Public Disclosure Authorized

Public Disclosure Authorized

Policy Research Working Paper 11205

Algorithms and Bureaucrats

Evidence from Tax Audit Selection in Senegal

Pierre Jean Bachas
Anne Brockmeyer
Alipio Ferreira
Bassirou Sarr

WORLD BANK GROUP
Development Economics
Development Research Group
September 2025

Reproducible Research Repository

A verified reproducibility package for this paper is available at http://reproducibility.worldbank.org, click **here** for direct access.

Policy Research Working Paper 11205

Abstract

Can algorithms enhance bureaucrats' work in developing countries? In data-poor environments, bureaucrats often exercise discretion over key decisions, such as audit selection. Exploiting newly digitized micro-data, this study conducted an at-scale field experiment whereby half of Senegal's annual audit program was selected by tax inspectors and the other half by a transparent risk-scoring algorithm. The algorithm-selected audits were 17 percentage points less likely to be conducted, detected 47% less evasion, were less cost-effective, and did not reduce corruption. Moreover, even a machine-learning algorithm would only have moderately raised detected evasion. These results are consistent with bureaucrats' expertise, the task complexity, and inherent data limitations.

This paper is a product of the Development Research Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at pbachas@worldbank.org and abrockmeyer@worldbank.org. A verified reproducibility package for this paper is available at http://reproducibility.worldbank.org, click **here** for direct access.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Produced by the Research Support Team

Algorithms and Bureaucrats: Evidence from Tax Audit Selection in Senegal

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

JEL codes: O17, H83, H26, D73
Keywords: Audits, Algorithms, Discretion, Risk-score, Tax Administration, Tax Evasion

*Pierre Bachas: World Bank Research, 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, Léo Czajka, Lucie Gadenne, Stephen Hansen, Janet Jiang, Nicola Limodio, Jan Loeprick, Markus Kitzmuller, Justine Knebelmann, Imran Rasul, Dan Rogger, Eduardo Souza-Rodrigues, Gabriel Zucman for helpful comments and discussions, and seminar 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 Economics Department, PUC Rio, University of Muenster, University of New Mexico, Norwegian School of Economics, NTA Conference, North Texas Economics Conference, RIDGE Public Economics, and Universidad del Pacífico. We thank Senegal's 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 and Rafael Vilarouca for excellent research assistance, the Paris School of Economics and CEPREMAP for administrative support. The authors are grateful for financial support from the Knowledge for Change (KCP) Program, administered by the World Bank, and currently funded by The Swedish International Development Cooperation Agency (SIDA), Agence Française de Développement (AFD) - French Development Agency, the Government of Japan, and the European Union; the research support budget of the World Bank's research group; UKAID via its Economic and Development Institutions call for proposal; the Centre for Tax Analysis in Developing Countries (TaxDev) at the Institute for Fiscal Studies; and the UKRI through Brockmeyer's Future Leaders Fellowship (grant reference MR/V025058/1). The findings, interpretations, and conclusions do not represent the views of the World Bank, its affiliated organizations, its Executive Directors, or the governments they represent, nor of the Government of Senegal.

1 Introduction

Lower-income countries are often data-poor environments, where policy decisions are taken in a discretionary rather than data-driven manner. For instance, individual bureaucrats may decide which taxpayers to audit, which water treatment facilities to inspect, and which manufacturing plants to monitor for pollution (Khwaja et al., 2011; OECD, 2023). In high-income countries, such decisions are often based on rules and data-driven, with limited input from individual agents. Data-driven decision models leverage available information in a systematic manner but require high-quality data. Discretionary decision-making, on the other hand, leverages bureaucrats' private information and experience but is vulnerable to bias. Although high-income countries commonly use data-driven methods for policy decisions, there is limited causal evidence on the conditions under which developing countries can benefit from replacing discretionary with data-driven decision making.

This paper examines whether a data-driven algorithm, applied at scale, can improve tax audit selection compared to discretionary case selection by tax inspectors. Our setting is the yearly tax audit program in Senegal, which aims to target firms with high amounts and high rates of tax evasion. Three features make Senegal an especially relevant context for this analysis. First, tax inspectors have traditionally exercised broad discretion in audit selection. Second, Senegal collected only 16.4% of its GDP in taxes as of 2018 (our base year), with tax evasion estimated to exceed 30% of true liabilities– suggesting large potential to improve tax compliance through audits. Third, recent investments in data infrastructure, including mandatory electronic tax returns and the digitization of third-party data sources, have greatly enhanced the information available for risk-based audit selection. Senegal's experience in leveraging these new data tools for enforcement is broadly relevant to other developing countries where similar tax digitization efforts are taking place.

In partnership with the Senegalese tax authority, we developed a *risk-scoring* algorithm to identify firms with high evasion amounts and evasion rates. This initiative marks the first systematic effort to leverage all available data for tax enforcement at scale in Senegal. The algorithm draws on taxpayers' self-assessment declarations across various taxes and on third-party records from customs, public procurement, and firm-to-firm transaction reports. Its design was subject to two key constraints: the selection criteria had to be simple and transparent to communicate, and it could not incorporate historical audit data, which were unavailable in digital format at the onset of the project. As a result, the risk-score algorithm is not a trained machine-learning algorithm, but rather built on intuitive compliance risk concepts and best practices, exploiting the data as thoroughly as possible.

For each firm, the algorithm calculates a risk score composed of multiple indicators designed 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 discussions with experienced Senegalese tax inspectors and foreign expert advisors, e.g., from the International Monetary Fund (IMF) and from other tax administrations.

The risk-score algorithm for audit selection was implemented at scale across tax audit offices in Senegal from 2018 to 2020. Each January, the tax administration approves a list of firms for each office that should be audited during the year. To experimentally assess the impact of data-driven selection on the administration's productivity, each list was split in two: half of the cases were selected at the discretion of tax inspectors, 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 based of how they were selected. By design, inspectors knew how a case was selected. While overlaps between inspector- and algorithm-selected cases were possible, they remained rare in practice.

To quantify the relationship between the selection method and audit outcomes, we compare the mean outcomes of algorithm-selected and inspector-selected audits within each audit list (i.e. tax office by year). The outcome measures stem from three new data sources. First, we digitized the universe of audit results records, which report the audit start and end date, inspectors working on the case, uncovered infractions, and fines to pay. Second, we surveyed a subset of audited firms to capture their perceptions of the audit process and corruption experiences. Finally, we surveyed inspectors to record their opinions and characteristics, such as skills and experience. These unique survey data allow us to examine how tax office managers deploy human resources based on the cases' selection methods.

We document five sets of results. First, we examine audit outcomes. Despite the protocol requiring full program implementation, the execution rate of the annual audit program is only 53%. This adds an additional margin of selection: from the list of planned audit cases, inspectors in practice pick a subset for implementation. Inspectors are more likely to audit cases which they selected: the execution rate is 17 percentage points lower for algorithm-selected audits. This is partly because algorithm-selected and inspector-selected firms differ on several dimensions: algorithm-selected firms are smaller, older, and less profitable. Yet, even after controlling for firm characteristics, an audit execution gap of 14 percentage points between the two audit case types remains. Conditional on implementation, 90% of audits uncover tax evasion, and the detection rate does not vary between inspector and algorithm-selected cases. However, conditional on detection, inspector-selected cases yield 47% higher assessments of taxes evaded on average. The results are robust across subsamples and time periods.[1]

Second, we put the audit process under the microscope. Using detailed administrative data, we find that algorithm-selected audits use less manpower, are implemented faster by various measures, and are conducted by less experienced teams. The implementation of algorithm-selected audits is evaluated less positively by firms, as captured by their perceptions of auditors' professionalism, competency

[1] A silver lining for the use of the algorithm is that algorithm-selected firms are 13 percentage points less likely to dispute audit results, compared to a mean share of 84% of firms that successfully contest the amount of confirmed evasion. The lower dispute rate reduces but does not close the gap in confirmed evasion amounts between selection methods.

and efficiency. We do not detect any significant difference in the incidence of corruption, as reported by taxpayers. We can reject differences in corruption of over 0.2 standard deviation. 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 investigate the reasons behind the execution and detection gaps between inspector-selected and algorithm-selected audits. We rule out two pre-registered hypotheses for why inspectors might favor their own discretionary selection: that the risk score is not predictive of evasion, and that inspectors lack relevant information on the algorithm-selected audits. In fact, the risk score strongly predicts both audit implementation and detected evasion for inspector-selected cases, which confirms that it contains relevant information for audits. We also show via a cross-randomized information treatment in the desk audit program that providing inspectors with descriptions of risk flags and Excel spreadsheets with the firms' data does not increase the likelihood of audit execution.

Fourth, using data from all full audits conducted between 2018 and 2020, we train a random forest model to predict audit execution, conditional on selection method. The key predictors of execution differ between algorithm- and inspector-selected cases, suggesting that inspectors apply different heuristics depending on how a case was selected. For both case types, however, inspectors' choice of which audit cases to execute is based on the expected detectable evasion. To show this, we train another random forest model, now predicting evasion amounts (rather than audit execution), and using only the sample of executed audits. Applying the model to predict evasion for all programmed audits (including those not implemented), we find that execution rates strongly rise with predicted evasion. In addition, conditional on a level of predicted evasion, execution rates are similar between algorithm and inspector-selected cases. This suggests that inspectors do not penalize the algorithm selection per se but rather equate the expected marginal returns on audits across the two selection methods.

Finally, using the predicted evasion from the random forest model, we measure the potential revenue gains from optimal audit execution among selected cases. We rank firms within each tax office, and pick the top cases, such that the number of cases implemented by each office equals their realized number of audits: by executing the cases selected by the machine-learning algorithm, the aggregate evasion uncovered could have risen by 18%. These potential revenue gains are moderate, and smaller than the returns from successful machine-learning applications in other contexts.

Overall, the results highlight the difficulty of improving bureaucratic decision-making via algorithms in lower-income countries when bureaucrats are highly skilled and engage in a complex task. The simple risk-score algorithm, designed without recourse to comprehensive audit outcome data, falls behind the performance of inspector-selected tax audits, but even the machine-learning algorithm might only improve audit performance moderately.

The risk-scoring selection of firms performed poorly despite our best efforts to draw on international

best-practice, our use of all available digital data, and the fine-tuning of the risk-score in three subsequent years based on the lessons learned; and despite the fact that inspectors ultimately chose which program cases to execute, de-facto adding their skill to the risk-scoring selection. Given the close collaboration with and endorsement from senior management at the tax administration, we believe that the results are unlikely to be driven by inspectors' desire to derail the algorithm. In fact, inspectors reported being highly motivated to use an algorithm for audit selection at the start of the program: in the 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 their bonuses are proportional to the uncovered evasion.

Inspectors might, in fact, outperform the risk-score algorithm in identifying high-evasion cases. Tax inspectors are among the top civil servants, highly trained, experienced, and incentivized.[2] Meanwhile, the risk-scoring algorithm relies on imperfect data: match rates between datasets are incomplete, key variables contain measurement errors, and evasion data was not systematically digitized before the experiment. In addition, although the intervention is implemented at scale in a medium-sized country, only 500 full audits were conducted between 2018 and 2020. Thus the training data for a machine-learning algorithm is small, while the number of predictors is large, given the set of variables in administrative tax declarations. Besides, in light of the small number of audits and the high stakes involved, randomly selecting audits to generate unbiased training data is impractical. This implies that the training dataset is selected, generating uncertainty about out-of-sample extrapolation.

As with any experiment, our design choices impose limits to the scope of our analysis. The empirical approach aimed to isolate differences in audit outcomes between algorithm and inspector-selected cases, conditional on any factor that affects implementation. However, our design does not directly address optimal audit selection, which requires considering not only the evasion detected through audits but also the medium-term effects on audited firms and potential deterrence impacts on non-audited firms. To explore the latter, we provide a back-of-the-envelope calculation of the size-dependent deterrence effects necessary for the risk-score algorithm to generate higher ex-ante compliance than inspector selection. We assume that firms know their size-dependent audit probabilities and that the only difference between selection methods is a reshuffling of audits along the firm size distribution, with the algorithm selecting smaller firms. Given the high concentration of revenue among the largest firms and the inspectors' focus on auditing them, we find that the inspector audit schedule yields larger aggregate declarations unless the deterrence elasticity is vanishingly small for larger firms.

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

[2] Anecdotally, as of 2025, both the current president and prime minister of Senegal are former tax inspectors.

1.1 Related Literature

Our work contributes to the rapidly growing literature on how algorithms can complement or substitute for human decision-making (see Table G.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 mainly from high-income countries. Our contribution is to test the performance of an algorithm in a low-income country context, at scale, and on a complex task. As in previous studies, the agents in our context (tax inspectors) are white-collar workers, but the skill difference between them and the other workers in the economy seems larger than in previous studies. This might explain why the agents in our setting perform so well compared to the algorithm.

In the context of tax audit selection, 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 available digitally and these countries can afford to have a share of their audit program selected randomly. In lower-income settings, in contrast, audit outcomes are often not digitized, the number of audits is smaller, and audits are rarely randomized.[4] In contrast to applications focused on machine learning, we hence test the performance of a risk-score algorithm as a potentially more suitable tool for lower-income 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 temporarily hired workers 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 to allocate cash transfers to the poor in Pakistan improved targeting and welfare. The decision-making agents were 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 literature that studies how the targeting of inspections and the allocation of inspectors to cases affect compliance. Duflo et al. (2018) show that the discretionary

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

[4] 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 audits in Rwanda. For audits of individuals, the welfare impacts can be large, especially at the top of the income distribution (Boning et al., 2023). This finding connects with the evidence from Senegal, where inspectors have a preference for auditing large firms and maximizing detected evasion amounts rather than detected evasion rates.

selection of audits lowers pollution offenses by more than random targeting, as random inspections found fewer extreme polluters than discretionary inspections. Building on this finding, we examine whether data-driven selection can further improve upon discretionary targeting. Furthermore, Duflo et al. (2013) shows that the elimination of discretion may reduce collusion between firms and auditors, who may hide detectable offenses in their reports. Reliance on an algorithm may hence influence audit outcomes by removing this potential conflict of interest.[5] Our study is the first to introduce a new selection method at scale in an experimental way in an enforcement context, allowing to isolate the effect of case selection on a bureaucracy's performance isolating other confounding aspects.

Finally, our work contributes to 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).[6] Another strand of this literature has focused on the power of tax administration (Basri et al., 2021) and 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, 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. Tax revenue as a share of GDP was 16.4% in 2018, as in many 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 has been 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 Value

[5] We also connect to 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).

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

Added Tax (VAT), and 26% of the Corporate Income Tax (CIT) (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 of a lower-middle-income country (see Table A.1). The Value Added Tax is the largest source of revenue, representing slightly more than 29% of total tax revenue in 2022, followed by corporate and personal 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 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 under 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 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 A.1). There are two types of audits: full audits, carried out by a team of inspectors at the taxpayer's premises; and desk audits, conducted by correspondence by an individual inspector using the firm's tax returns and third-party reported data. All audits are targeted at firms.

At the beginning of each calendar year, audit offices prepare a list of taxpayers to audit based on a target provided by the tax administration. This list is reviewed and approved by senior management, usually without any changes. Before our intervention, the selection of cases for tax audits was discretionary, meaning it did not follow an explicit rule and no rationale was required for desk audit selection. Although sometimes offices requested justification for the selection of full audits, such as a review of the firm's audit history or a summary of relevant indicators such as total sales and the profit margin, such justifications could take different forms and were not systematic. Thus, the criteria used for case selection varied across units.

