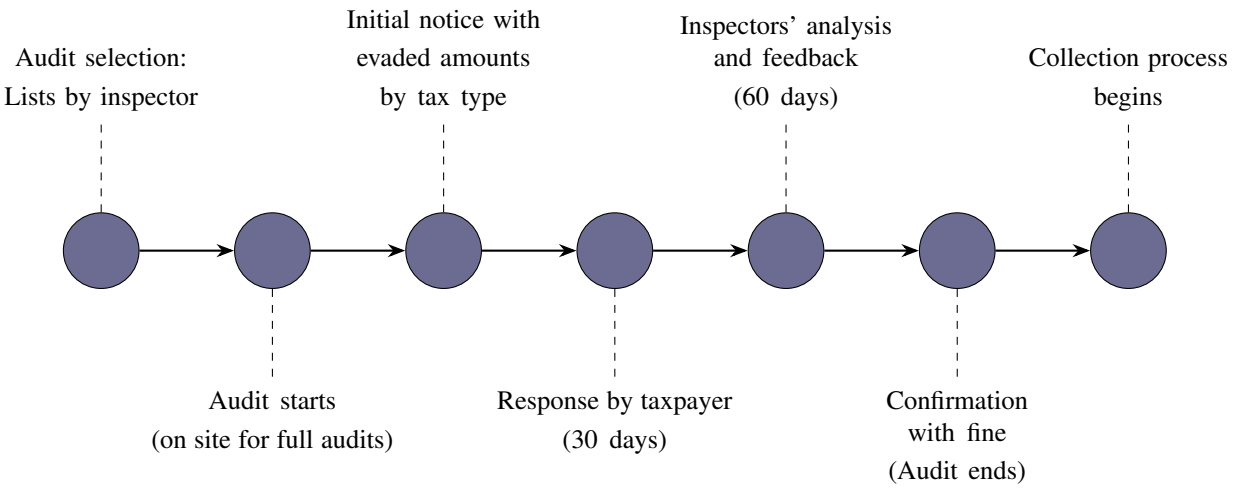
B.1 Tax Policy and Enforcement in Senegal

Figure B.1: Tax Offices Involved in Conducting Audits



Notes: This figure shows the units within the tax administration (DGID) in charge of conducting audits. The Large Taxpayer Unit (LTU) oversees firms with a turnover greater than 3 billion CFA francs (approximately 5.3 million USD). The LTU is divided into four units specialized by economic activity: LTU1 for mining and energy sectors, LTU2 for financial services and the telecommunications industry, LTU3 for real estate, and LTU4 Unit 4 for all other sectors. The medium taxpayer unit (MTU) oversees firms with a turnover between 100 million CFA francs and 3 billion CFA francs (MTU1) and those with a turnover between 100 million and 50 million CFA francs (MTU2), as well as regulated liberal professions, such as lawyers, notaries and medical practitioners. Other taxpayers are overseen by one of the 19 small taxpayer offices, organized by region. This figure is discussed in Section 2.2.

Figure B.2: Audit Process at the Senegalese Tax Authority



Notes: This figure shows the steps taken to inspect a taxpayer. At the beginning of the year, inspectors are given a list of audits agreed with their superior in the hierarchy. Upon conducting the audit, they draft an initial notice containing apparent infractions. The initial notice contains the value of presumed evasion and the corresponding fine. The taxpayer can respond to the notice providing evidence that they have complied, which the inspector analyses before sending a final notice. Shortly after the final notice, the taxpayer receives a request to pay the evaded amount plus fines. This figure is discussed in Section 2.2.

B.2 Audit Selection Around the World

Table B.2: Tax Audit Selection Methods in Selected Countries

Country	Discretionary selection	Risk analysis	Random selection
Kenya	Yes ; For all except large taxpayers	Yes ; Only for large taxpayers	No
Senegal	Yes	Yes, Introduced in FY 2018	Introduced in FY 2018
Zimbabwe	Yes; Inspectors rated on selection.	Yes; based on turnover variances	No
Lesotho	No	No	Yes ; Randomly by managers
Tanzania	Abandonned in 2007	Yes	
United Kingdom	Yes; For 55% of audit cases	Yes; Risk scoring	Yes ; Simple random sample
Switzerland	Yes for all cases	No	Yes, periodically for some taxes
United States	No	Yes	
France	Yes; For intelligence gathering	Yes; statistical techniques, data-mining	No
Bulgaria	Yes ; According to set criteria	Yes; Central risk analysis	No
Turkey	No	Yes; Analysis by tax type	Yes ; to collect unbiased data

Notes: This table is based on Khwaja et al. (2011) and our survey of select country tax officials.

B.3 Tax Inspector Performance Incentives in Senegal

Tax inspectors have a strong financial performance incentives. **Their remuneration is based on three elements: a base salary, a share from a "common fund", and an individual bonus. The common fund and the individual bonuses are paid based on the fines** (including also penalties, confiscations etc) that non-compliant taxpayers pay. The fines are typically 50% of the recovered amount of evasion. Less commonly, they can be 25% or 100% of the recovered amount.

For each audit, 10% of the fines paid are set aside for the **individual performance bonus** for auditors who worked on the specific case. Of these 10%, 2% goes the unit manager, 6% to the inspectors, and 2% to the recovery agents. When inspectors work as a team, the division of the sum in the team is determined in discussion with the manager. The remaining 90% of the fines enters the common fund, which is distributed among all agents in the relevant tax administration offices. This means all agents receive a share, from secretaries to managers.

An agent's **share of the common fund** depends on their function and level of seniority in the administration (rewarded based on "statutory points") and on their performance (rewarded based on "performance points"), as evaluated quarterly by their manager against pre-defined performance objectives. Managers and high-level technical advisors automatically receive the maximum number of 150 performance points. Exceptional performers (with performance of 80-100 out of 100) receive an increased shared of the common fund, while low performers (performance below 80) see their share of the common fund reduced or even eliminated (when performance is evaluated to be below 10). Detailed rules determine how seniority levels are determined and how performance evaluations are conducted. The rules also govern a set of special circumstances, e.g. moves between tax offices. In general, more senior inspectors have stronger performance incentives as they receive a larger share of the common fund.

This discussion is linked to the discussion on bureaucrat characteristics and incentives in Section 2.4.

C The Risk Scoring Algorithm

C.1 Motivation

A key feature of this project is to assist the Senegalese tax administration (DGID) to design a tool which assesses firms' tax evasion risk. Starting in 2017, the team held consultations with DGID leadership and former tax inspectors to map the compliance risks of Senegalese firms and to exploit all available data sources to assess this risk. Moreover, we discussed with experts in the field of taxation and risk management, who worked on tax evasion risk assessment in middle-income countries. With these inputs, we designed a risk-scoring tool, following best international practice, as implemented by the World Bank and its partner institutions.

Although the use of advanced machine-learning tools for prediction has exploded in economic analysis, it was decided together with DGID that the risk-score would be guided by simple variables which logically should predict evasion risk. The simplicity of the design is motivated by several factors, ranked by order of importance. First, the tool needed to be transparent, such that underlying compliance risks could be understood by tax inspectors, and explained to taxpayers when required. Second, the available data on historical audit results was sparse and not digitized, which limited the scope of our model calibration and model selection exercises (further details below). Finally, all cases concluded by 2017 were selected in a discretionary manner.

Thus, one should consider the risk-scoring tool as a transparent best-practice risk assessment, given the administrative capacity, rather than a fined-tool fully optimized algorithm. We note that the constraints faced by DGID are likely to bind in many low income countries, and especially in other West African countries, which often look at Senegal for administrative innovations.

Table C.1 below summarizes the steps we took in deriving the risk score.

Table C.1: Steps of Risk-Score Calculation

Step	Description
(1) Prepare database	The tax declarations of each taxpayer are merged across taxes (VAT, CIT, Payroll) and across years. Data from third parties is then merged in (customs, procurement, transaction network).
(2) Calculate inconsistency ratios	Inconsistencies are situations in which a self-reported tax liability can be considered as misreported or incomplete, by comparing different data sources. An example of an inconsistency ratio is third-party reported sales over self-reported sales.
(3) Calculate anomaly ratios	Anomalies correspond to abnormal reporting behavior, compared to peers. Anomalies may be associated with tax evasion, but do not indicate tax evasion behavior with certainty. An example of an anomaly ratio is the inverse of the profit rate.
(4) Define comparison clusters	Clusters regroup firms in the same economic sector and of comparable size. Peer comparisons of anomaly ratios are done within clusters.
(5) Transform ratios into risk indicators	For inconsistencies, the magnitude of the ratio is used to assign a value, ranging from one to ten (using deciles). For anomalies, firms within the top decile of a particular ratio within their cluster are assigned a value of one.
(6) Assign weights to indicators	Weights are assigned to each indicator reflecting our beliefs about their relative importance.
(7) Aggregate indicators and years	The weighted risk indicators are first aggregated for each year. Then the yearly scores are summed up to form a total risk score covering the past four years. More recent years are weighted higher than more distant years.
(8) Weigh risk score by declared turnover	The aggregated risk score is weighted by the log of turnover to give more importance to larger firms.

Notes: This table describes the steps taken in calculating the risk score based on which the algorithm selects firms for tax audits. This is discussed in Section 3.1.

C.2 Choosing Indicators and Weights

As explained above, the algorithm computes some ratios from the data of firms (declarations and third party data) and then calculates the value of the indicator based on the distribution of this ratio within a cluster of comparable firms. We tried several combinations of indicators before stabilizing the algorithm in a reduced set of them. The goal was to have a set of indicators that was sensible and correlated with evasion, but at the same time simple and understandable for the tax inspectors.

Table C.1 summarizes the steps that we took to conceptualize the algorithm. We tried out several possible indicators that could suggest under-declaration of tax liability. We discarded most based on some analysis of data availability or statistical relevance. In the end, we discarded indicators that required information that was available for a reduced set of firms and indicators that did not seem to have any correlation with evasion, as per past evasion data. We tested these indicators on data from historical audits data. We performed out of sample regressions with LASSO and OLS and computed the out of sample mean squared prediction errors to compare different models. This allowed us to assert that the ranking normalization performed well with respect to alternatives (meaning that it

presented a lower prediction error).

We decided to restrict the algorithm to a small list of indicators. Three of them are inconsistencies, plus a flag for inconsistent filing of taxes. On top of that, we have seven anomalies, of which two refer to value added tax, two refer to corporate income tax, one refers to third party data comparisons, one to share of imports from low tax countries and one refers to the financial services tax (only applicable to a reduces set of firms). The final list of indicators that is used in the algorithm, and the respective weights is summarized in the following table.

Some details for the calculation of the indicators are worth mentioning. In some cases of anomalies, the top decile within a cluster comprises more than 10% of cases. As long as the value is not zero, we include all these firms. Whenever there is not enough non-zero values that can fill un 10% of the firms, we only flag the non-zero values. We also top code (999 999 999) all values for which the denominator of te underlying ratio of the indicator is zero or missing. Therefore they belong by definition to the top decile. We also top code all values of negative tax liability, to make sure they also get flagged. The idea of the indicators is always that the larger the ratio, the less taxes the firm is paying.

We designed the risk-scoring scheme using best practices, drawing on policy documents from the World Bank (tax administration projects in Pakistan and Turkey), SKAT in Denmark, and the IMF's recommendations to DGID. We provide a high-level description of this process to preserve confidentiality around audit selection processes. We compute risk scores using information sets/tax returns submitted to DGID on corporate income taxes, VAT, personal income tax withholding remittance, as well external data from customs (imports/exports) and public procurement contracts, for the period 2013-2016.[45] The score relies on two types of risk indicators: discrepancies and anomalies. Discrepancy indicators flag taxpayers whose self-reported information according to their tax returns differs from information in datasets obtained from customs or the government budget department in charge of paying state procurement. For instance, a discrepancy indicator is logged when taxpayers' reported turnover over multiple years is lower than its aggregate costs, that its imports plus its wage bill over the same period. Anomaly indicators use industry/sector benchmarking to flag firms with unusual behavior relative to their peers. An example would be a firm in petroleum retail with low profit rate compared to its peers, which might be associated with evasion. Discrepancies and anomalies are aggregated to produce a risk-score for each taxpayer.

[45] We also attempted to apply predictive analytics from the machine learning literature on these datasets and on previous audit results was conducted to check whether risk indicators could predict DGID audit returns. This exercise was inconclusive because of the selected nature of the sample for whom audit returns are available, the small number of observations and noise in the data.