Figure A.2 outlines the steps of the audit process. After examining a case, inspectors list the detected irregularities and corresponding penalties, notifying the taxpayer via an initial notification. They may also request additional information. Upon receiving the notification, taxpayers have 30 days to respond. Failure to respond is considered acceptance of the inspector's findings. The inspector then reviews the response and has 60 days to issued a final confirmation with the irregularities, penalties, and the final amount due. A revenue order is then generated for the tax collection unit, requiring the taxpayer to make 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 twice as many.[7] The PAYE data provides 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 recorded in VAT annexes that firms file digitally 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

[7]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.

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, audit risk, and their view 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 of 1226 targeted firms. For the second survey wave, we targeted an additional 702 firms, all selected by stratified random sampling. 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).

2.4 Tax Inspectors

The key bureaucrats in the tax enforcement process are the tax inspectors and their managers, who are usually former tax inspectors. 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 doctorate degree. 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 are 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 A.3 discusses the bonuses and performance evaluation in more detail.

According to the survey, when selecting cases for audits, inspectors typically start from a long list of cases to consider, roughly half of which they chose with no particular suspicion. Inspectors then examine the long list of cases to derive a short list. Figure 1, Panel A, shows inspectors' reported objectives in audit selection. The most common objectives were a diversity of audits (i.e. create the perception of a non-zero audit probability for most taxpayers, which we interpret as an interest in the

deterrence function of audits (discussed in Section 5.4), 2) the detected evasion amount, and 3) the detected evasion rate. 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. Panel C shows that there is a clear ranking in the types of information used in audit selection. The most commonly used data sources are the self-assessed declarations (over 90% of all cases), followed by third-party data (70% of cases), with soft information being a distant third (around 40% of cases). Over 60% of inspectors report finding quantitative information more useful than soft information, while 22% prefer soft information, and the rest consider the two sources as equally important. Most inspectors use Excel 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.

Inspectors are critical of the status quo in audit selection and keen to improve. 55% of inspectors 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 and 70% support an automated selection based on risk indicators.[8]

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.

3 Experimental Design

3.1 Intervention: Risk-Based Audit Selection

The experiment was conceived at the request of the Senegalese tax administration to implement a more systematic data-driven audit selection method. We collaborated with the administration to design a risk-scoring algorithm for audit selection and to deploy it in a controlled manner, thus allowing for a rigorous evaluation of its impact. 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 evasion rates. The study was pre-registered in the AEA RCT Registry (Bachas et al., 2019). In the pre-analysis plan, we describe the algorithm, the outcomes of interest, the empirical design, and the intervention treatment discussed below. Our analysis closely follows the pre-analysis plan and deviations are presented in appendix B.5.

We faced two types of constraints when designing the algorithm. First, the algorithm needed to

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

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. Accessing and systematically digitizing all paper records of audit results, which were held in individual folders in each tax office, came in a variety of formats, and were considered sensitive information by the administration, took several years and a large investment of time and resources by the research team. Given these constraints, we designed a "best-practice" risk score 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.

The algorithm generates a risk score for each firm, allowing us to rank firms within a tax office. The risk score combines two classes of risk indicators: 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 declared by other firms. In contrast, anomaly indicators are *across-firm indicators* that flag outlying behavior 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. Another indicator flags firms trading with low tax jurisdictions (i.e., tax havens). Each inconsistency and anomaly is captured in the form of a ratio. We assign points from 0 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 by indicator. We weighted inconsistencies higher than anomalies as inconsistencies were more likely to reflect non-compliance, while anomalies could reflect temporary economic difficulties or poor management. Appendix B 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 the 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.[9] 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, due to time constraints, inspectors enjoyed discretion to pick which firms on the annual program list to audit. We discuss this incomplete execution 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. Inspectors were provided with a protocol emphasizing 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 digitize audit 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 B.6 and B.7 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 (and sometimes by pairs of inspectors), so the inspector is assigned at the selection stage, unlike in full audits. We took advantage of these features to introduce two additional treatments in desk audit lists: the inclusion of random audits, and a cross-randomized information treatment.

First, in 2018 and 2019, we added randomly selected cases to each inspector's list (drawn from the pool of firms of their respective tax office). Both randomly and algorithm-selected cases appeared in the lists as having been selected by the "new method," so inspectors were unaware that these cases were randomly selected. This approach avoids the possibility that inspectors ignore randomly selected cases, deeming them low-return.

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 inspec-

[9] Although our project ran from 2018 to 2020, DGID still asked us to generate the algorithm selection in 2021 and 2022.

tors 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

Our empirical exercise compares algorithm-selected cases' outcomes to 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 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.[10]

3.5 Descriptive Statistics of Firms Selected for Audits

Table 3 examines the characteristics of firms selected for a full audit. Across all tax offices, about 150 firms per year were selected by inspectors for a full audit, and a similar number was selected by the algorithm.[11] Panel A compares the characteristics of firms selected by inspectors to the general population of firms, pooling the years 2018-2020. Inspector-selected firms are orders of magnitude larger than the average firm appearing in the administrative tax data (columns 1 and 2), more likely to engage in import or export (column 4), and older (column 6). Inspector-selected firms do not significantly differ from other firms in their location (columns 5).

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 much larger than the average formal firm (columns 1-2).[12] They also have lower profit rates than inspector-selected firms but are more likely to engage in international trade (columns 3 and 4) and tend to be older (column 6). 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 hard to detect and document tax evasion on cash sales.

4 Audit Outcomes

We now examine the association between algorithm selection and audit outcomes. 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 analyze evasion detection, evasion amount, and taxpayer disputes. Finally, we examine the robustness of our results to potential sources of biases.

4.1 Audit Execution

Based on the inspectors' audit reports, we can observe which cases on the list were executed, that is, cases that were started, even if they did not lead to a fine. Overall, only 53% of cases selected for full

[10] Taxpayers were not informed of 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.
[11] Table B.10 shows the characteristics of firms selected for desk audits.
[12] Figure B.1 illustrates the gradient between firm size and audit selection within tax offices.

audits and 33% of cases selected for desk audits are actually audited. The execution rate improves slightly in 2019 from 2018, but then worsens in 2020, likely due to disruptions from the COVID-19 pandemic.[13] For this analysis, we consider all audits reported until the end of 2021, counting audits that were executed in the following fiscal year relative to their selection year.

Table 4, columns 1-3, shows how the execution of algorithm-selected cases compares to inspector-selected cases. Full audit cases selected by the algorithm are 17 percentage points less likely to be implemented than inspector-selected cases, corresponding to a 32% reduction compared to the mean implementation rate (column 1). "Overlap" cases, which were selected by both the inspectors and the algorithm (but which appeared as "Inspector selected" on the lists) were 16 percentage points more likely to be implemented than cases selected only by inspectors.[14]

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), because the treatment took place within every individual list of cases. Overlap cases are again more likely to be implemented, though the coefficient is not statistically significant. Random cases (only included in desk audits lists) 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").

4.2 Audit Yield

Table 4, columns 4-6, show the relationship between the selection method and the probability of evasion detection, conditional on execution.[15] Overall, 89% of executed full audits and 73% of desk audits detect evasion. For full audits (column 4), algorithm-selected cases are 4 percentage points more likely to detect evasion, and overlap cases 8 percentage points more likely, though only the coefficient on overlap cases is economically meaningful and statistically significant at the 10% level.

For desk audits (column 6), algorithm-selected cases are 4 percentage points less likely to detect

[13] As shown in Section 4.4, execution rates are higher in 2019, when the project team invested more resources in monitoring implementation and adherence to the experimental protocol, and lower in 2020, when the COVID-19 pandemic disrupted audit execution. Despite the incomplete implementation of the 2018 audit program, the tax administration emphasized their goal to fully implement each annual program. As a result, in 2019, the number of audit cases was lowered by a third for the large taxpayer office, and by 20% for the medium taxpayer offices (Table B.6).

[14] 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 inspectors, but 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.

[15] Intent-to-treat analyses that do not condition on audit implementation are shown in Table C.1. In these regressions, evasion detection and evasion amounts are zero for non-executed cases, which reduces the coefficient on "algorithm."

evasion, and overlap cases are 8 percentage points less likely to do so, although neither of these differences is statistically significant. The coefficient on random cases is zero, 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 to detect evasion.

Columns 7-9 show the results for the amount of detected evasion, conditional on positive detection. Algorithm-selected full audits detect significantly smaller amounts of tax evasion plus fines, though this result is not significant for desk audits once we control for inspector fixed effects (column 9).[16]

Since the analysis in columns 4-6 and 7-9 is conditional on audit execution and detection of evasion, the results are affected by non-random attrition: it is unclear what the results would look like if all planned audits had been conducted. To examine whether our results are affected by sample selection, we conduct a sensitivity analysis by computing upper and lower bounds on these effects as proposed by Lee (2009). This analysis assesses scenarios by eliminating the highest or lowest values in the group with fewer missing observations (algorithm or inspector-selected cases). 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 (upper) bound. When the outcome is 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 list-specific attrition rates. The estimated bounds are noted in Table 4 within square brackets below the main point estimate on algorithm selection. These 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, algorithm-selected audits are less likely to be executed but are similarly likely to detect evasion when implemented. Conditional on detection, algorithm audits uncover smaller amounts of evasion. Results are starker for full audits compared to desk audits, which is consistent with the fact that full audits involve higher stakes for both the administration and the taxpayer.

4.3 Dispute of Audit Outcomes

Once an audit concludes with the 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, possibly indicating errors in how the audit was conducted, uncertainty over the exact amount evaded, or collusive behavior between taxpayers and inspectors to alleviate the penalties.

[16] 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 (results in Table C.4).

Rules-based 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).[17] 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 shares are 40% and 18%, respectively. For desk audits, these figures are 62% and 58%, indicating slightly less push-back against the notification amount. For full audits, the figure suggests that inspector-selected cases are subject to more push-back against the notification amount than algorithm-selected cases.

Table C.5 shows these results in regression format. Among full audits for which both a confirmation and a notification are present in our data, algorithm audits are 12 percentage points more likely to have matching confirmation and notification amounts. 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 evasion amount than inspector-selected audits (columns 3 and 5).[18] Although the lower likelihood of disputes for algorithm audits slightly improves the relative return of algorithm-selected audits, it does not fully close the gap in detected evasion between algorithm and inspector-selected audits.

4.4 Robustness

We assess the robustness of our results to discrepancies in the number of algorithm versus inspector-selected cases within audit lists, and to particular periods and tax offices.

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

[17] The figure excludes cases with no confirmation, which represent about a third of these cases.

[18] The result is similar when we control for firm size. For desk audits, there are no significant differences in dispute of outcomes between inspector and algorithm-selected audits. The point estimates in this section are slightly different from those in Table 4, which does not impose the requirement that all audit cases have both a notification and a confirmation.

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.[19]

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. Moreover, firms in the Large Taxpayer Unit are audited roughly every four to five years, a much higher rate than any other office. Therefore, the association between audit selection method and outcomes may 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 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, equally likely to detect evasion (though point estimates are always positive), and conditional on detection 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 due to 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.

[19]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.

5 The Audit Process

5.1 Time and Human Resource Costs

We next examine how human resource costs and audit length vary by selection method among executed audits. Table 5, Panel A1, shows the results for the number of inspectors working on a case and for various proxies of audit length for full audits (Table D.1 shows desk audits). The most accurately captured measure of audit cost is the number of inspectors working on an audit, which is always 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 inspectors. In addition, teams for algorithm-selected audits have lower seniority, as measured by age and experience (Table D.2).[20]

Columns 2-4 of Table 5, Panel A1, show the association between audit selection and three different measures of audit length. Overall, all measures suggest that algorithm-selected full audits are shorter. In column 2, the outcome is the audit duration as reported by firms subject to a full audit from the taxpayer survey. The mean duration is 28 days, and algorithm-selected audits are approximately nine days shorter on average, which is an economically large difference. In column 3, we use 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 on the case, as they typically work on several cases simultaneously.[21] The average full audit takes 160 days until notification. Algorithm-selected audits are completed 30 days faster.

Finally, in column 4, 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. In this measure, full audits are completed in 177 inspector-days on average. Algorithm-selected audits take 27 inspector-days less, though the point estimate is not statistically significant. This is not surprising, given the reduced sample size and the presence of measurement error. The audit duration measures in columns 2-4 are not consistently available for all executed cases. For example, the self-reported number of days working on a case was only filed by a subset of inspectors. However, Panel A2 indicates that we can reject differential availability of outcome measures across case types for all but one outcome (column 4).[22]

In addition to the detected evasion amount, audit reports indicate the specific year of the infraction and

[20] These differences in team composition are consistent with the fact that, for desk audits, inspectors with above-median experience have lower execution rates and are less likely to favor the use of an algorithm for audit selection (Table D.3). Estimating inspector-quality fixed effects, we find substantial heterogeneity in inspector skills (Figure D.1), suggesting the algorithm may be more beneficial to low-skilled inspectors. We do not detect any correlation between overall list-level execution rates and algorithm execution rates (Figure D.2). However, we find that the number of inspectors on a (full) audit team mediates the association between algorithm selection and audit outcomes (Table D.4).

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

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

a short description.[23] 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 D.5, columns 1-4). All differences are statistically significant at conventional levels. There is little evidence of systematic differences in the fine to 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 D.5, columns 5-6).[24]

5.2 Audit Productivity

Given the 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 5, Panel B1, shows the results for our main specification in Equation 1, using audit productivity measures as outcomes (similar results for desk audits are shown in Table D.1). We consider the log of the ratio of uncovered evasion over different audit cost measures used in Panel A1: the number of inspectors working on a case (column 5), the audit length in number of days from the survey data (6), the number of days from opening to closing a case from the administrative data (7), and the self-reported number of days the inspectors worked on the case (8). The association between algorithm selection and audit productivity is negative in all estimations. 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. Together, these results imply 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.[25]

The averages captured by the regression coefficients in Table 5, Panel B1, mask heterogeneity across cases. Figure D.3 shows distributions of the three main audit productivity measures for algorithm and inspector-selected cases. The productivity distributions for inspector-selected cases are shifted slightly to the right for all measures. For our preferred measure of productivity–evasion amount divided by number of agents on the case–the two distributions are statistically significantly different as per the Kolmogorov-Smirnov test. Yet the two distributions overlap substantially, suggesting that many algorithm-selected cases have higher productivity than the marginal inspector-selected case.

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

[24] Table D.6 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.

[25] Panel B2 reports the availability of outcome measures across audit selection methods, showing small but significant negative point estimates in columns (7) and (8). The fact that less information is available for algorithm-selected cases is consistent with the idea that less time was spent on these cases (including less time on documenting what happened and generating the basic paper work, e.g. an audit announcement letter or audit closure report). Indeed, our results in Panel A, column (4) and Panel B, columns (7) and (8) would be even stronger if we replaced the resource and productivity outcomes by zero for audit cases that were implemented but do not have outcome data recorded.

This rationalizes the fact that execution of the inspector-selected set of audits is incomplete and algorithm audits are also implemented. Note that audit productivity is positively correlated with firm size, consistent with the preference for larger firms in audit selection and execution (Figure D.4).

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 audit 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 the perceived efficiency of audits and the perceived incidence of corruption in 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) a preferential 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% in 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 knowledge score of 7.3/10, but equal to their efficiency score.

Table 6 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.[26] Columns (1) to (3) show the 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 is no longer significant (Panel B). Firms selected by both inspectors and the algorithm reported higher evaluation grades for 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 perceived corruption. We don't

[26] Firms that were actually audited as per the administrative data account for a half of firms reporting an audit. 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.

find evidence of differential corruption perception across selection methods. For full audits, we can reject that the algorithm improved corruption 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. 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.[27] Yet, they do not lend support to algorithmically selected cases facing a different incidence of corruption.

5.4 Could Deterrence Effects Justify Using the Algorithm?

In addition to the productivity of audits, selection rules may also impact economy-wide compliance. To maximize tax revenue, the administration might consider how selection rules deter tax evasion. Deterrence implies that firms anticipating a higher audit probability may preemptively comply, even if they are not audited. Here, we examine whether the algorithm-based selection rule could foster greater ex-ante compliance compared to the discretionary selection.

Perceived audit probabilities are challenging to measure (Bergolo et al., 2023) and our experiment was not designed to estimate deterrence effects. We hence do not examine the revenue-maximizing audit policy. Instead, we focus on a simpler question: how strong would deterrence effects need to be for risk-score selection to enhance revenue collection? We analyze the potential deterrence effects of each selection method, and then calculate the counterfactual compliance (reported turnover) in scenarios where the audit program is either fully selected by the risk-score or fully selected by inspectors. We summarize our analysis below and detail it in Appendix E.

We focus on differences in audit probabilities by firm size (turnover), arguably the most important firm characteristic for audit selection. Risk-score selected firms are on average -1.78 log points smaller in turnover than inspector-selected firms (Table 3). Audit probabilities strongly rise with firm size for both selection methods, but the algorithm's schedule is less steep than the inspector's. Our analysis requires two strong assumptions: first, that firms know their size-specific audit probabilities, and second, that evasion rates by firm size in the firm population equate those observed among audited firms. Note that detected evasion rates decline with firm-size from a 50% evasion rate for the median-sized firm, to 20% for the largest firms (Figure E.2). We assume for simplicity that evasion takes the form of turnover under-reporting (as opposed to, e.g., cost under-reporting). The key parameter for our analysis is the deterrence elasticity: the response of a firm's evasion rate to its audit probability.

Considering different scenarios for the value of the deterrence elasticity, our analysis yields two insights. First, if the deterrence elasticity is uncorrelated with firm size, the inspector's schedule–which prioritizes larger firms–is always expected to yield higher tax revenue. This stems from the very high concentration of turnover at the top of the firm-size distribution. Second, if the deterrence elasticity

[27] 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).

is inversely related to firm size (i.e., smaller firms react more to audit threats than larger firms), then the algorithm selection could improve reporting behavior and revenue collection. Yet, even in this scenario, the potential reporting gain remains modest, peaking at 0.2% of baseline reported turnover, and occurs only when the firm-size dependence of the deterrence elasticity is very large.

To assess the plausibility of the latter scenario, we estimate evasion elasticities with respect to audit probability by leveraging differences in audit probabilities among firms of the same size, driven by their (conditionally quasi-random) assignment to different tax offices. We estimate substantial deterrence elasticities and size-dependence in elasticities. Yet, applying these estimates to generate counterfactual reporting of turnover, we still find that the inspector schedule produces larger deterrence effects (Figure E.3).[28] Thus, the ex-ante deterrence effects of audits do not seem to favor the algorithm selection. However, we acknowledge the uncertainty surrounding the magnitude of deterrence elasticities and that our study was not designed to measure these effects.

6 Mechanisms

The previous sections show that algorithm-selected audits were less likely to be executed, and, when implemented, detected lower evasion. We now examine the reasons why algorithm-selected cases were executed at a lower rate. We hypothesize that inspectors prioritize cases that they expect to yield a high amount of evasion, based on the firms' characteristics they observe.

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, providing additional information for algorithm audits does not raise execution rates, suggesting that inspectors' hesitance to implement algorithm cases does not stem from a lack of information. Third, observable characteristics can only explain a third of the execution gap between algorithm and inspector-selected audits. Instead, inspectors seem to use different criteria to decide whether to implement an inspector versus an algorithm-selected audit. Yet, for both case types, the execution rate strongly rises with the predicted detectable evasion. We conclude that algorithm-selected cases are executed at lower rates because inspectors predict that they yield lower average expected returns 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 the evasion rate. To test whether this is the case, Table 7 reruns our main estimation (Table 4, column 1),

[28] We estimate a semi-elasticity of -2: when the audit probability rises by 1 percentage point, the evasion rate falls by 2%. In this scenario, the deterrence effects under the inspector schedule raise reported turnover by 2% of baseline, compared to the current mixed selection. These deterrence effects may be another reason why inspectors prefer their selection.

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), 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 by construction in this sample and multiple risk indicators.

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).[29]

6.2 Providing Information does not Increase Audit Performance

Another factor that may determine inspectors' implementation of audits is their information about a case. Inspectors may be better informed about the cases they selected themselves and could have limited information about the compliance risks among algorithm-selected cases and, hence, struggle to implement algorithm-selected audits. To test this hypothesis, we conducted a randomized information treatment among desk audit cases. For a third of the cases, inspectors were provided with the risk flags from the algorithm. For another third of the cases, inspectors were provided with both the risk flags and 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 any additional information by the research team.[30]

To test the effect of the information treatment, we add treatment indicators to our main regression specification (Equation 1). We also add an interaction between the information treatment and the algorithm selection dummy, to test if 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 treatment arms into one, 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

[29] 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.

[30] Table B.9 shows that taxpayer characteristics were balanced across treatment groups. For full audits, inspectors were always provided with the risk flags for all cases.

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. Thus, a lack of information about algorithm-selected cases does not seem to contribute to explaining their lower execution rates.

6.3 Observable Characteristics do not Explain the Execution Gap

We now examine if the observable characteristics of algorithm-selected firms explain why they 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 firm characteristics as regressors. Table 8, Panel A shows the results. In the first column, we reproduce our main result on audit execution as a benchmark: algorithm-selected audits are 17 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 remains highly statistically significant and large at 12 percentage points.[31] 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 as additional regressors. Adding these regressors does not reduce the remaining audit execution gap and hardly changes the R2 of the regression.[32]

While our results indicate that a simple OLS model cannot explain the execution gap based on observable characteristics, it is possible that there are nonlinearities and interactions in the association between firm characteristics and audit execution. We hence turn to more flexible prediction models. As average execution rates vary 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 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:

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

[31] 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.

[32] Panels B and C in Table 8 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.

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 Figure F.1, for full audits in Panel A and for desk audits in Panel B. 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 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 (gray markers) are far higher than realized execution rates (gray 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.

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 turnover, 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, distance in meters ranks first among the predictors, followed by log turnover (with only 86% relative importance), lagged productivity (84%), duration of a trip in minutes (84%), and firm age (81%).[33]

[33]Table F.5 shows the point estimates on the predictors in a linear prediction model. 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.

6.4 Inspectors' Audit Execution Decisions are Strategic

Finally, we test the hypothesis that inspectors are strategic, i.e. they maximize detectable evasion when deciding which audits to implement. To do this, we train a random forest model to predict the amount of detected evasion (instead of the execution rate), in the sample of all cases that were audited. In addition to the predictors discussed in the previous section, we include a dummy indicating whether the case was selected by the algorithm. This means that we allow for observable characteristics to map into detectable evasion in different ways for algorithm-selected and inspector-selected cases.

We then use the model to predict the hypothetical detectable evasion amount for cases that inspectors didn't audit. Figure 4, Panel A, plots the results for full audits. For cases that were audited, the distribution of the predicted and realized evasion amounts are close, suggesting that the model yields reasonable predictions (we assess the quality of the prediction in the next section). The distribution of the expected detectable evasion for non-audited cases, however, is shifted leftwards, indicating that these cases have lower predicted potential on average.[34] The figure also shows that the realized execution rate is upward-sloping in predicted evasion amounts. Panel B shows that the level and slope of the realized execution rate in predicted evasion is very similar for algorithm and inspector-selected cases. Hence, inspectors do not prioritize inspector-selected cases, conditional on the amount of predicted evasion. The higher implementation rate of inspector-selected cases is because the distribution of detectable evasion is shifted out compared to the distribution of algorithm-selected cases.

We conclude that, regardless of the heuristics that inspectors apply to decide on audit execution, which seems different for inspector and algorithm cases, inspectors prioritize cases based on expected detectable evasion. The results suggest that, although the risk score has some power to predict audit outcomes, the inspectors are on average better than the risk-score in identifying high-evasion cases.

7 Performance of a Machine-Learning Algorithm

The analysis shows that our risk-scoring algorithm doesn't outperform inspectors' case selection. An open question is whether a machine learning algorithm, trained on audit outcome data to predict detectable evasion, could have improved audit performance. To answer this question, we lift the two constraints faced at the project's onset: that the audit selection mechanism needed to be transparent and simple, which led us to focus on best-practice risk flags, and that audit outcome data was not available for training purposes.

We leverage our previous random forest prediction of evasion and simulate the revenue potential of audits selected according to this machine-learning algorithm. Concretely, we first train a two-stage random forest to predict evasion on the sample of implemented audits. In step one, we predict

[34] The same holds also when considering inspector-selected audits and algorithm-selected audits separately (Figure F.3).

detection of any evasion; in step two, we predict evasion amounts, conditional on detection. We give tenfold higher weights to 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.[35] Second, we apply the resulting model to predict detectable evasion for all firms in the annual audit programs, i.e. those that could have faced an audit. Third, we calculate predicted aggregate revenue from several execution scenarios, holding the number of audits per office constant.

Table 9 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 shows the predicted aggregate revenue from the same audit program. This means that we sum up the predicted revenue from all implemented audit cases. The prediction is close to the realized aggregate revenue, but still slightly below. We hence compare the results of our simulations to this predicted revenue. Column 3 shows the potential revenue gain from a first simulation, 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 18.09%. Note that this is likely an overly optimistic estimate, as it could be achieved only if inspectors fully adhered to the proposed case selection. The experience of our intervention suggests that this is unlikely.[36]

Panel B shows that the revenue gains in both simulation scenarios are higher for desk audits than for full audits. This is consistent with inspectors investing less in desk audit selection and selecting part of their cases at random. For the risk-score algorithm, too, desk audit selection might be further away from the optimum. This is because firms with the highest risk score are selected for full audits; firms with medium risk scores would be selected for desk audits. Firms with relatively lower risk scores are likely a mix of firms that are truly less risky and firms with poor or missing data. Table G.2 shows the results 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 smaller in our context than in most other machine-learning applications (see Figure G.1 and Table G.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

[35] Alternative specifications of the weights do not achieve as close a match between realized and predicted aggregate revenue. After setting the weights, we tune the hyper-parameters (node size, number of features in split, number of trees) using semi-structured grid search, and choose parameters that maximize out-of-sample RMSE, which are also those that minimize out-of-sample R2. Table G.3 and Figure G.2 document the performance of our prediction models.

[36] 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, missing data would be a bigger problem in the population, and we would not be able to use the selection year as a predictor for evasion.

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.[37] 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 and firm names, which prevent matching across datasets (or lead to erroneous matches). Some data sources, such as the registry of contracts, the business-to-business payment and withholding records were digitized only after our intervention. Czajka (2023) shows the usefulness of these data for enforcement. It is likely that the inclusion of these data would strengthen the algorithm.

8 Conclusion

This paper examined whether an algorithm using newly digitized data improves tax audit outcomes in Senegal. Our results suggest that it is difficult to build a simple and transparent rule that outperforms tax inspectors in selecting high-evasion firms. Inspectors prioritize the audits they personally selected over those selected by a risk-scoring algorithm and their selection yields better results. 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. Even an ex-post machine-learning algorithm trained on outcome data might only have achieved a moderate improvement in aggregate evasion detection compared to the inspectors' selection.

Our findings contrast with a growing literature that demonstrates a large potential for rule based decisions 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 with strong financial incentives. The performance of tax inspectors in Senegal is consistent with evidence on the powerful role of specialized tax administrators in generating higher tax compliance (Basri et al., 2021). On the other hand, the data fed to our algorithm suffered from quality limitations, as is the case for most administrative datasets in lower-income countries. In addition, even in a mid-sized country such as Senegal, operating at scale, audit outcome data would have to be collected over an extended period to build a sufficiently large training dataset, especially given the vast number of potential predictors of evasion.

Our results also highlight the benefits of bureaucratic discretion and autonomy. In other contexts, discretion has limited the performance of algorithms, as agents may pursue different objectives than the

[37] 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.3 shows the importance ranking of predictors.

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 risk-scoring algorithm. The fact that overlap cases, selected by both inspectors and the risk-scoring algorithm performed better than the average inspector-selected case suggests that there is valuable information in the intersection of selection methods. How to best combine bureaucrats' skills and systematic data analysis is an interesting question for future research. In addition, given the demand for algorithms and the hopes placed in them, more work is needed to understand for which tasks and in which settings they can enhance the performance of organizations.

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. 22, 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

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

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. 30

Bachas, Pierre, Anne Brockmeyer, and Bassirou Sarr (2019) "Audit Selection Under Weak Fiscal Capacity.," *AEA RCT Registry, https://doi.org/10.1257/rct.2488-2.0*. 11

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

Basri, M. Chatib, Mayara Felix, Rema Hanna, and Benjamin A. Olken (2021) "Tax Administration versus Tax Rates: Evidence from Corporate Taxation in Indonesia," *American Economic Review*, 111 (12), 3827–71, 10.1257/aer.20201237. 7, 30

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

Bergolo, Marcelo, Rodrigo Ceni, Guillermo Cruces, Matias Giaccobasso, and Ricardo Perez-Truglia (2023) "Tax audits as scarecrows: Evidence from a large-scale field experiment," *American Economic Journal: Economic Policy*, 15 (1), 110–153. 23

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," *Working Paper*. 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

Czajka, Leo (2023) "Fraud Detection Under Limited State Capacity: Experimental Evidence from Senegal,"Technical report, Working Paper. 30

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. 6

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

D'Agosto, Elena, Marco Manzo, Stefano Pisani, and Filippo Maria D'Arcangelo (2018) "The effect of audit activity on tax declaration: Evidence on small businesses in Italy," *Public Finance Review*, 46 (1), 29–57.

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

Haseeb, Muhammad and Kate Vyborny (2022) "Data, discretion and institutional capacity: Evidence from cash transfers in Pakistan," *Journal of Public Economics*, 206 (C), S0047272721001717. 6

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

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

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

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. 31

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.