C.3 Statistics on Audit Program

Table C.2: Count of Selected Full Audits by Year, Tax Office, and Selection Method

		Algorithm	Discretion	Overlap	Total
DGE	2018	81	94	13	182
	2019	25	96	11	121
	2020	25	75	5	100
CME 1	2018	31	33	2	67
	2019	27	27	1	54
	2020	25	25	0	50
CME 2	2018	25	25	0	53
	2019	20	20	0	40
	2020	25	25	0	50
CPR	2018	15	15	0	32
	2019	15	15	1	30
	2020	20	16	2	36
Dakar P.	2018	0	0	0	0
	2019	14	15	2	29
	2020	7	7	0	14
Ngor A.	2018	0	0	0	0
	2019	11	10	0	21
	2020	8	8	0	16
Pikine G.	2018	0	0	0	0
	2019	8	7	0	15
	2020	8	8	0	16
G. Dakar	2018	0	0	0	0
	2019	9	8	0	17
	2020	8	8	1	16
All	2018	152	167	15	334
	2019	129	198	15	327
	2020	126	172	8	298
Total		407	537	38	959

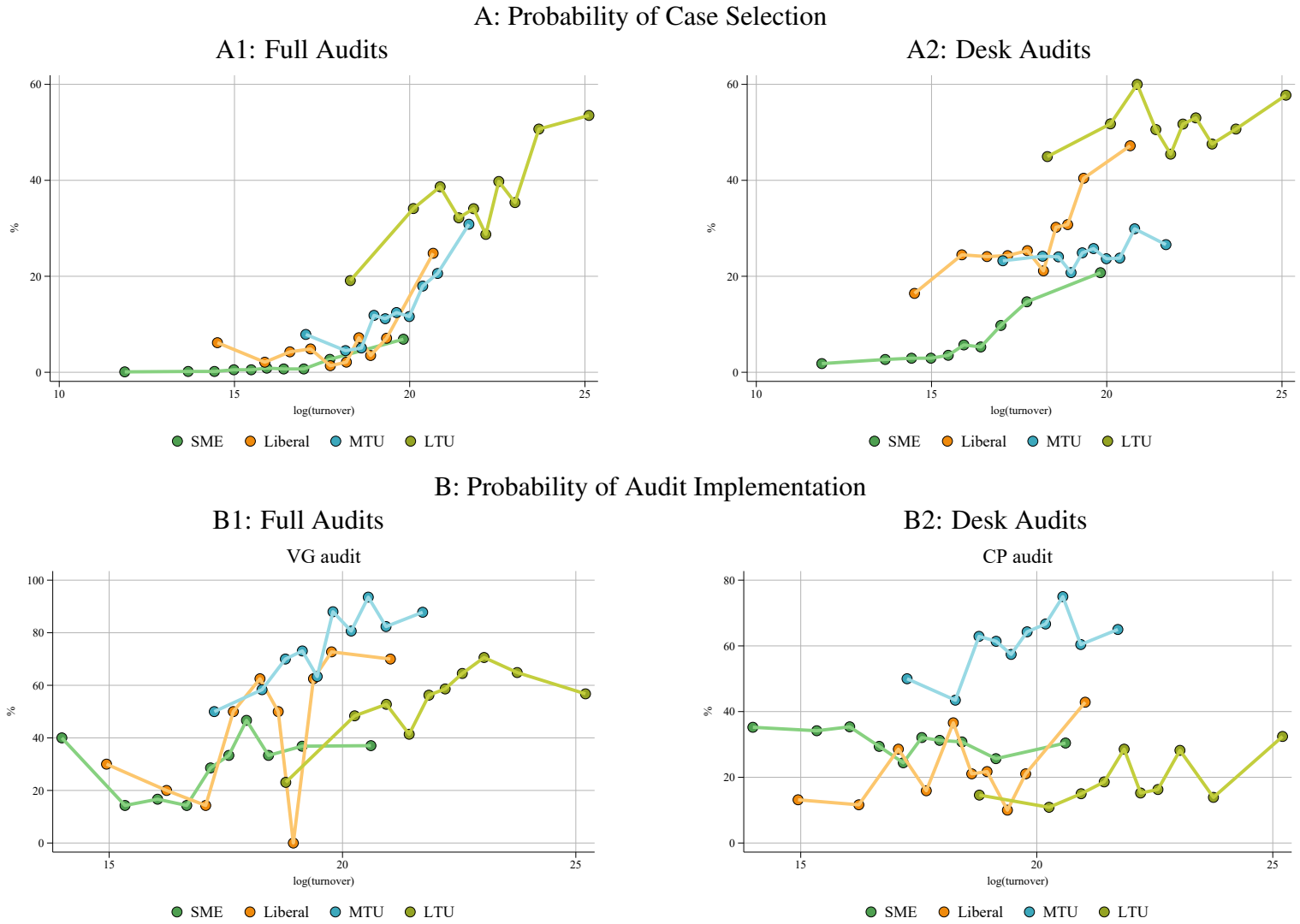
Notes: Number of selected full audits by year and tax office. The sum of the rows is larger than the total because there are overlapping cases between algorithm and discretion. This table is discussed in Section 3.2.

Table C.3: Count of Selected Desk Audits by Year, Tax Office, and Selection Method

		Random	Algorithm	Discretion	Overlap	Replacement	Total
DGE	2018	60	72	81	12	0	213
	2019	0	0	0	0	0	0
	2020	0	85	239	76	85	409
CME 1	2018	34	49	52	5	0	135
	2019	14	42	52	10	7	115
	2020	0	0	0	0	0	0
CME 2	2018	49	78	83	6	0	210
	2019	38	83	95	12	16	232
	2020	0	38	38	1	38	114
CPR	2018	59	86	95	10	0	240
	2019	26	66	70	4	12	174
	2020	0	70	70	7	70	210
Dakar P.	2018	0	0	0	0	0	0
	2019	37	71	78	7	15	201
	2020	0	72	72	1	72	216
Ngor A.	2018	0	0	0	0	0	0
	2019	19	57	59	2	10	145
	2020	0	49	59	2	49	157
Pikine G.	2018	0	0	0	0	0	0
	2019	14	26	26	0	8	74
	2020	0	63	63	2	63	189
G. Dakar	2018	0	0	0	0	0	0
	2019	14	10	12	2	8	44
	2020	0	53	53	2	53	159
All	2018	202	285	311	33	0	798
	2019	162	355	392	37	76	985
	2020	0	430	594	91	430	1454
Total		364	1070	1297	161	506	3237

Notes: Number of selected desk audits by year and tax office. The sum of the rows is larger than the total because there are overlapping cases between algorithm and discretion. This table is discussed in Section 3.2.

Figure C.1: Selection and Implementation of Audits Across the Firm-Size Distribution



Notes: This figure plots the share of selected and audited firms within the period 2018-2020 conditional on their decile of turnover and tax office. The values on the x-axis correspond to the mean of the decile within the four tax offices. The deciles were computed based on the mean declared turnover of the firms in the period 2017-2020, excluding firms that have zero turnover. For Panel A, we compute the deciles within tax office using the population of firms, and plot the share of firms within that decile-tax office. For Panel B, we condition the computation on the set of selected firms. This figure is mentioned in Section 3.5.

C.4 Balancing Tests

Table C.4: Balancing Test for Randomization of Ordering: Probability of Being on Top of the List

	(1) P(top)	(2) P(middle)	(3) P(top)	(4) P(middle)	(5) P(top)	(6) P(middle)	(7) P(top)	(8) P(middle)
Algorithm case	0.00264	0.00390						
	(0.0204)	(0.0139)						
log(Mean Turnover)			0.000889	0.000527				
			(0.00150)	(0.00148)				
log(Mean Tax Liability)					0.00115	0.00100		
					(0.00159)	(0.00156)		
Profit rate							0.0490	-0.104*
							(0.0592)	(0.0594)
N	3675	3675	3675	3675	3675	3675	3675	3675
R2	0.000523	0.000424	0.000609	0.000440	0.000664	0.000521	0.000720	0.00133
Mean outcome	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows the coefficients of a regression to predict the position of a case on the inspectors' list, conditional on characteristics. It predicts the probability that the case is located at the top third of the list, or in the middle third of the list. The table shows OLS results with fixed effects at the year X inspector level (tax office level for full audits). Robust standard errors (Huber-White) are shown in parentheses. This table is discussed in Section 4.

Table C.5: Balancing Tests of Information Intervention

Panel A: Firm Characteristics								
	Profit Rate	log(Turnover)	log(Payroll)	log(N. Employees)	log(Exports)	log(Tax Liability)	log(Firm's Age)	log(Distance to Firm)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Indicators	-0.01	0.12	-0.08	-0.12	-0.16	-0.24	-0.23	0.19
	(0.01)	(0.21)	(0.55)	(0.26)	(0.29)	(0.75)	(0.16)	(0.18)
+ data spreadsheets	0.01	0.00	0.79	-0.00	0.05	-0.22	0.02	0.05
	(0.01)	(0.21)	(0.54)	(0.26)	(0.31)	(0.75)	(0.16)	(0.18)
N	2136	2136	1452	1410	234	1420	2136	2136
R2	0.10	0.15	0.20	0.23	0.38	0.38	0.10	0.16
Mean outcome	-0.03	1.09	13.81	2.84	27.42	4.37	-0.17	-1.69
Panel B: Availability of Observations								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Indicators			0.01	0.01	-0.02	-0.01		
			(0.03)	(0.03)	(0.02)	(0.03)		
+ data spreadsheets			-0.02	-0.01	0.02	0.02		
			(0.03)	(0.03)	(0.02)	(0.03)		
N	2136	2136	2136	2136	2136	2136	2136	2136
R2	.	.	0.18	0.17	0.27	0.09	.	.
Mean outcome	1.00	1.00	0.68	0.66	0.12	0.66	1.00	1.00

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows results of OLS regressions of firm characteristics on the information treatments. The sample only includes desk audit cases for the years 2018, 2019, and 2020, which were the ones used in the intervention. The treatment was cross-randomized across algorithm and inspector cases in 2018 and 2019, but only used for algorithm cases in 2020. Therefore, we excluded the inspector-selected cases in 2020, which stemmed from a different selection method that is not directly comparable with the algorithm cases. The outcomes are based on tax declarations or firm registry data. For the tax data, we used information from the year before the firm was selected for audit. The outcomes are defined as follows: the profit rate is defined as total profits divided by total sales, obtained from CIT declarations (Column 1), log(turnover) is the natural logarithm of total sales (plus 1 to avoid dropping with zero turnover) obtained from CIT and VIT declarations (Column 2), log(payroll) is the log of total payroll (plus 1) obtained from the Pay-As-You-Earn declarations (Column 3), log(N. employees) is the log of the number of employees obtained from the Pay-As-You-Earn declarations (Column 4), log(exports) is the log of total value of exports (plus 1) obtained from customs data (Column 5), log(tax liability) is the log of the sum of VAT, CIT and PAYE liability as computed from the tax declarations (Column 6), log(Firm's age) is the log of the age in years of the firm obtained from the firm's date of creation (Column 7), and log(distance to firm) is the log of the distance from the tax office's location to the firm's premises in minutes computed using Google Maps on a Monday afternoon (Column 8). The regressions include fixed effects at the list level (inspector x year). Robust standard errors are shown in parentheses (Huber-White formula). This table is discussed in Section 6.2.

C.5 Characteristics of selection for Desk Audits

Table C.6: Firm Characteristics of Algorithm vs Discretionary Selection

Panel A: Characteristics of Inspector-Selected Cases

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Inspectors' Selection	4.71***	0.98***	0.04*	-0.28***	-1.02***	2.36***	-2.38	0.44	0.13
	(0.83)	(0.20)	(0.02)	(0.07)	(0.22)	(0.50)	(7.92)	(2.12)	(0.17)
N	22576	7433	5925	61238	60608	51992	696	702	640
R2	0.27	0.13	0.05	0.07	0.14	0.11	0.09	0.02	0.03
Mean outcome	10.82	14.81	-0.15	0.41	1.95	5.91	31.24	50.32	2.77

Panel B: Comparison Among Selected Cases

	Tax Declarations				Administrative		Taxpayer Survey		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	log(Turnover)	log(Payroll)	Profit Rate	P(Trade)	Duration Trip	Firm Age	Employees	% Sales in Cash	Audit Frequency
Algorithm	0.96***	-0.73**	-0.01	0.08***	0.57***	0.51***	1.53	3.49	-0.22
	(0.28)	(0.30)	(0.02)	(0.02)	(0.05)	(0.04)	(10.98)	(4.29)	(0.22)
Inspectors x Overlap	1.53***	-1.12*	-0.01	0.13***	0.23**	0.30***	5.54	1.61	-0.52
	(0.56)	(0.61)	(0.03)	(0.04)	(0.12)	(0.10)	(8.56)	(17.26)	(0.78)
Algorithm x Random	-0.79**	-0.12	-0.01	-0.11***	-0.32***	-0.25***	-5.27	4.95	0.04
	(0.39)	(0.45)	(0.03)	(0.03)	(0.09)	(0.07)	(7.75)	(5.72)	(0.32)
N	6369	4015	3304	13642	13411	13642	649	651	574
R2	0.29	0.17	0.15	0.26	0.24	0.35	0.29	0.23	0.22
Mean outcome	14.65	14.59	-0.06	0.17	1.44	1.31	22.61	51.43	2.71

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table depicts the OLS regression coefficients of characteristics on selected methods for desk audits. It reflects Table 3 on the main text, which shows the same results for full audits. Panel A uses the sample of selected cases across the three years of the experiment, including only desk audit selection. Panel B uses the sample of all firms observed in the administrative data, and the coefficient on Inspectors' Selection indicates that the firm was selected for audit at some point during 2018-2020. For Panel A, columns 1-6, the regression is a panel regression with fixed effects at the list level (year X tax office for full audits, year X inspector for desk audits). Otherwise, the regression is cross-sectional, with tax office fixed effects. The characteristics of the firms stem from three sources. From the tax declarations, we use the log of the yearly declared turnover, the log of the yearly declared payroll, the profit rate, and the probability that the firm has exports or imports. We use the value for the year before the firm was selected for audit. We use data from the firm registry on the firm's age and the distance between its location and DGID, in minutes of travel time (computed using GoogleMaps for a Monday at 3 PM). Finally, we use the taxpayer survey to compute the (self-reported) number of full-time employees, the share of total sales done in cash, and the perceived yearly frequency of full audits. Robust standard errors are shown in parentheses (Huber-White formula). This table is discussed in Section 3.5.

C.6 Deviations from Pre-Analysis Plan

Our analyses in Sections 4, 5, and 6 of the paper follow the pre-analysis plan, with small exceptions:

- We pre-specified the use of audit return outcomes (e.g. evasion) and the use of audit cost outcomes, but not the cost-effectiveness or productivity of audits, i.e. the ratio of the two pre-specified classes of outcomes. We nonetheless consider audit productivity in Section 5.2, as it is an important outcome from a policy perspective, especially given our results on audit returns and audit costs.
- We pre-specified but do not report self-reported difficulty of the audit and self-reported challenges encountered during the audit as outcomes, as inspectors' compliance with the reporting

requirement for these variables (submitted in pre-filed excel files) was limited. The data points are hence too few and selected to draw meaningful conclusions. We also did not use some audit process variables, such as whether inspectors requested additional information or a revised tax return, as the reporting of these outcomes is incomplete.

- We do not report heterogeneity, spillovers and learning effects of the information treatment in the desk audit sample, as the treatment had no statistically detectable effect on the main outcomes.
- The analysis in Sections 6.3 and 6.4 was not pre-specified. We think this extension of the analysis is justified as the sections are not about analyzing additional outcomes or additional heterogeneity dimensions, beyond those specified in the pre-analysis plan. Rather, the sections shed more light on the underlying reasons for the main results presented in Sections 4 and 5.
- We leave the analysis of medium-term outcomes, such as taxpayers' future compliance behavior, for a separate paper, given space constraints.

D Additional Results on Audit Outcomes

Table D.1: Algorithm Selection and Audit Outcomes, Intent-to-Treat Analysis

	P(Execution)		P(Detection | Execution)		log(Evasion) | Detection	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Panel A (Linear Regression)						
Algorithm	-0.18***	-0.04**	-0.14***	-0.05***	-3.63***	-0.82**
	(0.03)	(0.02)	(0.03)	(0.02)	(0.60)	(0.33)
Inspectors x Overlap	0.16**	0.04	0.20***	0.03	3.82***	0.67
	(0.07)	(0.04)	(0.07)	(0.04)	(1.46)	(0.68)
Algorithm x Random		-0.00		-0.01		-0.30
		(0.03)		(0.03)		(0.55)
N	944	2731	944	2731	890	2466
R2	0.26	0.32	0.24	0.19	0.27	0.25
Mean outcome	0.53	0.37	0.47	0.27	9.82	5.40
Panel B (Poisson Regression)						
Algorithm	-0.35***	-0.12**	-0.30***	-0.17***	-0.38***	-0.16**
	(0.06)	(0.05)	(0.07)	(0.06)	(0.06)	(0.06)
Inspectors x Overlap	0.27***	0.12	0.35***	0.11	0.32***	0.13
	(0.10)	(0.10)	(0.12)	(0.12)	(0.11)	(0.12)
Algorithm x Random		0.02		-0.01		-0.01
		(0.06)		(0.09)		(0.08)
N	925	2504	925	2493	871	2229
R2						
Mean outcome	0.54	0.40	0.48	0.30	10.03	5.97

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table is identical to Table 4 discussed in Section 4.1, except it shows intent-to-treat effects, where outcomes are set to zero for audits that are not implemented. Panel A shows ordinary least square results, and Panel B shows Poisson Pseudo-Maximum Likelihood estimates. Robust standard errors are shown in parentheses (Huber-White formula).

Table D.2: Algorithm Selection and Audit Outcomes, Controlling for List Slot Fixed Effect

	P(Execution)			P(Detection | Execution)			log(Evasion) | Detection		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	-0.18***	-0.05***	-0.04**	0.04	-0.05*	-0.04	-0.62***	-0.26**	-0.16
	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.18)	(0.13)	(0.14)
Inspectors x Overlap	0.16**	0.05	0.04	0.08	-0.08	-0.08	0.27	-0.60*	-0.41
	(0.07)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.47)	(0.32)	(0.37)
Algorithm x Random		-0.00	-0.00		0.00	0.01		-0.23	-0.22
		(0.03)	(0.03)		(0.04)	(0.04)		(0.18)	(0.19)
Tax Office x Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector x Year	No	No	Yes	No	No	Yes	No	No	Yes
N	944	2731	2731	507	1016	997	453	751	732
R2	0.27	0.23	0.32	0.16	0.32	0.43	0.32	0.29	0.43
Mean outcome	0.53	0.37	0.37	0.89	0.73	0.73	19.29	17.74	17.71

This table is identical to Table 4 discussed in Section 4.1, except that we now control for the placement of a case on the audit list, by adding list order quartile fixed effects. We use quartiles rather than list slot fixed effects, because the lists vary in length. The fixed effects help control for differential effort or attention over time within an audit program. The intent-to-treat effects are in Table D.1.

Table D.3: Algorithm Selection and the Evasion Rate

	% of Liability			% of Pre-Audit Turnover		
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	0.01	-0.03	-0.04	0.02	0.02	0.01
	(0.04)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)
Inspectors x Overlap	0.19**	-0.11**	-0.16***	0.14**	-0.05*	-0.06*
	(0.08)	(0.06)	(0.06)	(0.07)	(0.03)	(0.03)
Algorithm x Random		0.01	0.01		0.00	0.00
		(0.04)	(0.04)		(0.03)	(0.03)
Tax Office x Year	Yes	Yes	Yes	Yes	Yes	Yes
Inspector x Year	No	No	Yes	No	No	Yes
N	497	980	960	494	978	957
R2	0.10	0.19	0.34	0.11	0.14	0.31
Mean outcome	0.65	0.50	0.50	0.25	0.21	0.21

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Robust standard errors are shown in parentheses (Huber-White formula). These results are mentioned in Section 4.2.

Table D.4: Dispute of Audit Outcomes

	P(Confirmation= Notification)		log(Notification)| Both>0		log(Confirmation)| Both Not. and Conf. >0	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	0.13***	0.02	-0.35	-0.16	-0.19	-0.19
	(0.04)	(0.05)	(0.20)	(0.18)	(0.26)	(0.22)
Inspectors x Overlap	-0.03	0.03	0.66	-0.57	0.54	-0.06
	(0.07)	(0.10)	(0.40)	(0.40)	(0.38)	(0.48)
Algorithm x Random		-0.01		-0.11		0.05
		(0.09)		(0.26)		(0.27)
N	264	372	260	346	260	346
R2	0.16	0.24	0.32	0.55	0.19	0.51
Mean outcome	0.16	0.21	19.46	17.96	18.68	17.33

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case, controlling for list fixed effects (year x tax office for full audits, year x inspector for desk audits). To construct the outcomes we needed for each case data on both the initial notification of the evasion plus penalty amount, and its confirmation. Thus, the sample for the first outcome includes all cases with both **non-missing confirmation and notification**. This lowers the sample from 507 full audits to 264, and from 1016 desk audits to 372. For the second and third outcome, the sample is conditional on **both notification and confirmation being non-zero**. Robust standard errors (Huber-White) are shown in parentheses. This table is discussed in Section 4.3.

D.1 Estimates Underlying the Robustness Figure

Table D.5: Robustness of Main Results - Cluster S.E. at Tax Office level

	P(Execution)			P(Detect)|Execution			log(Evasion)|Detection		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	-0.18***	-0.05	-0.04	0.04	-0.05	-0.04	-0.64***	-0.25	-0.16
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.15)	(0.18)	(0.18)
Inspectors x Overlap	0.16***	0.05	0.04	0.08**	-0.07*	-0.08**	0.29	-0.58	-0.40
	(0.03)	(0.05)	(0.05)	(0.03)	(0.04)	(0.02)	(0.30)	(0.33)	(0.27)
Algorithm x Random		-0.00	-0.00		0.00	0.01		-0.22	-0.21*
		(0.03)	(0.03)		(0.05)	(0.04)		(0.13)	(0.10)
Tax Office x Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector x Year	No	No	Yes	No	No	Yes	No	No	Yes
N	944	2731	2731	507	1016	997	453	751	732
R2	0.26	0.23	0.32	0.15	0.32	0.43	0.31	0.29	0.43
Mean outcome	0.53	0.37	0.37	0.89	0.73	0.73	19.29	17.74	17.71

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. This table show the results for the main outcomes restricting to the lists in which the number of algorithm and discretionary cases were exactly the same. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Standard errors are clustered at the tax office x year levels, and inspector x year levels when inspector fixed effects are included. This table provides the estimates shown in Figure 2.

Table D.6: Robustness of Main Results - Equal Number of Cases (Definition 1)

	P(Execution)			P(Detect)|Execution			log(Evasion)|Detection		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	-0.23***	-0.02	-0.02	0.03	-0.02	-0.02	-0.51**	-0.11	-0.08
	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.24)	(0.24)	(0.25)
Inspectors x Overlap		0.08	0.07		0.04	-0.01		-1.42***	-1.12*
		(0.20)	(0.24)		(0.05)	(0.02)		(0.48)	(0.59)
Algorithm x Random		0.48***	0.47***		0.06	0.01		-1.61**	-1.43*
		(0.08)	(0.14)		(0.06)	(0.02)		(0.65)	(0.72)
Tax Office x Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector x Year	No	No	Yes	No	No	Yes	No	No	Yes
N	414	798	798	264	209	199	243	201	191
R2	0.19	0.17	0.23	0.09	0.07	0.18	0.27	0.39	0.46
Mean outcome	0.63	0.26	0.26	0.92	0.96	0.95	18.90	17.94	17.91

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. This table show the results for the main outcomes excluding the Large Taxpayer Unit. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Robust standard errors are shown in parentheses (Huber-White formula). This table provides the estimates shown in Figure 2.