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*. 6

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), 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. 22

Stevenson, Megan T and Jennifer L Doleac (2024) "Algorithmic risk assessment in the hands of humans," *American Economic Journal: Economic Policy*. 31

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. 7

Tables

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

		2014	2015	2016	2017	2018	2019	2020
A Self reported	CIT	5121	5954	6224	6575	6701	7210	6817
	VAT	11179	11905	12703	13355	13960	14190	13530
	PAYE	7055	7498	7870	8493	8757	8982	8619
	CGU	1589	1829	2033	2203	2649	2672	2799
	TAF	86	105	112	125	121	114	117
B Third party	Imports	8965	12404	13070	11850	13556	13676	10587
	Exports	1409	1727	1872	1830	1687	1660	1563
	Procurement	816	730	1387	1343	1904	1893	1690
	VAT annexes	6	9	21	804	3582	3199	NA
C Audits data	Digitized	NA	NA	1	3291	2756	2946	3717
	Self-reported (Excel)	NA	NA	NA	102	559	674	57

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 (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 (B) include transactions data from customs (imports and exports), procurement records, and VAT annexes. The audit data (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			
	Inspectors	Algorithm	Total	Inspectors	Algorithm	Random	Total
Large Taxpayer Unit	265	131	396	320	157	60	537
Medium Taxpayer Unit	155	153	308	320	290	135	745
Liberal Professions	46	50	96	235	222	85	542
SME(Regional)	71	73	144	422	401	84	907
Total	537	407	944	1297	1070	364	2731

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

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	3.25***	1.37***	0.01	0.18***	0.01	0.91*	-5.44	-7.41	-0.01
	(0.83)	(0.28)	(0.01)	(0.02)	(0.04)	(0.40)	(6.36)	(5.36)	(0.18)
N	22571	7420	5914	61513	60884	52035	697	703	640
R2	0.23	0.13	0.05	0.02	0.06	0.10	0.09	0.02	0.04
Mean outcome	10.81	14.82	-0.15	0.41	1.94	5.91	31.23	50.28	2.77

B: Comparison of Inspector versus Algorithm Selected Cases

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Algorithm	-1.68***	-1.54***	-0.06**	0.06*	0.14	0.18***	16.13	12.63**	-0.05
	(0.42)	(0.41)	(0.03)	(0.03)	(0.09)	(0.06)	(18.25)	(4.96)	(0.25)
Inspectors x Overlap	0.15	1.23**	0.01	0.09	0.35	0.10	-95.42	10.81	0.37
	(1.11)	(0.48)	(0.04)	(0.08)	(0.23)	(0.19)	(59.47)	(22.25)	(0.77)
N	1182	1016	899	1318	1296	1318	266	275	265
R2	0.12	0.12	0.09	0.14	0.05	0.06	0.28	0.13	0.12
Mean outcome	18.37	16.48	-0.02	0.51	2.04	2.13	44.71	45.50	2.80

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 B.10 shows the same exercise for desk audits). 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. 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 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	-0.17***	-0.05**	-0.04**	0.02	-0.04	-0.04	-0.65***	-0.24*	-0.12
	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.18)	(0.13)	(0.14)
				[-0.05 **,0.04*]		[-0.07***,-0.02]	[-0.77***,-0.52***]		[-0.31***,0.01]
Inspectors x Overlap	0.16**	0.05	0.05	0.08	-0.05	-0.04	0.29	-0.54*	-0.37
	(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		-0.02	-0.02		-0.25	-0.25
		(0.03)	(0.03)		(0.04)	(0.04)		(0.19)	(0.19)
Tax center 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	508	1011	992	457	747	728
R2	0.26	0.23	0.32	0.11	0.33	0.43	0.32	0.29	0.42
Mean outcome	0.53	0.37	0.37	0.89	0.73	0.73	19.26	17.76	17.73

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 C.1 shows intent-to-treat effects, Table C.3 shows robustness to controlling for audit slot fixed effects, and Table C.4 shows the results when using the evasion rate (instead of the evasion amount) as the outcome. Figure D.1 shows the distribution of audit list (=inspector×year) fixed effects for columns 3, 6 and 9. This table is discussed in Section 4.1.

Table 5: Resources Allocated to Audits and Audit Productivity, Full Audits

	A1: Resource Outcomes				**B1: Productivity Outcomes**			
	Number of Agents	Duration in Days (Taxpayer Survey)	Days from Start to Conf. (Admin. Data)	Days Working on Case (Self-Reported)	Evasion/ Number of Agents	Evasion/Duration (Taxpayer Survey)	Evasions/Duration (Admin. Data)	Evasion/Days Working (Self-Reported)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Algorithm	-0.26***	-8.54**	-30.18**	-27.26	-0.55***	-0.12	-0.50*	-0.74
	(0.08)	(3.72)	(13.75)	(30.60)	(0.18)	(0.39)	(0.26)	(0.49)
Inspectors x Overlap	0.04	-25.57	-9.83		0.30	0.96	-0.01	
	(0.17)	(21.89)	(27.66)		(0.44)	(2.20)	(0.57)	
N	508	214	282	51	457	98	253	40
R2	0.21	0.25	0.30	0.25	0.31	0.24	0.25	0.40
Mean outcome	2.87	28.06	160.07	177.07	18.19	15.90	15.00	13.62
	A2: Data Availability for Resource Outcomes				**B2: Data Availability for Productivity Outcomes**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Algorithm		-0.01	0.01	-0.05**	0.02	-0.00	-0.08**	-0.04**
		(0.04)	(0.01)	(0.02)	(0.03)	(0.04)	(0.04)	(0.02)
Inspectors x Overlap		-0.03	-0.02	-0.01	0.08	-0.01	0.07	0.00
		(0.06)	(0.04)	(0.04)	(0.05)	(0.06)	(0.07)	(0.03)
N		508	508	508	508	508	508	508
R2		0.22	0.96	0.47	0.11	0.22	0.49	0.41
Mean outcome	1.00	0.21	0.55	0.10	0.89	0.20	0.50	0.08

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of regressions of outcomes on the selection method of the case and tax office $\times$ year fixed effects. The sample includes all full audit cases selected and conducted from the 2018, 2019, and 2020 audit programs, for which we had data on the outcome variables. A shows the results for the resource and productivity outcomes, and B shows the results of regressions where data availability dummies are the outcomes. Column 1 shows the results for the number of agents involved in the audit, which is obtained from the audit records. 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. Column 3 uses 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, trim 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. Column 4 measures the duration of the audit using the self-reported number of days inspectors worked on a case, reported on a spreadsheet that inspectors needed to fill out during the 2019 audit program. Columns 5-8 use as outcomes the log of the ratio of evasion over the different cost measures used in columns 1-4. We multiplied the answer by the number of inspectors that worked in the case. Robust standard errors (Huber-White formula) are shown in parentheses. Table D.1 shows that same analysis for desk audits. This table is discussed in Sections 5.1 and 5.2.

Table 6: Algorithm Selection, Perceived Audit Efficiency and Corruption

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	-0.41***	0.01	-0.17	0.05	-0.05	-0.02
	(0.15)	(0.14)	(0.10)	(0.15)	(0.14)	(0.10)
Inspectors x Overlap	0.45**	0.04	0.16	-0.56*	-0.12	-0.18
	(0.18)	(0.26)	(0.21)	(0.29)	(0.27)	(0.20)
Algorithm x Random		-0.09	-0.05		-0.08	-0.04
		(0.19)	(0.18)		(0.18)	(0.16)
N	199	273	472	199	272	471
R2	0.10	0.02	0.03	0.11	0.10	0.05
Mean outcome	0.00	0.00	0.00	-0.07	0.05	0.00

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

	(1)	(2)	(3)	(4)	(5)	(6)
Algorithm	-0.36	0.37	-0.02	0.29	-0.24	0.02
	(0.26)	(0.23)	(0.17)	(0.24)	(0.22)	(0.17)
Inspectors x Overlap	0.19	0.40	-0.07	-0.33	-0.13	0.09
	(0.24)	(0.35)	(0.29)	(0.48)	(0.31)	(0.29)
Algorithm x Random		-0.01	0.08		-0.03	-0.04
		(0.27)	(0.24)		(0.26)	(0.23)
N	89	120	211	88	122	212
R2	0.10	0.08	0.03	0.16	0.13	0.05
Mean outcome	0.00	-0.06	-0.02	-0.14	-0.03	-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. A shows the results for all firms interviewed in the survey, conditional on them self-reporting an audit in recent years. 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 D.7. This table and the survey questions the indices are based on are discussed in Section 5.3.

Table 7: 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	-0.29***	-0.07***	0.01	-0.08*	-1.09***	-0.26
	(0.05)	(0.03)	(0.04)	(0.04)	(0.28)	(0.21)
Risk score	0.05**	-0.00	0.00	0.01	0.36***	0.38***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.13)	(0.13)
Alg. x Risk score	-0.02	0.02	0.00	0.01	-0.29*	-0.43***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.15)	(0.13)
N	944	2731	508	992	457	728
R2	0.27	0.32	0.11	0.43	0.33	0.43
Mean outcome	0.53	0.37	0.89	0.73	19.26	17.73

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. Tables F.1 to F.3 show robustness tests for these results. This table is discussed in Section 6.1.

Table 8: Explaining the Execution Gap for Full Audits: The Role of Observable Characteristics

A: Outcome P(Execution)					
	(1)	(2)	(3)	(4)	(5)
Algorithm	-0.172***	-0.123***	-0.121***	-0.117***	-0.115***
	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)
N	944	944	944	944	944
R2	0.262	0.332	0.334	0.338	0.344
Mean outcome	0.53	0.53	0.53	0.53	0.53
B: Outcome P(Detection|Execution)					
	(1)	(2)	(3)	(4)	(5)
Algorithm	0.019	0.024	0.027	0.027	0.027
	(0.029)	(0.030)	(0.030)	(0.029)	(0.030)
N	508	508	508	508	508
R2	0.110	0.122	0.130	0.130	0.137
Mean outcome	0.89	0.89	0.89	0.89	0.89
C: Outcome log(evasion)|Detection					
	(1)	(2)	(3)	(4)	(5)
Algorithm	-0.651***	-0.528***	-0.556***	-0.533***	-0.532***
	(0.183)	(0.186)	(0.187)	(0.186)	(0.186)
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	457	457	457	457	457
R2	0.323	0.368	0.386	0.404	0.406
Mean outcome	19.26	19.26	19.26	19.26	19.26

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 (A) is the probability that the case on the list was conducted by the inspectors. The second outcome (B) is the probability that an audit results in a penalty (initial or final notice), conditional on the case being conducted. The third outcome (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.

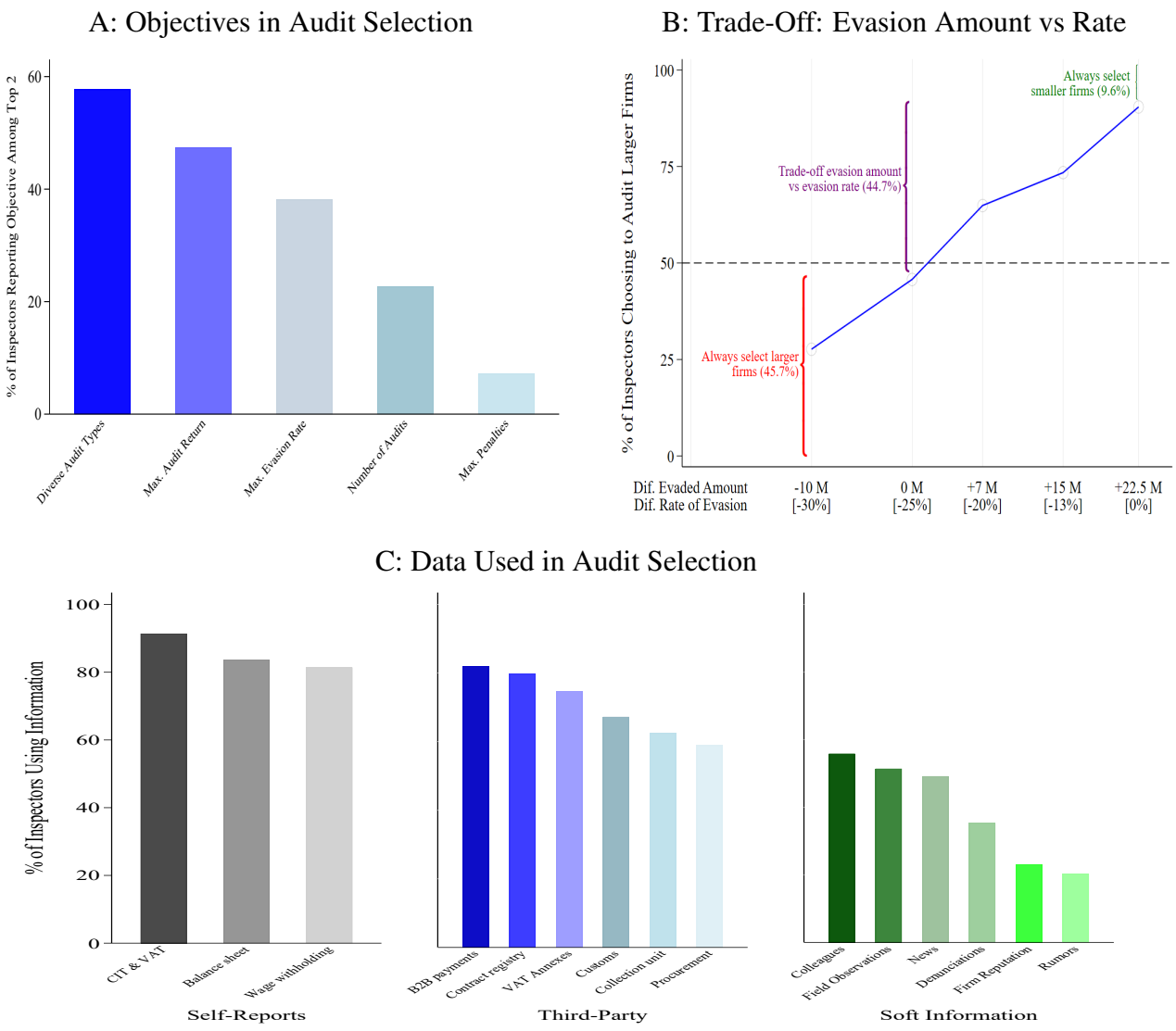
Table 9: Optimization Gains from Machine-Learning Algorithm Trained on Audit Outcome Data

Realized Revenue Log(mean)	Predicted Revenue Log(mean)	Δ Revenue vs Predicted w/ RF Selection Among Program Cases	Overlap Between Optimized and Realized Audit Program
(1)	(2)	(3)	(4)
A: Full Audits			
21.24	21.14	+ 18.09%	65%
B: Desk Audits			
19.00	18.73	+ 31.01%	61%

Notes: This table shows the results from our optimization exercise in which we use the audit results from the realized set of audits to train a machine-learning (random forest) algorithm to predict evasion. We optimize which audits were conducted in 2018 and 2019. Column 1 shows the realized revenue (evasion and fines) from audits across all tax offices. Column 2 shows the predicted revenue, based on the random forest model, 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 within tax office and year, 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 share of cases audited in the realized program that were also chosen to be audited under the optimized program. This table is discussed in Section 7. Table G.2 shows the same results disaggregated by tax office.