Table D.7: Robustness of Main Results - Equal Number of Cases (Definition 2)

	P(Execution)			P(Detect)|Execution			log(Evasion)|Detection		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	-0.22***	-0.03	-0.03	0.05	-0.06**	-0.04	-0.64***	-0.16	0.07
	(0.05)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.23)	(0.26)	(0.28)
Inspectors x Overlap	0.01	0.10	0.09	0.21*	-0.08	-0.12	2.21***	-1.64***	-1.28**
	(0.14)	(0.08)	(0.09)	(0.12)	(0.09)	(0.10)	(0.48)	(0.51)	(0.52)
Algorithm x Random		0.03	-0.02		0.09*	0.04		-0.46	-0.71
		(0.13)	(0.13)		(0.05)	(0.07)		(0.77)	(0.74)
Tax Office x Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector x Year	No	No	Yes	No	No	Yes	No	No	Yes
N	366	838	838	224	180	173	200	175	168
R2	0.21	0.13	0.19	0.12	0.07	0.25	0.18	0.27	0.43
Mean outcome	0.61	0.21	0.21	0.89	0.97	0.97	18.58	17.91	17.93

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. This table show the results for the main outcomes only for the year 2019, which we consider the implementation of the experiment to be best. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Robust standard errors are shown in parentheses (Huber-White formula). This table provides the estimates shown in Figure 2.

Table D.8: Robustness of Main Results - More Algorithm Cases

	P(Execution)			P(Detect)|Execution			log(Evasion)|Detection		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	-0.17***	-0.04**	-0.04*	0.02	-0.05*	-0.04	-0.68***	-0.17	-0.03
	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.19)	(0.23)	(0.27)
Inspectors x Overlap	0.13*	0.08*	0.07	0.08	-0.08	-0.09	0.27	-1.23***	-0.98*
	(0.07)	(0.04)	(0.04)	(0.05)	(0.06)	(0.08)	(0.49)	(0.43)	(0.55)
Algorithm x Random		0.03	-0.01		0.08*	0.05		-0.38	-0.56
		(0.12)	(0.13)		(0.04)	(0.07)		(0.77)	(0.72)
Tax Office x Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector x Year	No	No	Yes	No	No	Yes	No	No	Yes
N	855	1180	1180	474	226	208	428	220	203
R2	0.26	0.12	0.20	0.12	0.07	0.25	0.30	0.29	0.46
Mean outcome	0.55	0.19	0.19	0.90	0.97	0.97	19.34	18.14	18.10

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. This table show the results for the main outcomes. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Robust standard errors are shown in parentheses (Huber-White formula). This table provides the estimates shown in Figure 2.

Table D.9: Robustness of Main Results - Excluding Large Taxpayer Unit

	P(Execution)			P(Detect)|Execution			log(Evasion)|Detection		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	-0.19***	-0.05**	-0.05**	0.07**	-0.05*	-0.04	-0.61***	-0.31**	-0.21
	(0.04)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.20)	(0.14)	(0.14)
Inspectors x Overlap	0.22*	0.01	-0.00	0.15**	-0.08	-0.08	1.01	-0.23	-0.18
	(0.13)	(0.05)	(0.05)	(0.07)	(0.07)	(0.06)	(0.87)	(0.36)	(0.40)
Algorithm x Random		-0.01	-0.01		0.01	0.01		-0.19	-0.17
		(0.03)	(0.03)		(0.04)	(0.04)		(0.19)	(0.20)
Tax Office x Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector x Year	No	No	Yes	No	No	Yes	No	No	Yes
N	548	2194	2194	319	906	900	285	645	639
R2	0.28	0.22	0.30	0.20	0.31	0.41	0.17	0.16	0.29
Mean outcome	0.58	0.41	0.41	0.89	0.71	0.71	18.63	17.43	17.43

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. This table show the results for the main outcomes. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Robust standard errors are shown in parentheses (Huber-White formula). This table provides the estimates shown in Figure 2.

Table D.10: Robustness of Main Results - Only 2019 Program

	P(Execution)			P(Detect)|Execution			log(Evasion)|Detection		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	-0.14**	-0.06*	-0.06*	0.08	-0.07	-0.06	-0.87***	-0.42**	-0.27
	(0.05)	(0.04)	(0.03)	(0.06)	(0.05)	(0.04)	(0.28)	(0.19)	(0.20)
Inspectors x Overlap	0.23**	-0.02	-0.01	0.14***	-0.15	-0.15*	0.52	-0.77**	-0.48
	(0.11)	(0.08)	(0.08)	(0.05)	(0.10)	(0.09)	(0.71)	(0.39)	(0.43)
Algorithm x Random		0.06	0.06		0.00	0.00		-0.13	-0.14
		(0.04)	(0.04)		(0.06)	(0.06)		(0.25)	(0.25)
Tax Office x Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector x Year	No	No	Yes	No	No	Yes	No	No	Yes
N	327	909	909	195	547	547	163	299	299
R2	0.16	0.10	0.24	0.18	0.14	0.28	0.31	0.20	0.33
Mean outcome	0.59	0.60	0.60	0.83	0.54	0.54	19.32	17.43	17.43

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. This table show the results for the main outcomes. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Robust standard errors are shown in parentheses (Huber-White formula). This table provides the estimates shown in Figure 2.

E Additional Results on the Audit Process

Table E.1: Inspector Characteristics by Case Selection Method, Full Audits

Panel A: Full-Audit Cases in Which Some Inspector Reported Information							
	(1) Mean Age	(2) Max Age	(3) Mean Years of Experience	(4) Max Years of Experience	(5) Share with Masters/PhD	(6) Max Education	(7) Share in Favor of Algorithm
Algorithm	-0.28	-0.86**	-0.03	-0.38	0.04	-0.08	-0.02
	(0.32)	(0.41)	(0.21)	(0.24)	(0.03)	(0.05)	(0.03)
N	459	459	422	422	422	422	422
R2	0.36	0.40	0.37	0.42	0.33	0.34	0.32
Mean outcome	37.10	40.52	8.29	9.54	0.81	3.23	0.82
Panel B: Only Cases in Which All Inspectors Reported Information							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Algorithm	-0.17	-0.65	0.15	-0.12	0.01	-0.04	-0.02
	(0.36)	(0.48)	(0.36)	(0.36)	(0.04)	(0.04)	(0.05)
N	230	230	117	117	117	117	117
R2	0.42	0.36	0.29	0.21	0.60	0.85	0.18
Mean outcome	37.39	41.85	9.35	11.49	0.76	3.35	0.75
Panel C: Probability that All Inspectors Reported Information							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Algorithm	-0.01	-0.01	0.02	0.02	0.02	0.02	0.02
	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
N	507	507	507	507	507	507	507
R2	0.27	0.27	0.41	0.41	0.41	0.41	0.41
Mean outcome	0.45	0.45	0.23	0.23	0.23	0.23	0.23

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows the OLS estimation results for the mean characteristics of the team members working on cases. The regression is done at the case level only for started full audits. Full audits are assigned at the tax office level, and tax office leaders compose teams to work on the cases. Since there is no assignment of cases to inspectors at the selection phase, we can only run this regression for the cases that were effectively conducted. The outcomes are obtained from the inspector survey, and are averaged across the inspectors that worked in the case. Columns 1 and 2 show the results using as outcomes the mean and max age of the team members. Columns 3 and 4 use the share of team members with a Masters/PhD and the max of a categorical education variable that goes from 1 (only high school) to 4 (Masters/PhD). Column 5 uses the mean response of team members to whether they would favor the use of an algorithm to automate selection of cases. Columns 7, 8, and 9 use the share of highly experience (more than median) team members in the team, the mean experience, and the max experience among team members. The regressions control for fixed effects at the tax office-year level. Robust standard errors are shown in parentheses (Huber-White formula). Coefficients on the inspectoroverlap dummy are included in the regression but omitted from the display. This table is discussed in Section 5.1.

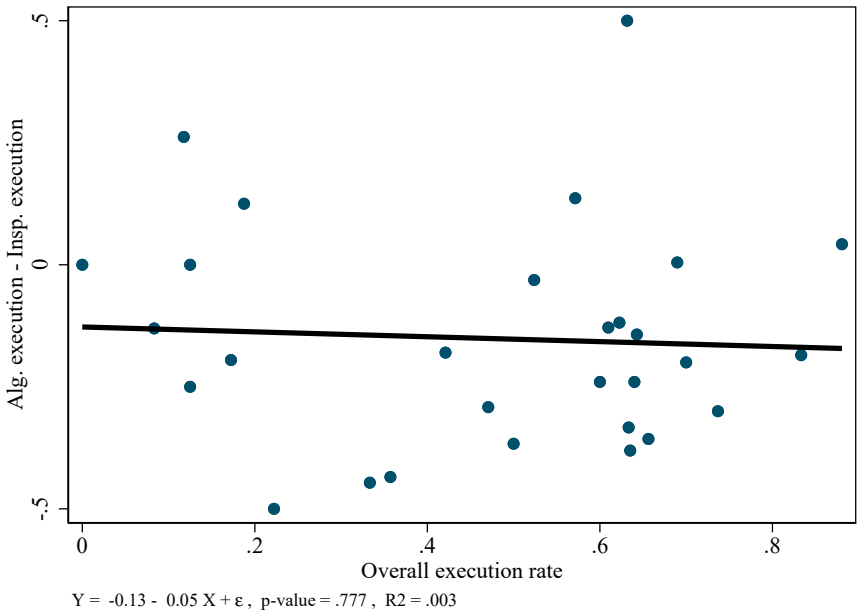
Table E.2: Inspector Characteristics and Performance, Desk Audits

	(1) # cases Executed	(2) P(execution)	(3) P(detection|execution)	(4) log(evasion) (Mean)	(5) Executed Alg. cases	(6) P(execution| Alg.)	(7) Executed Alg. > Executed Insp.	(8) P(In favor of Alg.)
Masters/PhD	-0.76	-0.01	0.06	0.79	-0.05	0.03	0.18**	0.12
	(0.72)	(0.04)	(0.06)	(1.29)	(0.45)	(0.05)	(0.08)	(0.11)
Above median age	0.50	0.05	0.02	0.23	0.11	0.04	-0.05	-0.06
	(0.53)	(0.04)	(0.04)	(0.83)	(0.36)	(0.06)	(0.09)	(0.13)
Above median experience	-0.52	-0.05	-0.00	-0.51	0.18	-0.01	0.11	-0.12
	(0.44)	(0.04)	(0.04)	(0.92)	(0.31)	(0.05)	(0.09)	(0.10)
N	86	86	76	76	86	86	86	86
R2	0.81	0.54	0.56	0.56	0.75	0.54	0.47	0.07
Mean outcome	4.77	0.32	0.90	16.92	2.53	0.29	0.34	0.77

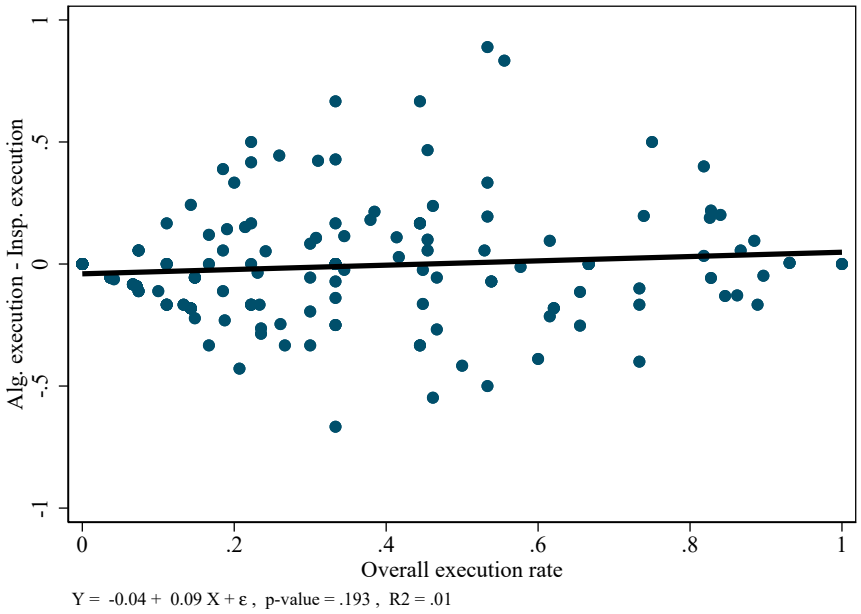
Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows the estimation of the OLS regressions of outcomes of inspectors on inspector characteristics. The analysis is done at the list (inspector x year) level and only includes the lists of desk audits. These lists were assigned individually to inspectors. The regressors are three individual characteristics of the inspectors collected using the inspector survey (whether the inspector had a Masters/PhD and whether they had higher than median experience working at the tax authority) and administrative data (age). Column 1's outcome is the total started cases for each list, Columns 2-4 are averages within list of the main outcomes (see Table 4), Column 5 uses the number of started algorithm cases, Column 6 the share of started algorithm cases among algorithm-selected cases, Column 7 uses an indicator of whether the number of started algorithm cases was larger than the number of started inspector cases in the list, and Column 8 uses the response of the inspector to whether they would favor the automation of case selection using an algorithm. The regressions control for fixed effects at the tax office-year level. Robust standard errors are shown in parentheses (Huber-White formula). This table is discussed in Section 5.1.

Figure E.1: Absence of Correlation Between Audit Execution Rate and Execution Rate for Algorithm-Selected Cases

(a) Full Audits



(b) Desk Audits



Notes: This figure plots outcomes at the audit-list level (i.e. the tax office × year level for full audits and the tax inspector × year level for desk audits). The X-axis shows the mean execution rate for all cases on the annual list. The Y-axis shows the difference in the execution rate between algorithm-selected cases and inspector-selected cases. This figure is mentioned in Section 5.1.

Table E.3: Algorithm Selection and Scope of Audits (1/2)

	N. infractions		N. years		Fine/Evasion	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	-1.21***	-0.47***	-0.36***	-0.28**	0.01	0.03***
	(0.30)	(0.15)	(0.13)	(0.13)	(0.01)	(0.01)
Inspectors x Overlap	0.13	-0.33	0.28	-0.18	0.00	-0.00
	(0.93)	(0.56)	(0.25)	(0.30)	(0.03)	(0.02)
Algorithm x Random		0.40*		0.41**		-0.00
		(0.23)		(0.18)		(0.02)
N	453	732	453	732	450	716
R2	0.32	0.33	0.19	0.29	0.21	0.27
Mean outcome	4.92	3.08	3.48	2.58	0.41	0.44

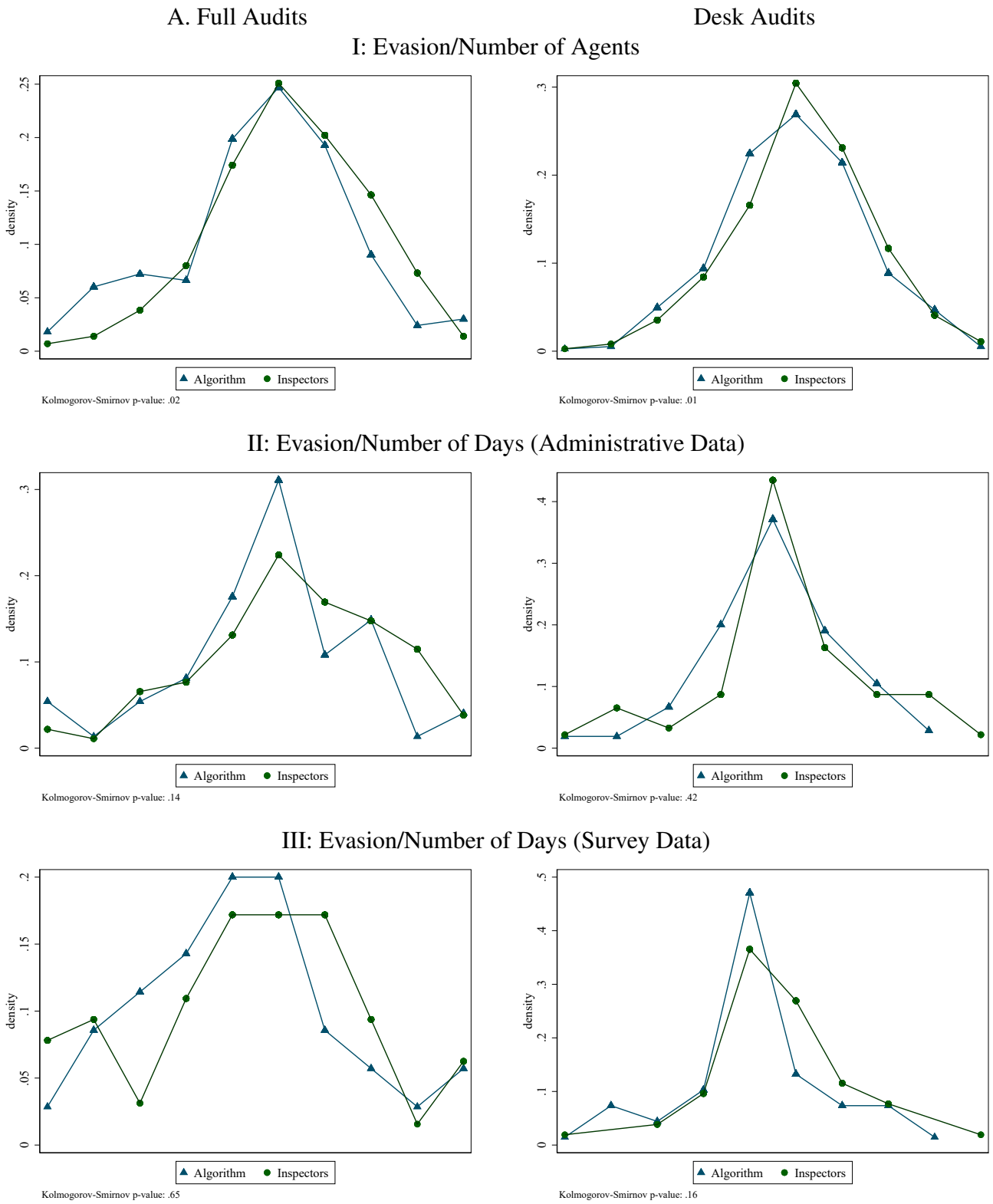
Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case, controlling for list fixed effects (year x tax office for full audits, year x inspector for desk audits). The outcomes are extracted from the audit reports. The three outcomes are i) the number of years from the taxpayers' declarations in which the inspector found an infraction (notice that the inspection can investigate tax declarations up to four years before the audit date according to Senegalese law), ii) the number of different infractions found by the inspector, and iii) the severity of the infraction as indicated by the ratio of fine to the evaded amount. Robust standard errors are shown in parentheses (Huber-White formula). Sample is conditioned on cases that started an audit. This table is discussed in Section 5.1.

Table E.4: Algorithm Selection and Scope of Audits (2/2)

	Share severe infractions (main taxes)		Share severe infractions (all taxes)		P(infraction in main taxes)		P(severe infraction in main taxes)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	1.38	2.77	1.72	4.44	0.02	-0.04	-0.02	-0.01
	(3.71)	(3.75)	(3.20)	(3.05)	(0.02)	(0.03)	(0.04)	(0.04)
Inspectors x Overlap	-2.99	3.92	-13.68**	5.19	0.05**	0.04	-0.03	0.04
	(7.81)	(8.34)	(5.39)	(6.80)	(0.03)	(0.06)	(0.09)	(0.09)
Algorithm x Random		4.19		3.33		-0.01		0.03
		(5.22)		(4.23)		(0.05)		(0.05)
N	453	732	453	732	453	732	453	732
R2	0.18	0.31	0.29	0.37	0.18	0.28	0.11	0.26
Mean outcome	60.84	64.31	60.76	71.59	0.94	0.86	0.78	0.72

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case, controlling for list fixed effects (year x tax office for full audits, year x inspector for desk audits). The outcomes are extracted from the audit reports. The three outcomes are i) the number of years from the taxpayers' declarations in which the inspector found an infraction (notice that the inspection can investigate tax declarations up to four years before the audit date according to Senegalese law), ii) the number of different infractions found by the inspector, and iii) the severity of the infraction as indicated by the ratio of fine to the evaded amount. Robust standard errors are shown in parentheses (Huber-White formula). Sample is conditioned on cases that started an audit. This table provides additional outcomes for the analysis discussed in Section 5.1.

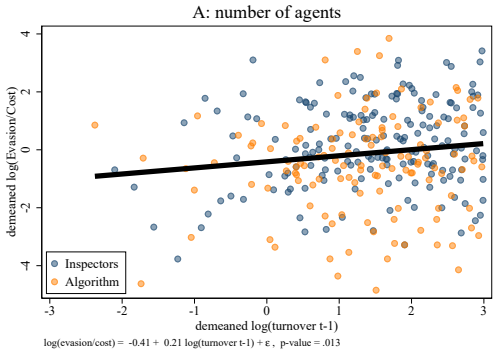
Figure E.2: Distribution of Audit Productivity



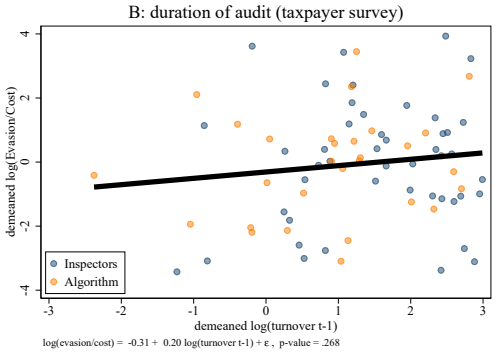
Notes: This figure plots density distributions of audit productivity outcomes, as per the panel titles. The outcomes are demeaned at the list level. The data is split into ten equally-spaced bins for each audit outcome and type. This figure is discussed in Section 5.1.

Figure E.3: Productivity of Uncovering Evasion Using Different Definitions of Cost

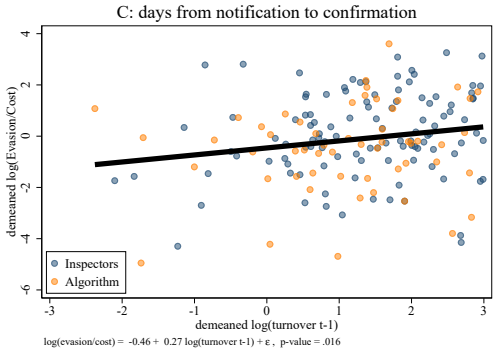
(a) Number of Agents



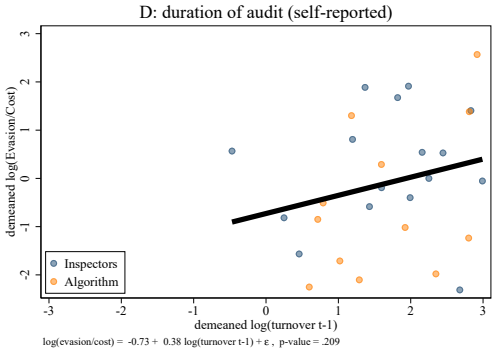
(b) Duration of Audit (Taxpayer Survey)



(c) Days from Start to Notification



(d) Duration of Audit (Self-Reported)



Notes: This figure plots the data points of audit cases. The Y-axis represents the productivity of uncovering evasion dividing evasion by three different cost measures. The X-axis represents the log of the firm's turnover as declared one year before the audit. All the variables were demeaned at the list level. The lines show a local non-parametric regression around the points, run separately for the algorithm-selected and inspector-selected cases. This figure is discussed in Section 5.1.