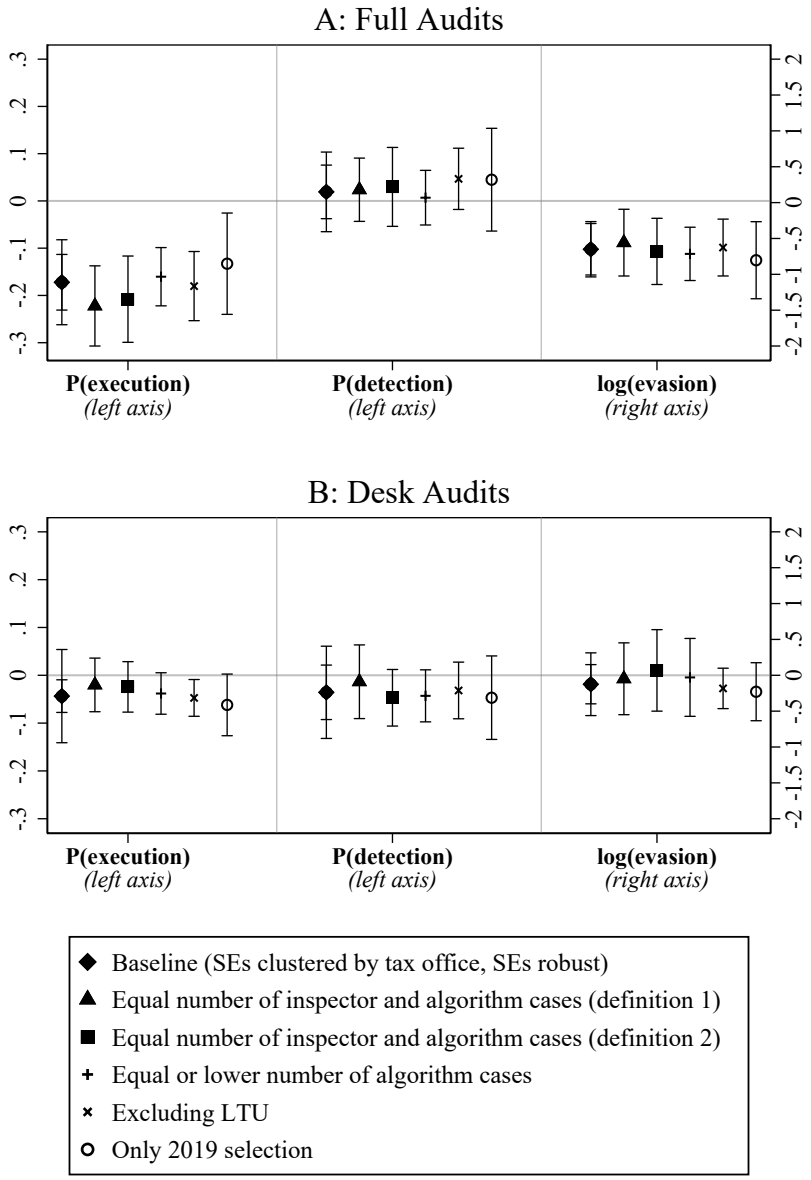
Figures

Figure 1: Tax Inspector Objectives and Practices in Audit Selection



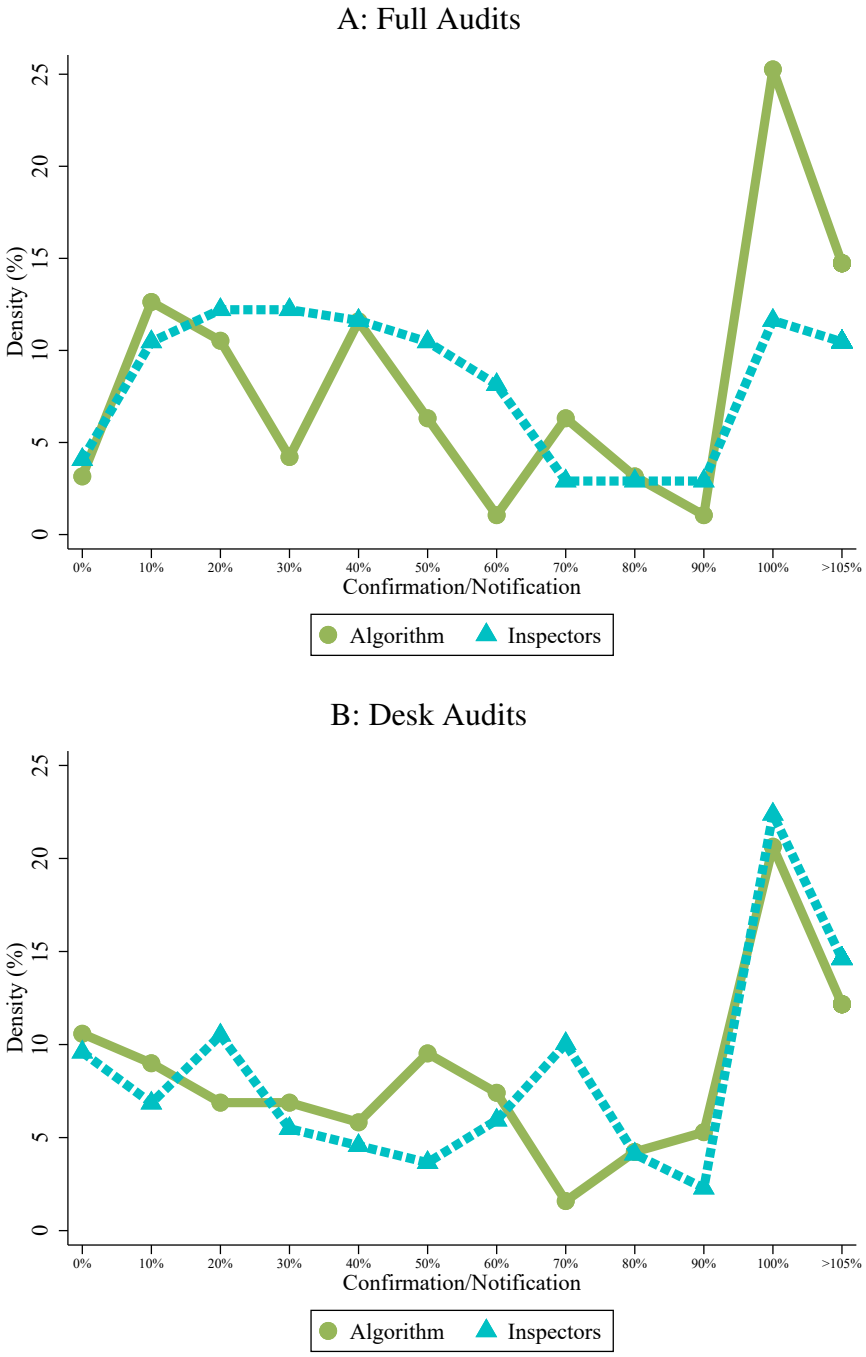
Notes: This figure shows inspectors' objectives, trade-offs and the information they use for audit selection. The data is collected at baseline (December 2017-January 2018) from our survey of 97 tax inspectors actively involved in conducting audits. Panel (a) shows the top two reported personal objectives 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. A shows results for full audits, and 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 robust standard errors, as in our main analysis, and standard errors clustered at the tax office × year level. The remaining coefficients use the same empirical specifications as Table 4, employing tax office × year fixed effects for full audits and inspector × year fixed-effects for the desk audits. We limited 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 19. Tables C.6 to C.11 show the details of all these estimations with numbers of observations in table format. This figure is discussed in Section 4.4.

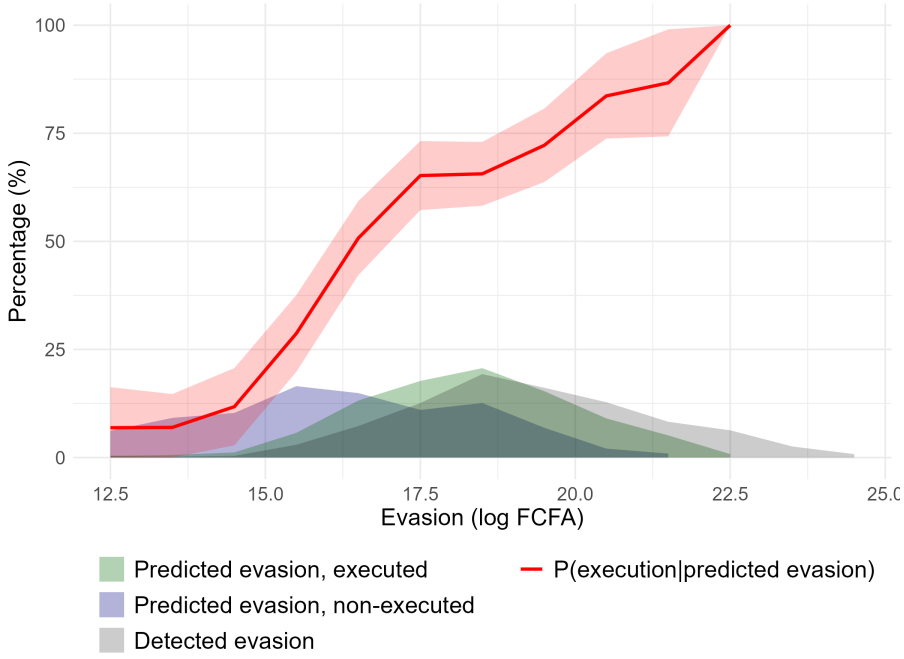
Figure 3: Dispute of Audit Results
Distribution of Confirmed Amount/Notified Amount of Evasion



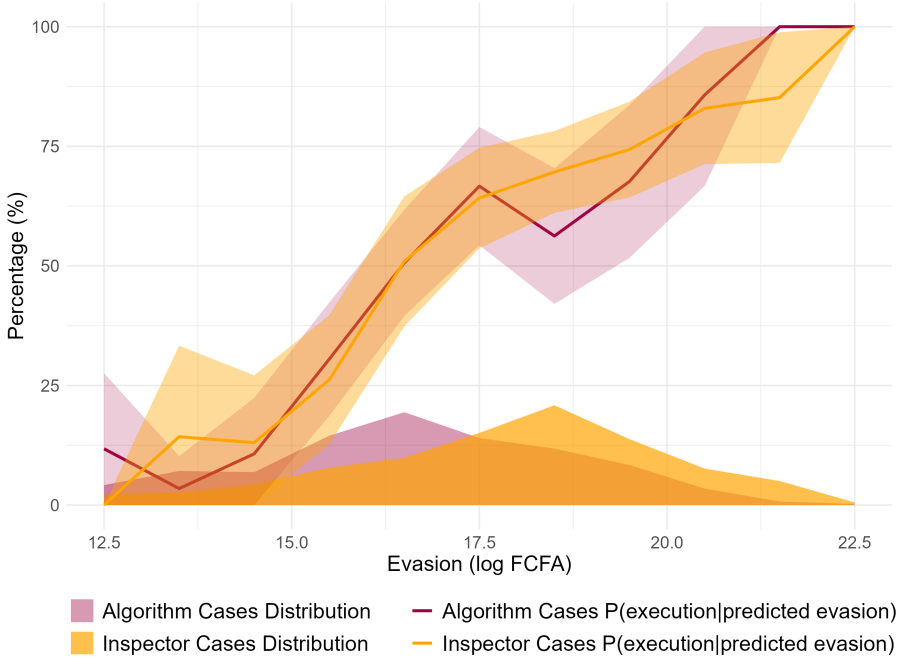
Notes: This figure shows the distribution of the confirmed amount over the notified amount of evasion, for each selection method and audit type. The share is calculated for audit cases were both a non-zero notified amount and a non-missing confirmed amount are reported. This reduces the sample of cases we can consider, for full audits from 508 to 264, and for desk audits from 1,011 to 371. 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: Predicted Evasion and Audit Execution, Full Audits

A: Predicted Evasion Amount and Realized Execution Rate, All Cases



B: Realized Execution Rate for Algorithm Cases and Inspector Cases



Notes: A plots the distribution of detected evasion for all full audits (log of FCFA amounts, red), as well as the results from a random forest model predicting detected evasion based on firms' observable characteristics. The blue distribution is predicted evasion for cases that were part of the annual audit program but not actually audited. The green distribution is predicted evasion for cases that were audited. The black solid line plots the realized audit execution rate by level of predicted evasion. B shows the distribution of predicted evasion for all algorithm-selected cases (orange) and for all inspector-selected cases (purple). The solid lines (orange and purple) plot the realized audit execution rates by level of predicted evasion. This figure is for full audits only. Figure F.2 shows the same results for desk audits, and Figure F.3 shows the distributions for inspector-selected and algorithm-selected cases separately. This figure is discussed in Section 6.4.

Online Appendix: Not for Publication

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

A Context Appendix

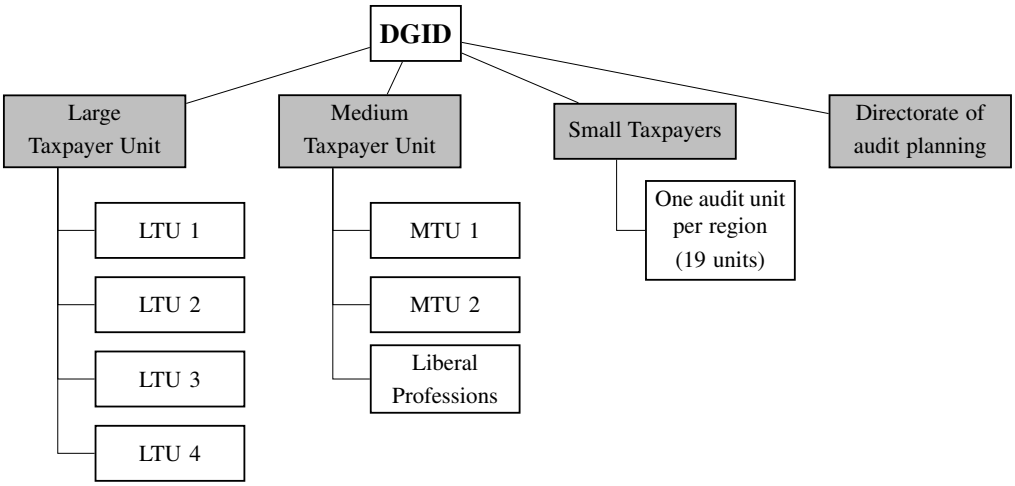
A.1 Tax Policy and Enforcement in Senegal

Table A.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

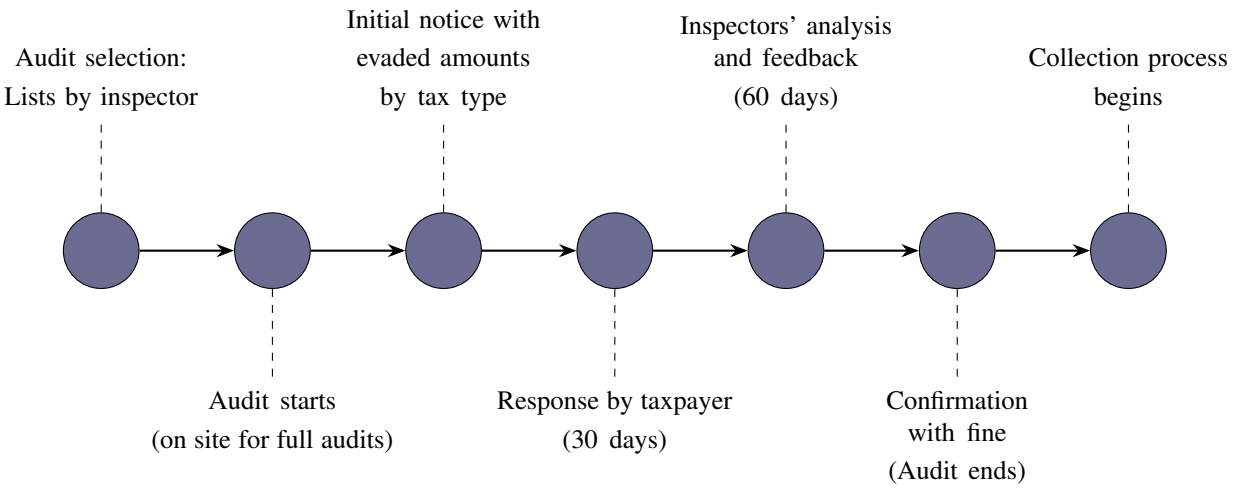
Notes: This table shows total tax revenues, including social contributions, as a percentage of GDP, along with the breakdown of total revenue by category (as a percentage of total revenue) for Senegal, Sub-Saharan Africa, lower-middle-income countries, and high-income countries. The data includes observations 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.

Figure A.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 A.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.

A.2 Audit Selection Around the World

Table A.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.

A.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.