Table E.5: Selection Method and Survey Results for Corruption - Probability of Answering the Question

Panel A: All Interviewed Firms

	Corruption in General			Corruption Experience			Grade on Honesty			Friend at Tax Auth.		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Full audits	Desk audits	All	Full audits	Desk audits	All	Full audits	Desk audits	All	Full audits	Desk audits	All
Algorithm	-0.0218	0.00369	-0.00758	-0.0134	-0.00530	-0.00904	0.00375	0.0196	0.0122	0.0332	0.0469	0.0409
	(0.0743)	(0.0700)	(0.0508)	(0.0161)	(0.0233)	(0.0145)	(0.0518)	(0.0574)	(0.0386)	(0.0353)	(0.0361)	(0.0252)
Inspectors x Overlap	-0.138	-0.263	-0.209	0.0213	0.0140	0.0168	-0.0237	-0.0114	-0.0176	0.142*	0.0609	0.0959
	(0.302)	(0.189)	(0.168)	(0.0176)	(0.0169)	(0.0120)	(0.147)	(0.139)	(0.101)	(0.0690)	(0.124)	(0.0754)
Algorithm x Random		-0.0346	-0.0259		0.0123	0.0145		-0.0870	-0.0832		0.00195	0.00686
		(0.0831)	(0.0789)		(0.0227)	(0.0205)		(0.0574)	(0.0549)		(0.0626)	(0.0606)
N	254	369	623	254	369	623	254	369	623	254	369	623
R2	0.0847	0.231	0.174	0.133	0.156	0.146	0.146	0.252	0.226	0.0866	0.137	0.117
Mean outcome	0.64	0.60	0.61	0.96	0.97	0.97	0.81	0.71	0.75	0.87	0.86	0.86

Panel B: Only Audited Firms

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Algorithm	-0.0336	-0.137	-0.0820	-0.0314	-0.0173	-0.0240	-0.0128	-0.0345	-0.0250	0.0251	0.0962*	0.0671*
	(0.111)	(0.0826)	(0.0664)	(0.0371)	(0.0354)	(0.0252)	(0.0700)	(0.0802)	(0.0535)	(0.0487)	(0.0567)	(0.0389)
Inspectors x Overlap	0.529***	-0.186	0.105	-0.00192	0.0328	0.0173	-0.000782	0.00671	0.00701	0.226*	0.0489	0.107
	(0.164)	(0.310)	(0.229)	(0.00264)	(0.0278)	(0.0182)	(0.00432)	(0.178)	(0.105)	(0.115)	(0.241)	(0.158)
Algorithm x Random		0.115	0.0974		0.0417	0.0447**		-0.00276	-0.00787		0.0572	0.0753
		(0.110)	(0.105)		(0.0265)	(0.0214)		(0.0821)	(0.0756)		(0.0892)	(0.0858)
N	116	179	295	116	179	295	116	179	295	116	179	295
R2	0.0932	0.275	0.207	0.233	0.172	0.203	0.0720	0.260	0.236	0.102	0.198	0.168
Mean outcome	0.70	0.60	0.64	0.95	0.97	0.96	0.87	0.73	0.78	0.87	0.82	0.84

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results with fixed effects at the year x tax office level for outcomes from the taxpayer survey. Given the low number of observations in the taxpayer survey, it is not possible to run the regression with year x inspector fixed effects for desk audits, as is done with outcomes extracted from administrative data. Panel A is for all selected firms interviewed in the survey, whereas Panel B is restricted to firms that were audited according to the administrative audit reports. Robust standard errors are shown in parentheses (Huber-White formula). The outcomes are as follows: "Corruption in general" (Columns 1 to 3) is the declared belief of the percentage of audits in Senegal that are affected by corruption; "Corruption experience" (Columns 4 to 6) means that the respondent declared to have experienced an instance of bribery to obtain tax favors; "Grade on honesty" (Columns 7 to 9) is a grade from 0 to 10 to the honesty of the inspector the respondent last interacted with during a full audit; and "Friend at tax authority" captures the degree to which the respondent agrees with the statement that "if the firm's boss has a friend at the tax authority, the firm will rarely be audited". This table is connected to Table 7 discussed in Section 5.3.

F Additional Results on Mechanisms

Table F.1: Association of Risk Score with Audit Outcomes, Controlling for Firm Size

	P(Execution)		P(Detect)|Execution		log(Evasion)|Detection	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	-0.25***	-0.07**	0.01	-0.07*	-0.99***	-0.32
	(0.05)	(0.03)	(0.04)	(0.04)	(0.28)	(0.21)
Risk score	0.02	-0.01	-0.00	-0.00	0.34***	0.38***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.13)	(0.13)
Alg. x Risk score	0.01	0.04*	0.02	0.02	-0.29*	-0.42***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.15)	(0.13)
Lagged log(turnover)	0.02***	0.01***	0.01*	0.01***	0.04**	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)	(0.01)
N	944	2731	507	997	453	732
R2	0.33	0.32	0.16	0.44	0.33	0.44
Mean outcome	0.53	0.37	0.89	0.73	19.29	17.71

This table is identical to Table 8, except that we had lagged log turnover as an additional regressor. This table is discussed in Section 6.1.

Table F.2: Association of Risk Score with Audit Outcomes, Using the Unweighted Risk Score

	P(Execution)		P(Detect)|Execution		log(Evasion)|Detection	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	-0.25***	-0.09***	0.08*	-0.07	-0.86**	-0.32
	(0.06)	(0.03)	(0.05)	(0.04)	(0.33)	(0.23)
Unw. Risk score	0.06**	0.01	0.02	0.02	0.31*	0.09
	(0.03)	(0.02)	(0.02)	(0.02)	(0.17)	(0.17)
Alg. x Unw. Risk score	-0.04	0.03	-0.04	0.00	-0.28	0.02
	(0.04)	(0.02)	(0.03)	(0.03)	(0.22)	(0.19)
N	944	2731	507	997	453	732
R2	0.27	0.32	0.15	0.43	0.32	0.43
Mean outcome	0.53	0.37	0.89	0.73	19.29	17.71

This table is identical to Table 8, except that we use the raw risk score, prior to reweighting in by turnover, as a regressor. This table is discussed in Section 6.1.

Table F.3: Association of Risk Score Components with Audit Outcomes

	P(Execution)		P(Detect)|Execution		log(Evasion)|Detection							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Full audits	Full audits	Desk audits	Desk audits	Full audits	Full audits	Desk audits	Desk audits	Full audits	Full audits	Desk audits	Desk audits
Algorithm	-0.28***	-0.28***	-0.07***	-0.08**	-0.01	-0.01	-0.03	-0.00	-1.09***	-1.26***	0.03	-0.07
	(0.05)	(0.06)	(0.03)	(0.03)	(0.04)	(0.05)	(0.04)	(0.05)	(0.35)	(0.38)	(0.24)	(0.28)
VAT anomalies risk score		0.02		0.01		0.02		-0.00		0.27		0.21
		(0.03)		(0.02)		(0.02)		(0.02)		(0.22)		(0.17)
CIT anomalies risk score		-0.04		0.02		-0.01		0.00		0.12		-0.07
		(0.03)		(0.02)		(0.03)		(0.02)		(0.20)		(0.17)
Inconsistencies risk score	0.04	0.04	0.01	0.01	0.02	0.02	-0.00	-0.00	0.26	0.24	-0.13	-0.15
	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.03)	(0.20)	(0.21)	(0.14)	(0.15)
Anomalies risk score	0.03*	0.05	0.01	-0.00	0.03*	0.02	-0.01	-0.01	0.18	-0.04	-0.03	-0.11
	(0.02)	(0.04)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.12)	(0.21)	(0.12)	(0.21)
Algorithm x Inconsistencies	-0.02	-0.02	0.01	0.01	-0.02	-0.02	-0.00	0.00	-0.37	-0.38	0.18	0.21
	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.25)	(0.26)	(0.18)	(0.18)
Algorithm x Anomalies	-0.04	-0.05	-0.02	-0.01	-0.04	-0.05	0.01	-0.02	-0.01	0.24	0.06	0.22
	(0.03)	(0.04)	(0.02)	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)	(0.18)	(0.27)	(0.15)	(0.26)
Algorithm x VAT anomalies		-0.06		-0.01		-0.01		0.06		-0.45*		-0.42*
		(0.04)		(0.02)		(0.04)		(0.04)		(0.27)		(0.23)
Algorithm x CIT anomalies		0.06*		-0.01		0.03		-0.02		0.01		0.21
		(0.03)		(0.02)		(0.03)		(0.03)		(0.24)		(0.20)
N	944	944	2731	2731	507	507	997	997	453	453	732	732
R2	0.26	0.27	0.32	0.32	0.15	0.16	0.43	0.44	0.32	0.33	0.43	0.43
Mean outcome	0.53	0.53	0.37	0.37	0.89	0.89	0.73	0.73	19.29	19.29	17.71	17.71

This table is similar to Table 8, but shows the association between audit outcomes and components of the risk score. This table is discussed in Section 6.1.

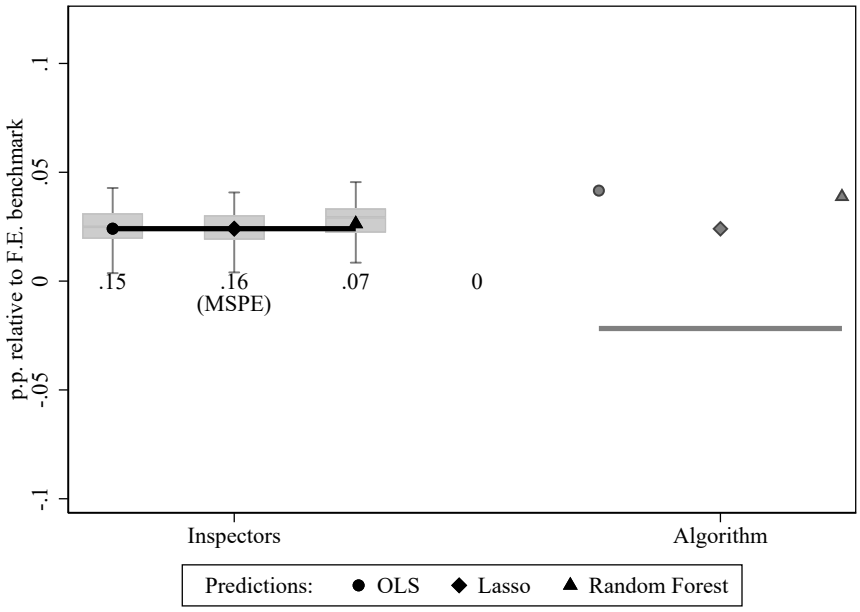
Table F.4: Treatment Effect of Information Intervention

	(1) P(Execution)	(2) P(Detection| Execution)	(3) log(Evasion)| Detection	(4) P(Start)	(5) P(Detection| Execution)	(6) log(Evasion)| Detection
Algorithm	-0.05	-0.08*	-0.22	-0.05	-0.08*	-0.22
	(0.04)	(0.05)	(0.22)	(0.04)	(0.05)	(0.22)
Algorithm x Random	-0.00	0.01	-0.24	-0.00	0.01	-0.25
	(0.03)	(0.04)	(0.19)	(0.03)	(0.04)	(0.19)
Information	0.02	-0.05	-0.22			
	(0.04)	(0.04)	(0.20)			
Algorithm x Information	0.01	0.06	0.27			
	(0.04)	(0.06)	(0.26)			
Info. (indicators)				0.03	-0.05	-0.07
				(0.04)	(0.05)	(0.23)
Info. (indicators+data)				0.01	-0.05	-0.37
				(0.04)	(0.05)	(0.24)
Alg. x Info. (Indicators)				0.02	0.07	0.07
				(0.05)	(0.07)	(0.31)
Alg. x Info. (Indicators+data)				0.01	0.06	0.45
				(0.05)	(0.06)	(0.31)
N	2136	896	631	2136	896	631
R2	0.32	0.41	0.43	0.32	0.41	0.43
Mean outcome	0.42	0.70	17.66	0.42	0.70	17.66

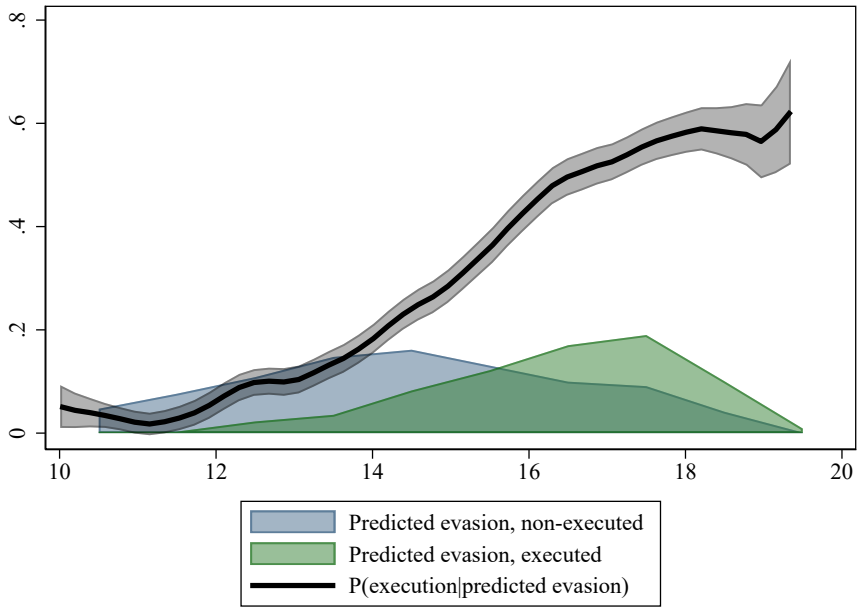
Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows the estimation of the randomized information treatment intervention's effect on the probability of audit execution, and the correlation of the information treatment with subsequent audit outcomes. Only desk audit cases were used in the information intervention, such that the sample does not include full audit cases. The treatment was cross-randomized across algorithm and inspector cases in 2018 and 2019, but only used for algorithm cases in 2020. Therefore, we excluded the inspector-selected cases in 2020. Columns 1, 2, and 3 show the results for a specification containing a dummy indicating whether the case was treated. Columns 4, 5, and 6 distinguish between two modalities of the treatment: providing only indicators of risk about the case to the inspectors and providing risk indicators plus a spreadsheet with data on the taxpayers' tax declarations and third-party data. OLS results with fixed effects at the year X inspector level. Robust standard errors are shown in parentheses (Huber-White formula). This table is discussed in Section 6.2.

Figure F.1: Predicting Inspector Choices and Audit Outcomes Using Observable Characteristics

Panel A: Predicted and Realized Execution of Cases



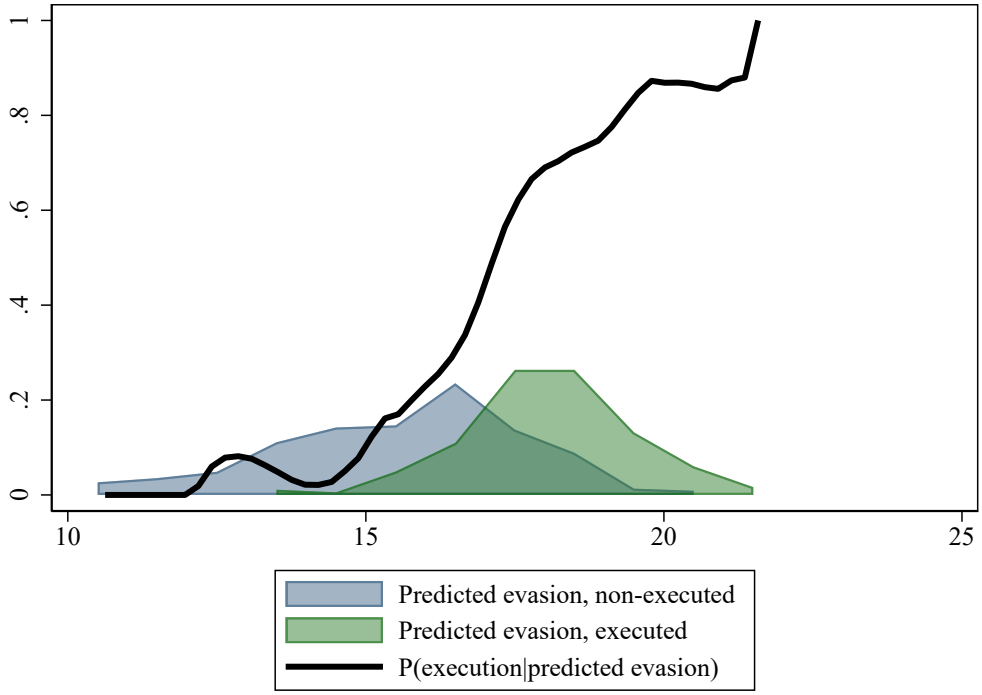
Panel B: Predicted and Realized Detected Evasion Amount



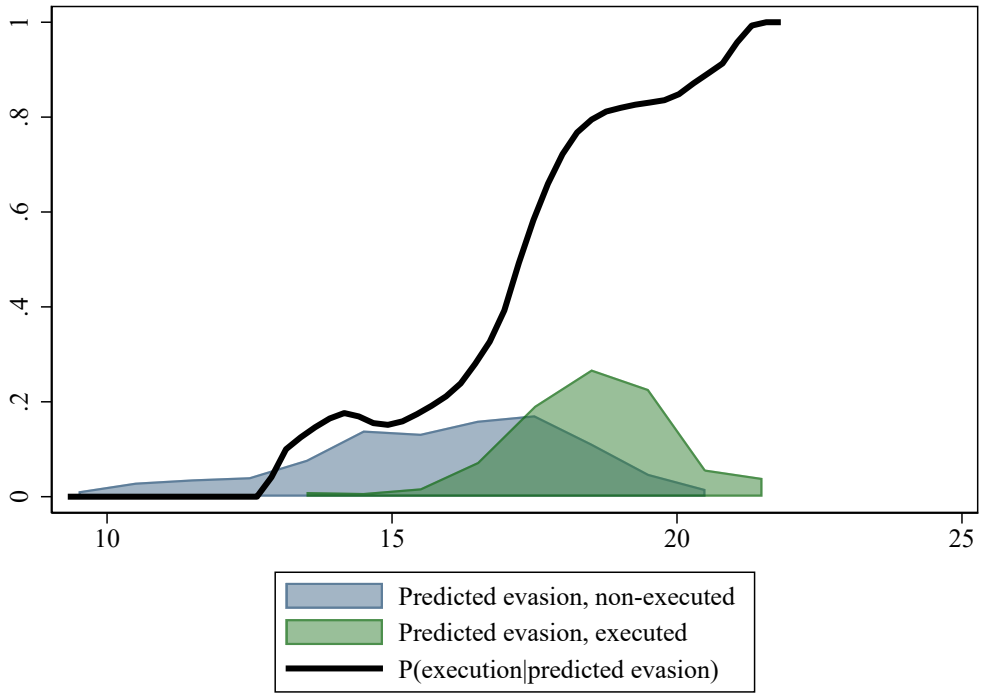
Notes: This figure is identical to Figure 4 but focuses on desk audits rather than full audits.

Figure F.2: Predicted and Realized Detected Evasion Amount

(a) Only Algorithm Cases

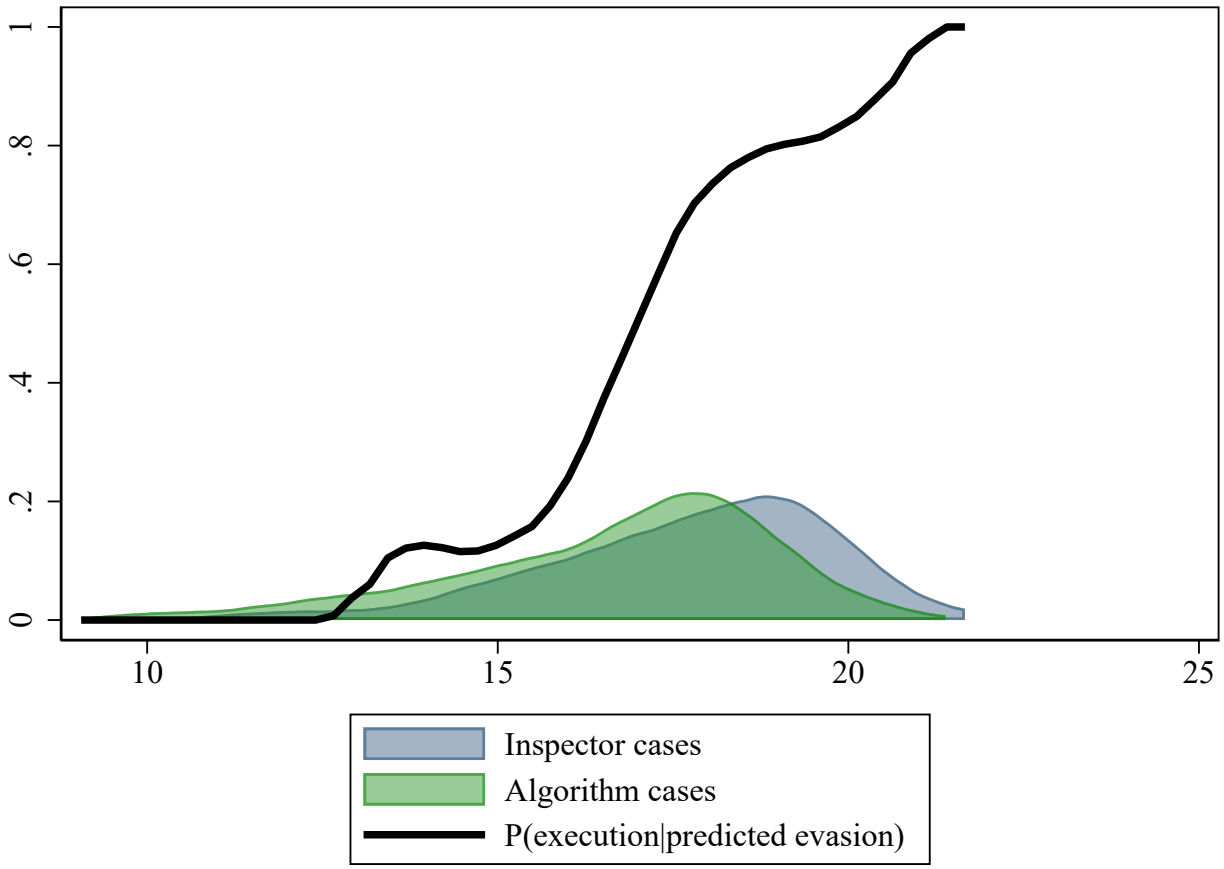


(b) Only Inspector Cases



Notes: This figure is the same as Figure 4, Panel B, but distinguishes between audits selected by the algorithm (panel a) and audits selected by inspectors (panel b). This figure is discussed in Section 6.4.

Figure F.3: Predicted Detectable Evasion Amount for All Cases in the Lists



Notes: This figure plots the distribution of detectable evasion for all full audits (log of FCFA amounts) estimated with a random forest model trained on the executed cases. The graph plots the distributions of the predicted amounts for the algorithm and the inspectors' lists. The black solid line plots the distribution of the realized audit execution rate by level of predicted evasion. This figure is discussed in Section 6.4.

Table F.5: Predictive Coefficients of Execution for Algorithm and Inspector Cases

	Algorithm	Inspector
	(1)	(2)
L1 turnover	-0.19**	-0.21***
	(0.07)	(0.07)
L1 turnover sq.	0.02**	0.02***
	(0.01)	(0.01)
L1 turnover cu.	-0.00**	-0.00***
	(0.00)	(0.00)
L1 profit rate	-0.08	0.09
	(0.20)	(0.16)
L1 profit rate sq	-0.07	-0.00
	(0.12)	(0.12)
L1 profit rate cu	-0.02	-0.12
	(0.24)	(0.21)
L1 productivity	0.05	0.04
	(0.05)	(0.05)
L1 productivity sq	-0.01	-0.01
	(0.01)	(0.01)
L1 productivity cu	0.00	0.00
	(0.00)	(0.00)
Distance km	-0.12	0.62***
	(0.30)	(0.23)
Distance min	0.04	-0.27***
	(0.13)	(0.10)
Firm's age	0.00	0.02
	(0.02)	(0.02)
N	407.00	537.00
R2	0.35	0.38

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows the correlation of firm characteristics and the execution rate of the full audit in two different OLS regressions. Column 1 shows the relationship between firm characteristics and the execution of audit cases selected by the algorithm. Column 2 shows the same for cases selected by the inspectors. The regressions include year and tax office fixed effects. Robust standard errors are shown in parentheses (Huber-White formula). This table is discussed in Section 6.3.

Table F.6: Prediction Model Using Ranking of Turnover Only

	Full Audits			Desk Audits		
	All executed	Algorithm	Inspectors	All executed	Algorithm	Inspectors
DGE	69	58	73	43	62	48
CME1	86	82	88	65	66	70
CME2	82	79	81	73	71	80
CPR	73	57	88	44	43	53
DSF	54	53	63	62	68	61

Notes: This table shows the percentage of audited cases that are correctly predicted by a simple turnover ranking within tax office and year, using declared turnover of the year prior to selection. The columns show the percentages of the N executed cases within each office that are also among the N largest cases in the selection list. For the algorithm and inspectors' columns, N is defined as the number of executed cases among algorithm-selected and inspector-selected cases. This table is discussed in Section 6.3.