B The Risk Scoring Algorithm

B.1 Designing the risk-score

This project was designed to assist the Senegalese tax administration (DGID) in assessing firms' tax evasion risk. Starting in 2017, the research team held consultations with DGID leadership and former tax inspectors to map the compliance risks of firms and to exploit all available data sources to assess this risk. We designed the risk-score 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 data submitted to DGID on corporate income taxes, VAT, personal income tax withholding remittance, as well as external data from customs (imports/exports) and public procurement contracts, for the period 2013-2016.[38]

The resulting risk-score algorithm is a parametric formula that aggregates indicators suggestive of non-compliance. It is not a non-parametric machine learning model to predict evasion. Instead, it selects firms based on simple and interpretable indicators. As explained in the main text, this choice was dictated by the lack of historical evasion data on firms and by the administration's wish to have an interpretable selection rule. Thus, one should consider the risk-score as a transparent best-practice risk assessment, given the administrative capacity, rather than a 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 to Senegal for administrative innovations.

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 data obtained from third-party reported data. 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. The definition of peers is based on the administrative division by DGID and economic sectors. Thus, for example, a firm that sells clothes may be flagged if its profit rate is at the bottom of the distribution of the profits of similar-sized firms that sell clothes.

Discrepancies and anomalies are aggregated to produce a risk-score for each taxpayer. Table B.1 summarizes the steps that we took to conceptualize the risk-score algorithm. We tried out several indicators that suggest under-declaration of tax liability. We discarded most based on analysis of data availability and statistical relevance.

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

Table B.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.

The computation of risk score changed in each edition of the algorithm. In 2018, we computed the ratios of the indicators and used these ratios as the risk score, adding them up for each firm. In the case of the 2019 edition, we computed deciles of the ratios and of the absolute deviations. For example, a firm would have 10 points if it were in the bottom decile of profit rate within its cluster, plus 7 points if the total profits were in the third lowest decile in absolute values. Finally, in 2020, we reduced the number of indicators and computed the risk score merely by flagging the top decile of each anomaly indicator, plus a flag if any inconsistency indicator was active.

In all years, we weighted the indicators to aggregate them for a single firm. The inconsistency indicators obtained high weights (of 5, 7, or even 12), whereas the anomaly indicators were weighted at 1. In the end, we also multiplied the total points by the log of turnover of the firm to allow the algorithm to give more weight to larger firms.

Table B.2: Inconsistencies Indicators

Indicator	**Description**	**Source**	**Year**
I1: (Treasury payments + Exports) > declared turnover	Firms that file a turnover lower than the sum of their exports plus treasury payments. The turnover should be at least equal to this sum.	L5 of CIT declaration or L5 of VAT declaration.	2018, 2019, 2020
I2: VAT turnover> CIT turnover	Firms that file a turnover value in the CIT declaration that is lower than that of the VAT declaration.	L5 of CIT declaration and L5 of VAT declaration.	2018, 2019, 2020
I3: max(Payroll + Imports, Total Sales based on VAT annexes) > declared turnover	Firms that declare a turnover lower than the sum of the payroll and imports, or of the total sales as obtained from VAT annexes (third party). This indicator is defined by taking the average values over the four previous years, both for the sum of payroll/imports and for the turnover. The idea is that in the medium run it is not possible to have costs that are always superior to the turnover.	L5 of CIT declaration, L5 of VAT declaration, payroll (RAS IRPP), customs, and VAT annexes.	2018, 2019, 2020
I4: max(Treasury payments + Export, Treasury payments + Imports) > 50 Million FCFA and firm is in CGU	Firms are eligible for the CGU regime if their turnover is lower than 50 Million F CFA. Here we check whether their turnover is likely below or above that threshold.	Treasury payments, customs.	2018, 2019
I5: inconsistent filing	Firms that have not declared their taxes regularly. Here, we flag firms that declared VAT but not CIT, or else firms that declared CIT, have declared VAT in the past, but do not declare VAT anymore.	CIT and VAT declarations	2018, 2019, 2020
I6: (Imports VAT) > (Imports customs)	Value of imports declared in VAT is greater than what is observed with customs data.	VAT declarations and customs.	2020
I7: Imports from low tax jurisdictions	This indicator chooses companies having a large share of their imports coming from a tax haven (CN, AD, AI, AF, BS, BH, BB, BZ, BM, VG, CR, CY, DJ, DM, GI, GD, HK, IE, JO, LB, LR, LU, MV, MH, MU, MS, NR, AN, PA, WS, SM, SG, CH, TO, VU)	Customs data.	2018, 2019, 2020

Table B.3: Anomaly Indicators - VAT

Indicator	Description	Source	Year
VAT A1: (turnover)/(Net VAT liability)	Inverse of the effective tax rate. If this indicator is high, it means that the company has a low tax burden.	Domestic turnover: L5-L10 Amount to be paid: L110	2018, 2019, 2020
VAT A2: (Non-taxable domestic turnover)/(Domestic turnover)	This indicator shows the proportion of turnover that is made up of non-taxable sales.	Non-taxable domestic turnover: Max (L15 + L20, L25-L10), domestic turnover: L5-L10	2018, 2019
VAT A3: (Reduced tax turnover)/(Domestic turnover)	This indicator indicates the relationship between the taxable turnover at a reduced rate and the taxable turnover at a normal rate.	turnover Reduced Rate: L50 or L50/0.1; turnover Normal rate: L45 or L55/0.18	2018, 2019
VAT A4: (Deductible VAT)/(Domestic turnover)	This indicator calculates the proportion of the tax deductions in the total sales (turnover).	Deductible VAT: Sum of lines L70 +L75 + L85 + L90 + L95; Domestic CA: L105	2018, 2019
VAT A5: (Reimbursement Credit claims)/(Deductible VAT)	Companies whose VAT deductions come largely from the request for reimbursement of credits. This mechanism is intertemporal and, therefore, more complex to increase VAT deductions.	Reimbursement application: L95 Deductible VAT: Sum of L70 +L75 + L85 + L90 + L95	2018, 2019
VAT A6: (Gross tax liability)/(Turnover)	Effective tax rate before tax deductions (like A1)	VAT declaration L60	2020

Special cases: VAT indicators are set to 0 for the following cases:

- Exporting firms benefiting from special tax regimes (in Senegal this regime is called EFE – Entreprise Franche d'exportation).
- Firms with advance payments (precompute) VAT or suspended VAT.
- Firms in sectors of health, banking, insurance, education, telecommunications and tourism in Senegal's southern Casamance (defined by sector codes)
- Firms that declare the financial activity tax(FAT). FAT indicators used instead of TVA indicators for firms that file for this tax

Table B.4: Anomaly Indicators - CIT

Indicator	Description	Source	Year
CIT A1: turnover / (net tax liability)	Business with a significant turnover with respect to the tax to be paid. This indicator measures the inverse of the effective tax rate. If this indicator is high, it means that the company bears a low tax burden in CIT. Notice, however, that the CIT is taxable on profits, not turnover. (In 2020, we computed this ratio as liability/turnover for ease of interpretation.)	turnover: L5 Gross amount to be paid: L470	2018, 2019, 2020
CIT A2: Taxable profit/(net tax liability)	Companies with a high taxable profit relative to the tax to be paid. This indicates the effective tax rate is low, that is, the amount actually paid in relation to the taxable profit. If this indicator is high, it means that the company pays little tax.	Taxable profit: L400 Gross amount to be paid: L470	2018, 2019
CIT A3: (Gross tax liability)/(Net tax liability)	This indicator shows the firms using much deductions applied to the gross amount of taxes to be paid (Decrease of taxable base after application of tax rate). If the company does not use deductions, this relationship is equal to 1.	Gross amount to be paid: L470 Net amount to be paid: L490	2018, 2019
CIT A4: Deductions/ turnover	Companies using a lot of deductions applied to turnover (decrease of taxable base before application of tax rate). This indicator shows the size of the deductions as a proportion of the turnover.	Deductions: L400-L405; turnover: L5	2018, 2019
CIT A5: Amortizations/Turnover	Companies using a lot of depreciation deductions as a share of their turnover. Depreciations being a type of deduction, this indicator of risk should be correlated with the previous one. Four-year averages are used to calculate this indicator, since it is an intertemporal variable.	Depreciation: L321 turnover: L5	2018, 2019
CIT A6: (Gross tax liability)/Turnover	This indicator shows the firms using much deductions applied to the gross amount of taxes to be paid (Decrease of taxable base after application of tax rate). If the company does not use deductions, this relationship is equal to 1.	Gross amount to be paid: L470; Net amount to be paid: L490	2020

Table B.5: Anomaly Indicators - TAF

Indicator	**Description**	**Source**	**Year**
TAF A1: High ratio of domestic to TAF	Companies with large domestic sales in relation to the tax to be paid. This indicates the actual tax, that is, the amount actually paid in relation to the CA. If this indicator is high, it means that the company pays can tax.	Domestic turnover: L10-L20 Tax payable: L110	2018, 2019
TAF A2: High ratio Non-taxable turnover / domestic TAF	Companies with a low effective tax rate of the exemption from Taf (Declaring a large portion of their non-taxable turnover). This indicator shows the proportion of turnover that is made up of non-taxable sales, but only for domestic turnover.	turnover Not taxable: L30 turnover Domestic L10-L20	2018, 2019
TAF A3: High ratio Taxable turnover low rate/ turnover normal rate	Companies with a low effective rate of taxation due to a Tax Reduced (report a large portion of their low-rate turnover). This indicator indicates the relationship between the taxable CA at a reduced rate and the taxable CA at a normal rate.	Low rate Taxable turnover: L70/0.07 CA Taxable Normal Rate: L50/0.17	2018, 2019
TAF A4: High Ratio gross deductible/TAF	Companies using a lot of deductions applied to value added (decrease of taxable basis). This indicator calculates the proportion of the deductions in the CA.	Deductible TAF: L100-L90 Gross amount to be paid: MaxL80, L50 + L70)	2018, 2019

B.2 Statistics on Audit Program

Table B.6: 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

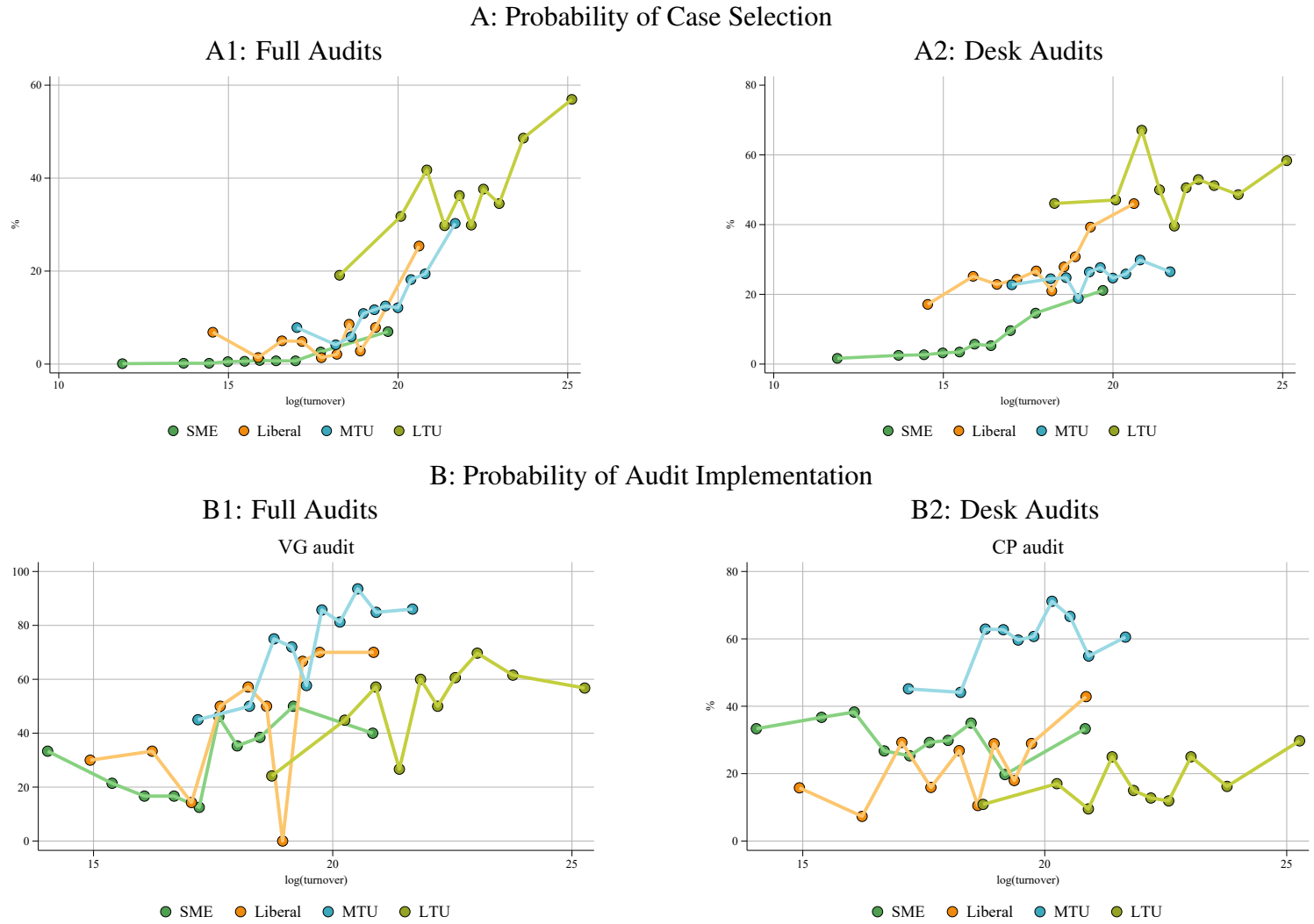
Notes: Number of selected full audits by year and tax office. The sum of the rows is larger than the total because there are overlapping cases between algorithm and discretion. This table is discussed in Section 3.2.

Table B.7: 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 B.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 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 B, we condition the computation on the set of selected firms. This figure is mentioned in Section 3.5.

B.3 Balancing Tests

Table B.8: 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	-0.0153	0.0110						
	(0.0208)	(0.0169)						
log(Mean Turnover)			0.00155	-0.000622				
			(0.00167)	(0.00176)				
log(Mean Tax Liability)					0.00133	0.00113		
					(0.00170)	(0.00168)		
Profit rate							-0.0126	-0.0176
							(0.0663)	(0.0614)
N	3034	3034	3034	3034	3034	3034	3034	3034
R2	0.00301	0.00217	0.00305	0.00209	0.00297	0.00219	0.00279	0.00207
Mean outcome	0.35	0.33	0.35	0.33	0.35	0.33	0.35	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 B.9: Balancing Tests of Information Intervention

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	-0.02	-0.07	-0.11	-0.17	-0.25	-0.27	0.05	0.01
	(0.02)	(0.19)	(0.56)	(0.26)	(0.28)	(0.75)	(0.04)	(0.00)
+ data spreadsheets	0.01	0.12	0.86	0.05	0.05	-0.28	-0.07*	0.00
	(0.02)	(0.19)	(0.54)	(0.26)	(0.30)	(0.75)	(0.04)	(0.00)
N	1172	1889	1451	1408	231	1419	1840	1531
R2	0.17	0.14	0.20	0.23	0.39	0.38	0.20	0.14
Mean outcome	-0.06	1.81	13.80	2.84	27.33	4.38	2.44	0.32
B: Availability of Observations								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Indicators	0.01	0.00	0.01	0.01	-0.02	-0.01	-0.03	0.02
	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)
+ data spreadsheets	-0.00	0.00	-0.02	-0.02	0.01	0.02	0.01	0.00
	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)
N	2136	2136	2136	2136	2136	2136	2136	2136
R2	0.26	0.12	0.18	0.17	0.27	0.09	0.10	0.15
Mean outcome	0.54	0.88	0.68	0.66	0.12	0.66	0.86	0.72

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.

B.4 Characteristics of selection for Desk Audits

Table B.10: Firm Characteristics of Algorithm vs Discretionary Selection

A: Characteristics of Inspector-Selected Cases									
Data source:	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
Inspectors' Selection	4.68***	0.99***	0.04*	-0.27***	-1.02***	2.37***	0.03	1.62	0.11
	(0.81)	(0.18)	(0.02)	(0.07)	(0.22)	(0.49)	(8.13)	(2.25)	(0.16)
N	22571	7420	5914	61513	60884	52035	697	703	640
R2	0.27	0.13	0.05	0.07	0.14	0.11	0.09	0.02	0.04
Mean outcome	10.81	14.82	-0.15	0.41	1.94	5.91	31.23	50.28	2.77
B: Comparison Among Selected Cases									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Algorithm	1.04***	-0.64**	-0.01	0.08***	0.58***	0.51***	1.00	3.59	-0.23
	(0.28)	(0.30)	(0.02)	(0.02)	(0.05)	(0.04)	(10.75)	(4.27)	(0.22)
Inspectors x Overlap	1.48***	-1.11*	-0.01	0.13***	0.28**	0.30***	5.24	1.59	-0.53
	(0.57)	(0.61)	(0.03)	(0.04)	(0.12)	(0.10)	(8.50)	(17.22)	(0.78)
Algorithm x Random	-0.81**	-0.11	-0.01	-0.10***	-0.29***	-0.25***	-5.21	4.72	0.04
	(0.39)	(0.45)	(0.03)	(0.03)	(0.09)	(0.07)	(7.74)	(5.71)	(0.32)
N	6422	4019	3292	13653	13427	13653	649	650	574
R2	0.29	0.17	0.15	0.26	0.24	0.33	0.30	0.22	0.21
Mean outcome	14.55	14.53	-0.06	0.17	1.43	1.33	22.59	51.15	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. A uses the sample of selected cases across the three years of the experiment, including only desk audit selection. 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 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.

B.5 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.

C Additional Results on Audit Outcomes

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

A (Linear Regression)	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	-0.17***	-0.04**	-0.14***	-0.04**	-3.54***	-0.81**
	(0.03)	(0.02)	(0.03)	(0.02)	(0.60)	(0.33)
Inspectors x Overlap	0.16**	0.05	0.19***	0.04	3.80***	0.82
	(0.07)	(0.04)	(0.07)	(0.04)	(1.46)	(0.67)
Algorithm x Random		-0.00		-0.02		-0.35
		(0.03)		(0.03)		(0.55)
N	944	2731	944	2731	893	2467
R2	0.26	0.32	0.24	0.19	0.27	0.26
Mean outcome	0.53	0.37	0.48	0.27	9.85	5.37
B (Poisson Regression)						
Algorithm	-0.32***	-0.12**	-0.30***	-0.17***	-0.37***	-0.15**
	(0.06)	(0.05)	(0.07)	(0.06)	(0.06)	(0.06)
Inspectors x Overlap	0.27***	0.14	0.34***	0.15	0.32***	0.16
	(0.10)	(0.10)	(0.12)	(0.12)	(0.11)	(0.12)
Algorithm x Random		0.02		-0.03		-0.02
		(0.06)		(0.09)		(0.08)
N	925	2504	925	2493	874	2230
Pseudo-R2	0.07	0.11	0.08	0.08	0.18	0.17
Mean outcome	0.54	0.40	0.49	0.29	10.07	5.94

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. A shows ordinary least square results, and B shows Poisson Pseudo-Maximum Likelihood estimates. Robust standard errors are shown in parentheses (Huber-White formula).

Table C.2: Algorithm Selection and Audit Outcomes, Controlling for Number of Agents and Years of Experience

	P(Detection | Execution)			log(Evasion) | Detection		
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	0.04*	-0.06**	-0.05**	-0.37**	-0.05	0.03
	(0.02)	(0.03)	(0.03)	(0.18)	(0.16)	(0.16)
Inspectors x Overlap	0.06***	-0.01	-0.01	0.24	-0.32	-0.08
	(0.02)	(0.02)	(0.03)	(0.43)	(0.33)	(0.37)
Algorithm x Random		0.06**	0.06*		-0.32	-0.33*
		(0.03)	(0.03)		(0.21)	(0.20)
Number agents	0.06***	0.01	0.00	1.08***	0.37**	0.36**
	(0.02)	(0.02)	(0.03)	(0.14)	(0.16)	(0.16)
Years experience	0.00	-0.01**	-0.01	0.08*	0.04	0.09***
	(0.00)	(0.00)	(0.01)	(0.05)	(0.03)	(0.03)
Tax center x Year	Yes	Yes	Yes	Yes	Yes	Yes
Inspector x Year	No	No	Yes	No	No	Yes
N	421	597	572	402	528	501
R2	0.14	0.45	0.55	0.41	0.28	0.44
Mean outcome	0.95	0.88	0.88	19.44	18.09	18.09

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 C.1 shows intent-to-treat effects, Table C.3 shows robustness to controlling for audit slot fixed effects, and Table C.4 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 C.3: 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	-0.17***	-0.05***	-0.04**	0.02	-0.04	-0.04	-0.63***	-0.25*	-0.13
	(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	0.05	0.08	-0.05	-0.04	0.27	-0.56*	-0.39
	(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		-0.02	-0.02		-0.26	-0.26
		(0.03)	(0.03)		(0.04)	(0.04)		(0.19)	(0.19)
Tax center 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	508	1011	992	457	747	728
R2	0.27	0.23	0.32	0.11	0.33	0.43	0.33	0.29	0.43
Mean outcome	0.53	0.37	0.37	0.89	0.73	0.73	19.26	17.76	17.73

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 C.1.

Table C.4: 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	-0.03	0.02	0.01	0.01
	(0.04)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)
Inspectors x Overlap	0.18**	-0.07	-0.13**	0.10	-0.02	-0.04
	(0.08)	(0.06)	(0.06)	(0.07)	(0.03)	(0.03)
Algorithm x Random		-0.00	-0.01		-0.00	-0.01
		(0.04)	(0.04)		(0.03)	(0.03)
Tax center x Year	Yes	Yes	Yes	Yes	Yes	Yes
Inspector x Year	No	No	Yes	No	No	Yes
N	499	976	956	496	974	953
R2	0.08	0.19	0.33	0.11	0.14	0.30
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 C.5: 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	0.12***	-0.00	-0.35	-0.12	-0.23	-0.21
	(0.04)	(0.05)	(0.22)	(0.18)	(0.26)	(0.22)
Inspectors x Overlap	-0.04	0.01	0.66	-0.41	0.51	0.06
	(0.06)	(0.09)	(0.39)	(0.40)	(0.37)	(0.47)
Algorithm x Random		-0.00		-0.20		0.02
		(0.09)		(0.26)		(0.28)
N	264	371	260	346	260	346
R2	0.17	0.23	0.34	0.54	0.20	0.50
Mean outcome	0.16	0.21	19.43	17.97	18.66	17.35

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 508 full audits to 264, and from 1011 desk audits to 371. 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.

C.1 Estimates Underlying the Robustness Figure

Table C.6: 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	-0.17***	-0.05	-0.04	0.02	-0.04	-0.04	-0.65***	-0.24	-0.12
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.16)	(0.18)	(0.19)
Inspectors x Overlap	0.16***	0.05	0.05	0.08**	-0.05	-0.04	0.29	-0.54	-0.37
	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)	(0.31)	(0.33)	(0.29)
Algorithm x Random		-0.00	-0.00		-0.02	-0.02		-0.25*	-0.25**
		(0.02)	(0.03)		(0.05)	(0.04)		(0.12)	(0.09)
Tax center 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	508	1011	992	457	747	728
R2	0.26	0.23	0.32	0.11	0.33	0.43	0.32	0.29	0.42
Mean outcome	0.53	0.37	0.37	0.89	0.73	0.73	19.26	17.76	17.73

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 C.7: 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	-0.22***	-0.02	-0.02	0.02	-0.02	-0.01	-0.56**	-0.10	-0.05
	(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		-1.41***	-1.11*
		(0.20)	(0.24)		(0.05)	(0.02)		(0.48)	(0.59)
Algorithm x Random		0.48***	0.47***		0.06	0.01		-1.62**	-1.45**
		(0.08)	(0.14)		(0.06)	(0.02)		(0.65)	(0.72)
Tax center 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	207	197	242	199	189
R2	0.19	0.17	0.23	0.09	0.07	0.18	0.27	0.38	0.46
Mean outcome	0.63	0.25	0.25	0.91	0.96	0.95	18.87	17.95	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 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 C.8: 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	-0.21***	-0.02	-0.02	0.03	-0.06*	-0.05	-0.68***	-0.17	0.07
	(0.05)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.23)	(0.26)	(0.29)
Inspectors x Overlap	-0.02	0.07	0.06	0.21*	0.01	-0.01	2.19***	-1.64***	-1.40***
	(0.16)	(0.08)	(0.09)	(0.11)	(0.02)	(0.01)	(0.48)	(0.51)	(0.51)
Algorithm x Random		-0.04	-0.08		0.08*	0.06		-0.26	-0.46
		(0.13)	(0.13)		(0.04)	(0.06)		(0.85)	(0.78)
Tax center 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	226	177	170	200	173	166
R2	0.20	0.12	0.19	0.10	0.07	0.24	0.20	0.28	0.43
Mean outcome	0.61	0.21	0.21	0.88	0.97	0.97	18.55	17.92	17.94

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 C.9: 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	-0.16***	-0.04*	-0.04*	0.01	-0.05*	-0.04	-0.71***	-0.18	-0.03
	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.19)	(0.23)	(0.27)
Inspectors x Overlap	0.13*	0.07	0.06	0.08	-0.03	-0.02	0.26	-1.24***	-1.06*
	(0.07)	(0.04)	(0.04)	(0.05)	(0.03)	(0.01)	(0.49)	(0.43)	(0.55)
Algorithm x Random		-0.03	-0.07		0.06	0.06		-0.18	-0.33
		(0.13)	(0.13)		(0.04)	(0.06)		(0.85)	(0.77)
Tax center 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	475	223	205	427	218	201
R2	0.26	0.12	0.20	0.11	0.07	0.25	0.31	0.29	0.46
Mean outcome	0.55	0.18	0.18	0.89	0.97	0.98	19.32	18.15	18.11

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 C.10: 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	-0.18***	-0.05**	-0.05**	0.05	-0.04	-0.03	-0.63***	-0.30**	-0.18
	(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	0.14*	-0.04	-0.04	1.03	-0.18	-0.13
	(0.13)	(0.05)	(0.05)	(0.07)	(0.06)	(0.06)	(0.87)	(0.35)	(0.40)
Algorithm x Random		-0.01	-0.01		-0.01	-0.02		-0.22	-0.20
		(0.03)	(0.03)		(0.04)	(0.04)		(0.19)	(0.20)
Tax center 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	321	901	895	290	642	636
R2	0.27	0.22	0.31	0.14	0.31	0.42	0.18	0.15	0.29
Mean outcome	0.58	0.41	0.41	0.90	0.71	0.71	18.58	17.45	17.45

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 C.11: 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	-0.13**	-0.06*	-0.06*	0.04	-0.06	-0.05	-0.80***	-0.40**	-0.23
	(0.05)	(0.04)	(0.03)	(0.06)	(0.05)	(0.04)	(0.27)	(0.19)	(0.21)
Inspectors x Overlap	0.23**	-0.02	-0.00	0.13**	-0.09	-0.10	0.54	-0.62	-0.35
	(0.11)	(0.08)	(0.08)	(0.05)	(0.10)	(0.09)	(0.71)	(0.39)	(0.45)
Algorithm x Random		0.05	0.06		-0.03	-0.03		-0.17	-0.17
		(0.04)	(0.04)		(0.06)	(0.06)		(0.25)	(0.25)
Tax center 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	197	546	546	169	299	299
R2	0.17	0.10	0.25	0.12	0.15	0.28	0.32	0.19	0.32
Mean outcome	0.60	0.60	0.60	0.85	0.54	0.54	19.27	17.46	17.46

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.

D Additional Results on the Audit Process

Table D.1: Resources Allocated to Audits and Audit Productivity, Desk Audits

A: Resource and Productivity Outcomes						
	Duration in Days (Taxpayer Survey)	Days from Start to Conf. (Admin. Data)	Days Working on Case (Self-Reported)	Evasion/Duration (Taxpayer Survey)	Evasions/Duration (Admin. Data)	Evasion/Days Working (Self-Reported)
	(1)	(2)	(3)	(4)	(5)	(6)
Algorithm	-1.08	17.94	3.25	-0.42	-0.47	0.32
	(2.12)	(19.11)	(4.62)	(0.48)	(0.34)	(0.44)
Inspectors x Overlap	13.61	22.35	-3.51	0.64	-1.49	0.26
	(12.54)	(40.58)	(3.67)	(0.96)	(1.15)	(1.28)
N	294	247	115	89	167	68
R2	0.28	0.40	0.54	0.43	0.58	0.41
Mean outcome	20.21	110.68	9.88	14.90	13.60	15.87
B: Data Availability						
	(1)	(2)	(3)	(4)	(5)	(6)
Algorithm	0.01	0.00	-0.03	0.01	-0.02	-0.04*
	(0.03)	(0.01)	(0.02)	(0.02)	(0.03)	(0.02)
Inspectors x Overlap	-0.10*	-0.00	-0.01	-0.05	-0.10*	-0.00
	(0.05)	(0.00)	(0.04)	(0.05)	(0.05)	(0.04)
N	992	992	992	992	992	992
R2	0.19	0.98	0.46	0.21	0.29	0.29
Mean outcome	0.16	0.27	0.12	0.11	0.20	0.07

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table is similar to Table 5, but focuses on desk audits. The regressions control for inspector $\times$ year fixed effects. We do not consider the number of agents working on a or evasion / number of agents as outcomes, as desk audit lists are selected and executed by individual inspectors or pairs of inspectors. There is hence no variation in the number of inspectors working on a case within a desk audit list. This table is discussed in Section 5.1.

Table D.2: Inspector Characteristics by Case Selection Method, Full Audits

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	-0.31	-0.95**	-0.04	-0.39	0.04	-0.09*	-0.02
	(0.33)	(0.42)	(0.21)	(0.24)	(0.03)	(0.05)	(0.03)
N	463	463	421	421	421	421	421
R2	0.35	0.39	0.36	0.37	0.31	0.34	0.35
Mean outcome	37.11	40.56	8.15	9.34	0.81	3.23	0.83

B: Only Cases in Which All Inspectors Reported Information

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Algorithm	-0.23	-0.65	0.19	-0.09	0.01	-0.02	-0.00
	(0.37)	(0.48)	(0.35)	(0.36)	(0.04)	(0.04)	(0.05)
N	226	226	115	115	115	115	115
R2	0.38	0.31	0.30	0.18	0.59	0.88	0.19
Mean outcome	37.54	42.09	9.24	11.31	0.76	3.35	0.76

C: Probability that All Inspectors Reported Information

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Algorithm	-0.03	-0.03	0.01	0.01	0.01	0.01	0.01
	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
N	508	508	508	508	508	508	508
R2	0.27	0.27	0.40	0.40	0.40	0.40	0.40
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.