G Additional Results on Machine-Learning Algorithm

Table G.1: Optimization Gains from Machine-Learning Algorithm Trained on Audit Outcome Data, Results Disaggregated by Tax Office, for Full Audits

Panel A: Large Taxpayer Office			
(1)	(2)	(3)	(4)
Realized Revenue Log(mean)	Predicted Revenue Log(mean)	Δ Revenue vs Predicted w/ RF Selection Among Program Cases	Δ Revenue vs Predicted w/ Size-Ranking Among Program Cases
21.89	21.81	22.13	8.53
Panel B: Medium Taxpayer Office			
(1)	(2)	(3)	(4)
Realized Revenue Log(mean)	Predicted Revenue Log(mean)	Δ Revenue vs Predicted w/ RF Selection Among Program Cases	Δ Revenue vs Predicted w/ Size-Ranking Among Program Cases
19.78	19.82	15.51	4.21
Panel C: Liberal Professions Office			
(1)	(2)	(3)	(4)
Realized Revenue Log(mean)	Predicted Revenue Log(mean)	Δ Revenue vs Predicted w/ RF Selection Among Program Cases	Δ Revenue vs Predicted w/ Size-Ranking Among Program Cases
20.66	20.53	18.94	-6.4
Panel D: Small Taxpayer Office			
(1)	(2)	(3)	(4)
Realized Revenue Log(mean)	Predicted Revenue Log(mean)	Δ Revenue vs Predicted w/ RF Selection Among Program Cases	Δ Revenue vs Predicted w/ Size-Ranking Among Program Cases
18.31	18.38	117.77	31.07

Notes: This table is similar to Table 10, Panel A, but shows the results disaggregated by tax office.

Table G.2: Performance of Random Forest Prediction

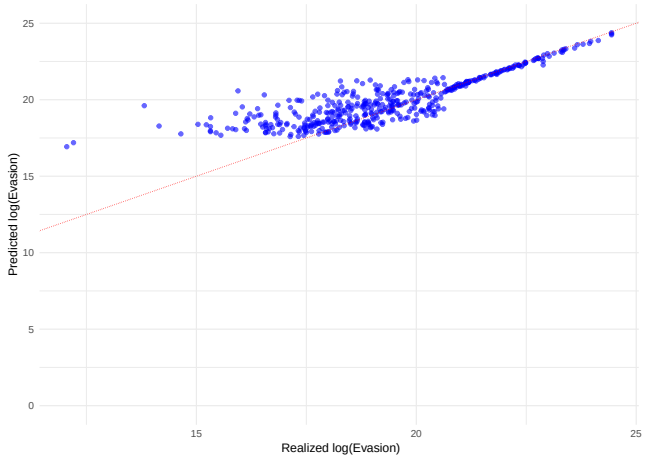
	Panel A: Full Audits			
	A1: Predicting Detection | Execution		A2: Predicting Evasion Amount | Detection	
	(1)	(2)	(3)	(4)
	Precision	Recall	MSPE	R2
Overall OOS	0.88	1.00	2.28	0.33
Overall OOB	0.89	1.00	3.28	0.25
2018	0.94	1.00	2.39	0.40
2019	0.78	1.00	2.72	0.18
2020	0.94	1.00	1.70	-0.01
	Panel B: Desk Audits			
	B1: Predicting Detection | Execution		B2: Predicting Evasion Amount | Detection	
	(1)	(2)	(3)	(4)
	Precision	Recall	MSPE	R2
Overall OOS	0.75	0.96	2.96	0.22
Overall OOB	0.78	0.93	2.38	0.21
2018	0.95	1.00	3.51	0.22
2019	0.58	0.95	2.92	0.12
2020	0.97	0.93	2.57	0.05

Notes: This table documents the performance of our random forest algorithm for predicting detection of evasion conditional on audit execution (panels A1 and B1) and predicting the amount of evasion conditional on detection (panels A2 and B2). The performance metrics are either calculated out-of-sample (i.e. in the hold-out/testing sample, rows 1 and 3-5 in each panel), or out-of-bag (row 2). The precision rate (column 1) is number of true positives over the sum of true positives and false positives. The recall rate (column 2) is the number of true positives over the sum of true positives and false negatives. A true positive is a case with detected evasion that is predicted to detect evasion. A false positive is a case with predicted evasion in which no evasion was detected in reality. The mean squared prediction error (MSPE, column 3) is the average squared difference between the predicted and observed values. The out-of-sample R2 (column 4) is ratio of the variance explained to the total variance in the outcome. Figure G.1 further evaluates the accuracy of predicted evasion compared to realized evasion. This table is mentioned in Section 7.

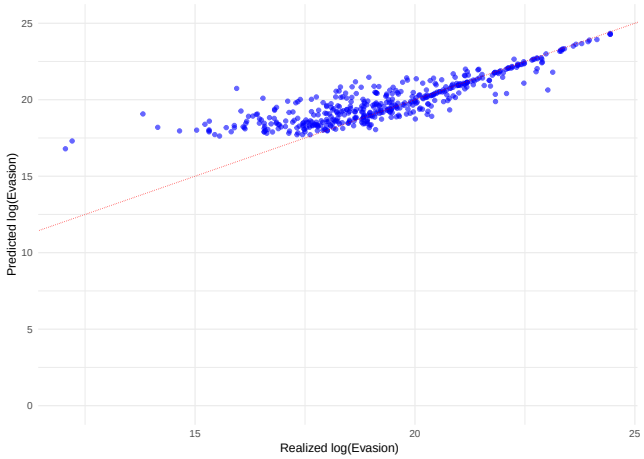
Figure G.1: Within-Sample Model Fit: Comparison of Predicted and Realized Evasion

Panel A: Full Audits

A1: Top Quartile Weights

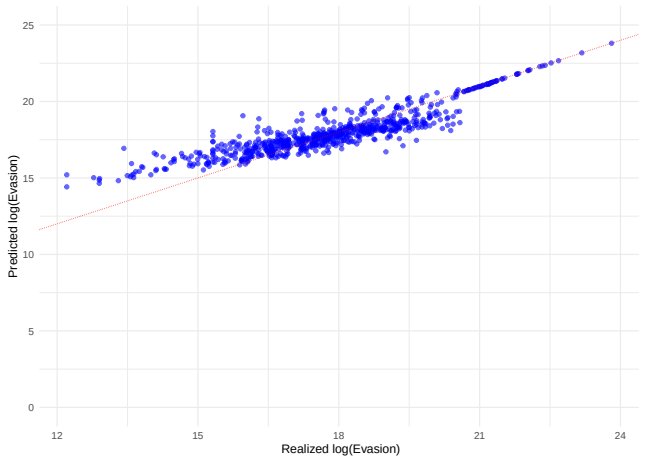


A2: Top Quartile Weights Within Office

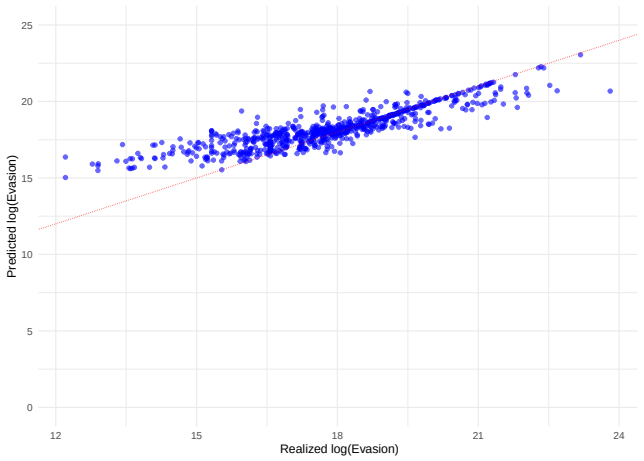


Panel B: Desk Audits

B1: Top Quartile Weights



B2: Top Quartile Weights Within Office

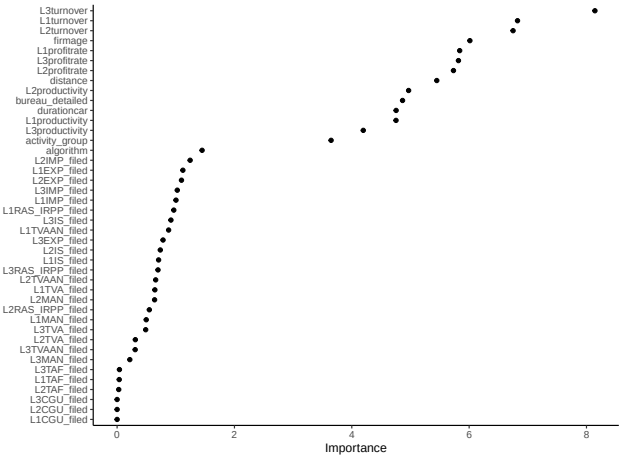


Notes: This figure shows scatterplots of the predicted and realized evasion, for full audits (panel A) and for desk audits (panel B), conditional on predicted evasion being non-zero. In each panel, we show results for a model where we increase the weight ten-fold on all cases in the top quartile of realized evasion (left column), or on all cases in the top quartile of realized evasion within each tax office (right column). Table G.2 shows other performance metrics of the prediction. This figure is mentioned in Section 7.

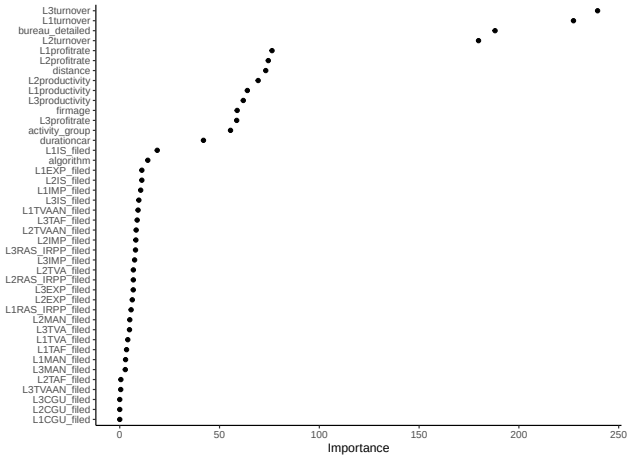
Figure G.2: Importance Ranking of Predictors in Random Forest Algorithm

Panel A: Full Audits

A1: Predicting Detection

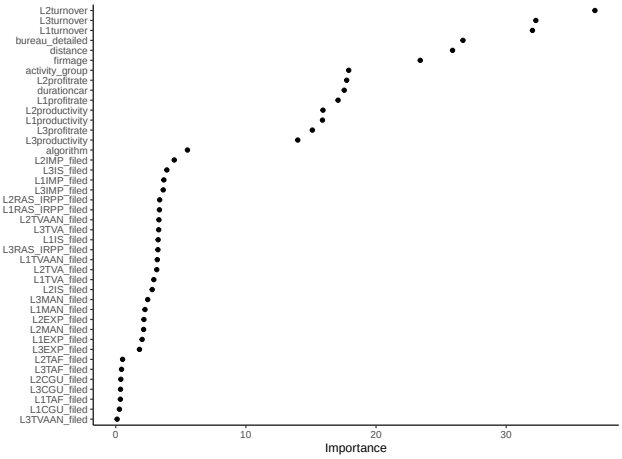


A2: Predicting Evasion Amount | Detection

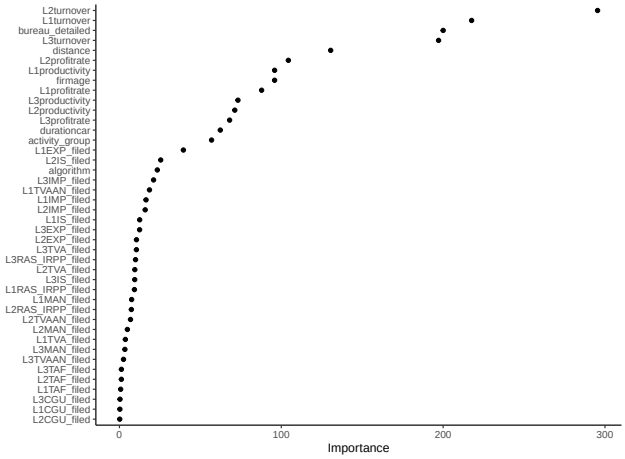


Panel B: Desk Audits

B1: Predicting Detection



B2: Predicting Evasion Amount | Detection



Notes: This figure shows the importance ranking of predictors in the random forest algorithm used to predicted whether or not an audit case detects any evasion (column 1), and for predicting the amount of evasion among the cases with non-zero evasion (column 2), for full audits (panel A) and desk audits (panel B). This figure is mentioned in Section 7.