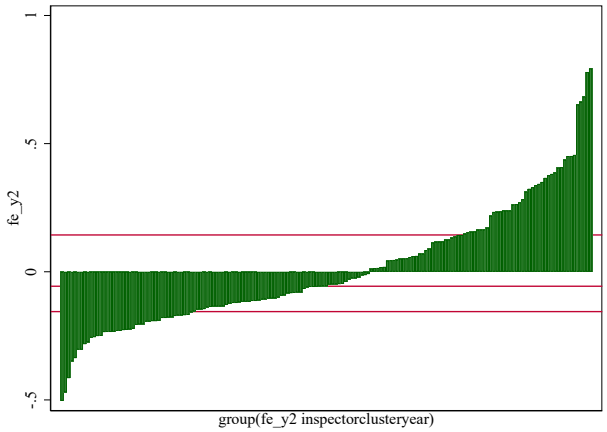
Table D.3: Inspector Characteristics and Performance, Desk Audits

	(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	-0.82	-0.02	0.08	1.23	-0.10	0.00	0.12	0.11
	(0.71)	(0.04)	(0.07)	(1.32)	(0.44)	(0.06)	(0.09)	(0.11)
Above median age	0.50	0.05	0.03	0.34	0.14	0.05	-0.05	-0.08
	(0.52)	(0.04)	(0.04)	(0.78)	(0.35)	(0.06)	(0.09)	(0.12)
Above median experience	-0.37	-0.03	-0.02	-0.68	0.24	-0.00	0.11	-0.16
	(0.44)	(0.04)	(0.04)	(0.92)	(0.30)	(0.05)	(0.09)	(0.10)
N	86	86	76	76	86	86	86	86
R2	0.82	0.53	0.57	0.56	0.74	0.53	0.45	0.09
Mean outcome	4.75	0.32	0.90	16.87	2.52	0.29	0.36	0.77

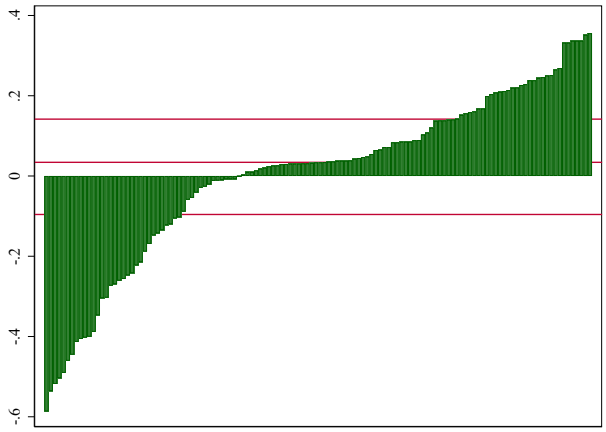
Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows the estimation of the OLS regressions of outcomes of inspectors on inspector characteristics. The analysis is done at the list (inspector x year) level and only includes the lists of desk audits. These lists were assigned individually 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, footnote 20.

Figure D.1: Inspector Quality Fixed Effects for Desk Audit

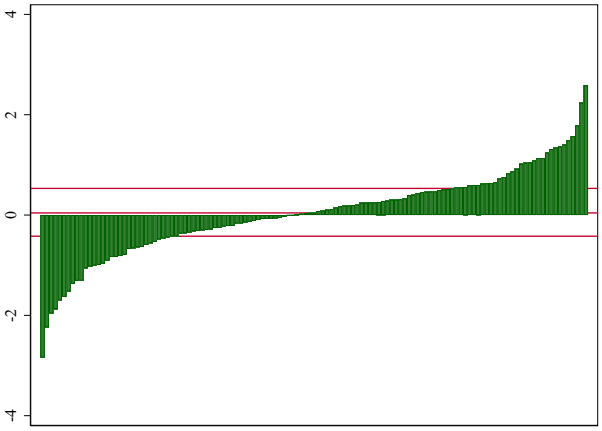
A: Execution of Audit



B: Detection of Evasion



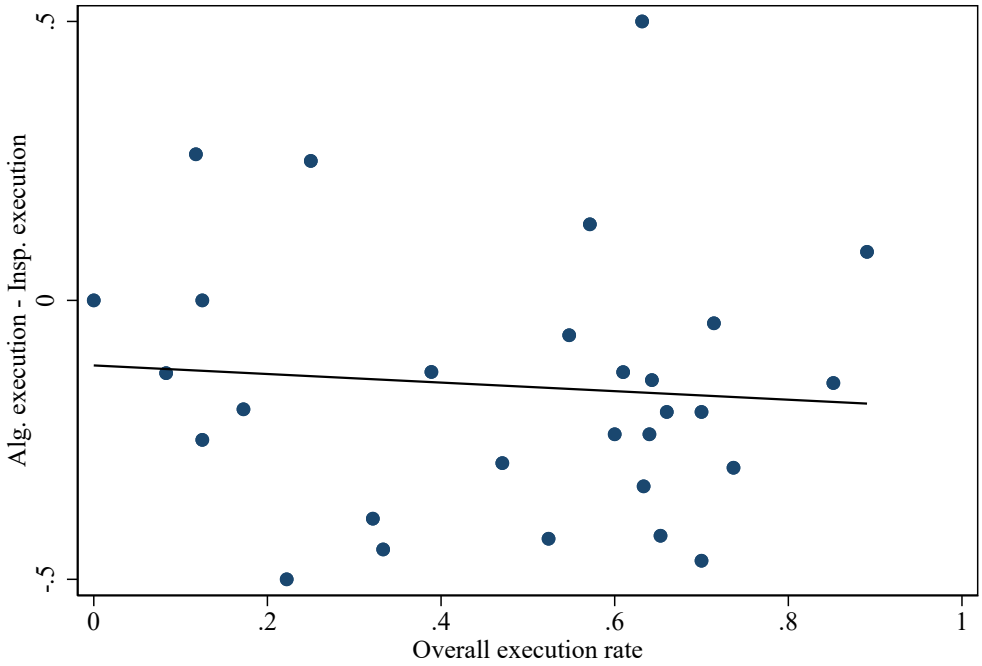
C: log(Evasion)



Notes: This figure shows the distribution of audit list (=inspector×year) fixed effects γ_l from estimating a version of Equation 1 where we additionally control for office fixed effects. This ensures that the list (i.e. inspector) fixed effects pick up variation within rather than across tax offices. The red lines mark p25, p50 and p75 of the distributions. The figure documents substantial heterogeneity in inspector skills, with long tails on either end of the distribution. To gauge the magnitude of the variation, consider that the mean outcome is 0.37 in A, 0.73 in B and 17.74 log points in C. An improvement in the selection of audit cases through an algorithm would be most valuable for inspectors in the bottom of the skill distribution. This figure is mentioned in Section 5.1, footnote 20.

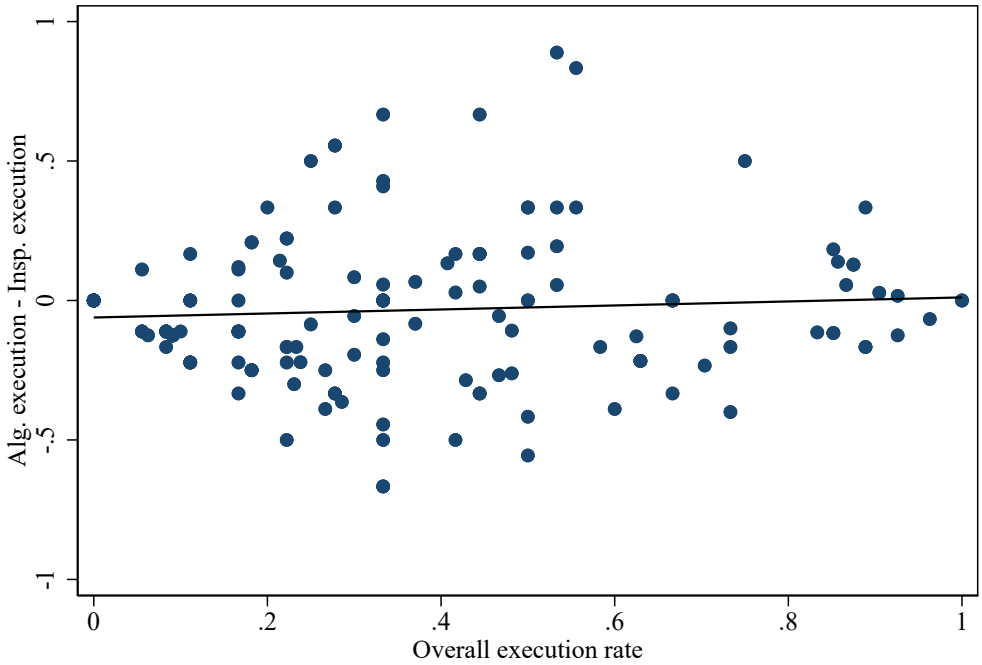
Figure D.2: Absence of Correlation Between Audit Execution Rate and Execution Rate for Algorithm-Selected Cases

A: Full Audits



Y = -0.12 - 0.08 X + ε , p-value = .675, R2 = 0.007

B: Desk Audits



Y = -0.06 + 0.07 X + ε , p-value = .325, R2 = 0.006

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, footnote 20.

Table D.4: Algorithm Selection and Audit Outcomes, Controlling for Number of Agents and Years of Experience

	P(Detection | Execution)				log(Evasion) | Detection			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Algorithm	0.02	0.08***	0.04*	0.04*	-0.63***	-0.35**	-0.36*	-0.36*
	(0.03)	(0.02)	(0.02)	(0.02)	(0.18)	(0.18)	(0.18)	(0.18)
Inspectors x Overlap	0.08	0.07*	0.06***	0.06**	0.29	0.36	0.27	0.24
	(0.05)	(0.04)	(0.02)	(0.02)	(0.46)	(0.43)	(0.44)	(0.43)
Lagged Log(Turnover)	0.01*	0.00	0.00	0.00	0.03	0.02	0.02	0.01
	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.02)	(0.02)
N. Agents		0.21***	0.06***	0.06***		0.96***	1.09***	1.01***
		(0.02)	(0.02)	(0.02)		(0.13)	(0.14)	(0.15)
Max. Years Experience				-0.00				0.09**
				(0.00)				(0.04)
N	508	508	421	421	457	457	402	402
R2	0.12	0.41	0.15	0.15	0.33	0.41	0.41	0.42
Mean outcome	0.89	0.89	0.95	0.95	19.26	19.26	19.44	19.44

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table is similar to Table 4, columns 4 and 7, but adds additional controls: firm size (turnover), the number of agents working on an audit team, the maximum number of years of experience on the team, and the maximum ability on the team, where ability is the inspector fixed effect estimated from the desk audit program (see Figure D.1). This table is discussed in Section 5.1, footnote 20.

Table D.5: 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	-1.24***	-0.46***	-0.32**	-0.24*	0.01	0.03***
	(0.29)	(0.15)	(0.13)	(0.13)	(0.01)	(0.01)
Inspectors x Overlap	0.14	-0.29	0.29	-0.14	0.00	-0.00
	(0.93)	(0.55)	(0.25)	(0.29)	(0.03)	(0.02)
Algorithm x Random		0.33		0.34**		-0.00
		(0.23)		(0.17)		(0.02)
N	457	728	457	728	454	712
R2	0.33	0.33	0.14	0.28	0.21	0.28
Mean outcome	4.91	3.08	3.53	2.65	0.41	0.44

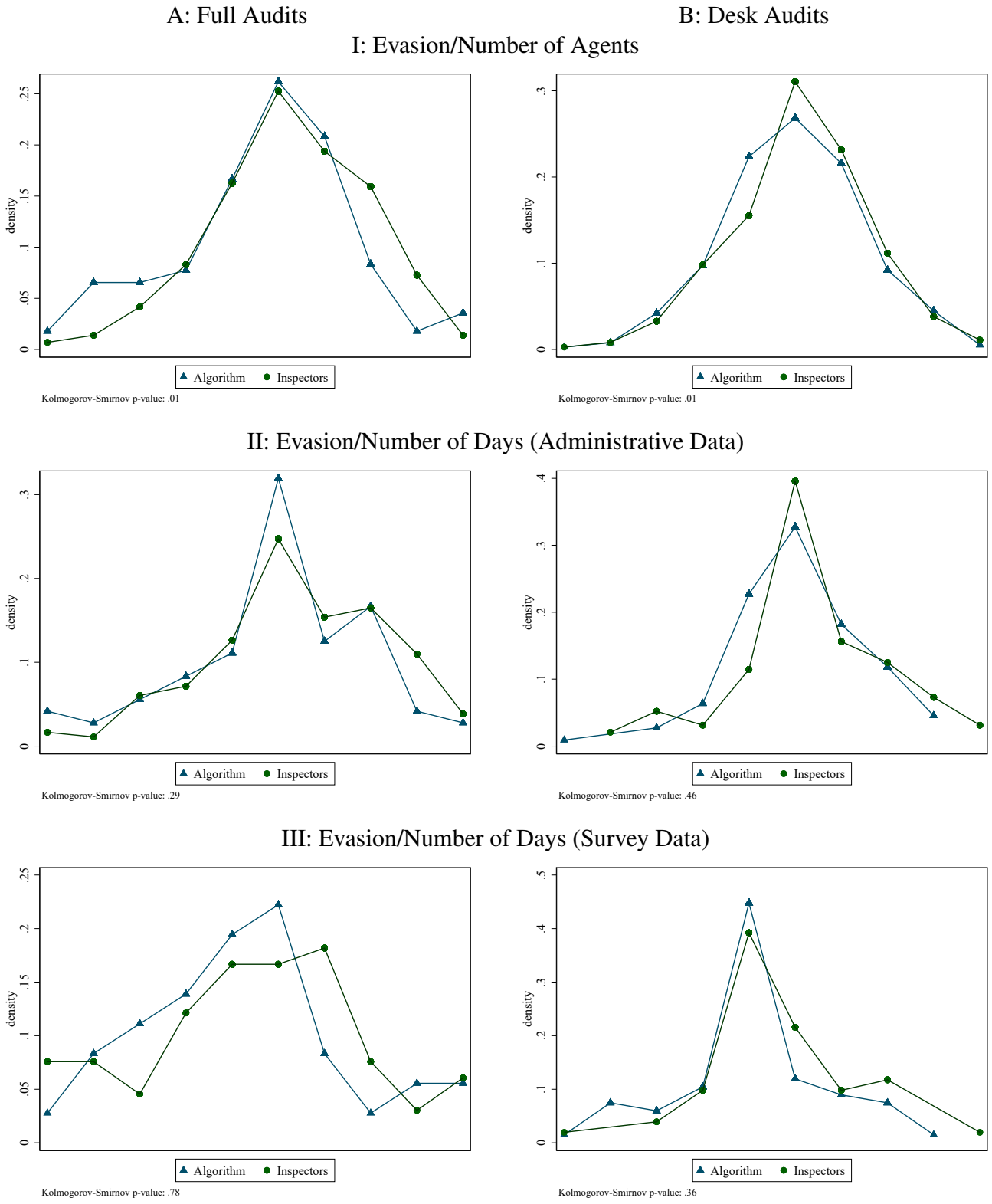
Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case, controlling for list fixed effects (year x tax office for full audits, year x inspector for desk audits). The outcomes are extracted from the audit reports. The three 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 D.6: 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	1.79	2.25	2.27	4.36	0.01	-0.04	-0.02	-0.01
	(3.71)	(3.75)	(3.19)	(3.03)	(0.02)	(0.03)	(0.04)	(0.04)
Inspectors x Overlap	-2.84	4.70	-13.54**	5.62	0.05**	0.04	-0.03	0.05
	(7.81)	(8.13)	(5.38)	(6.54)	(0.03)	(0.06)	(0.09)	(0.09)
Algorithm x Random		4.95		3.82		-0.01		0.04
		(5.32)		(4.32)		(0.05)		(0.06)
N	457	728	457	728	457	728	457	728
R2	0.18	0.31	0.29	0.37	0.17	0.28	0.11	0.26
Mean outcome	60.67	64.15	60.71	71.38	0.94	0.86	0.78	0.71

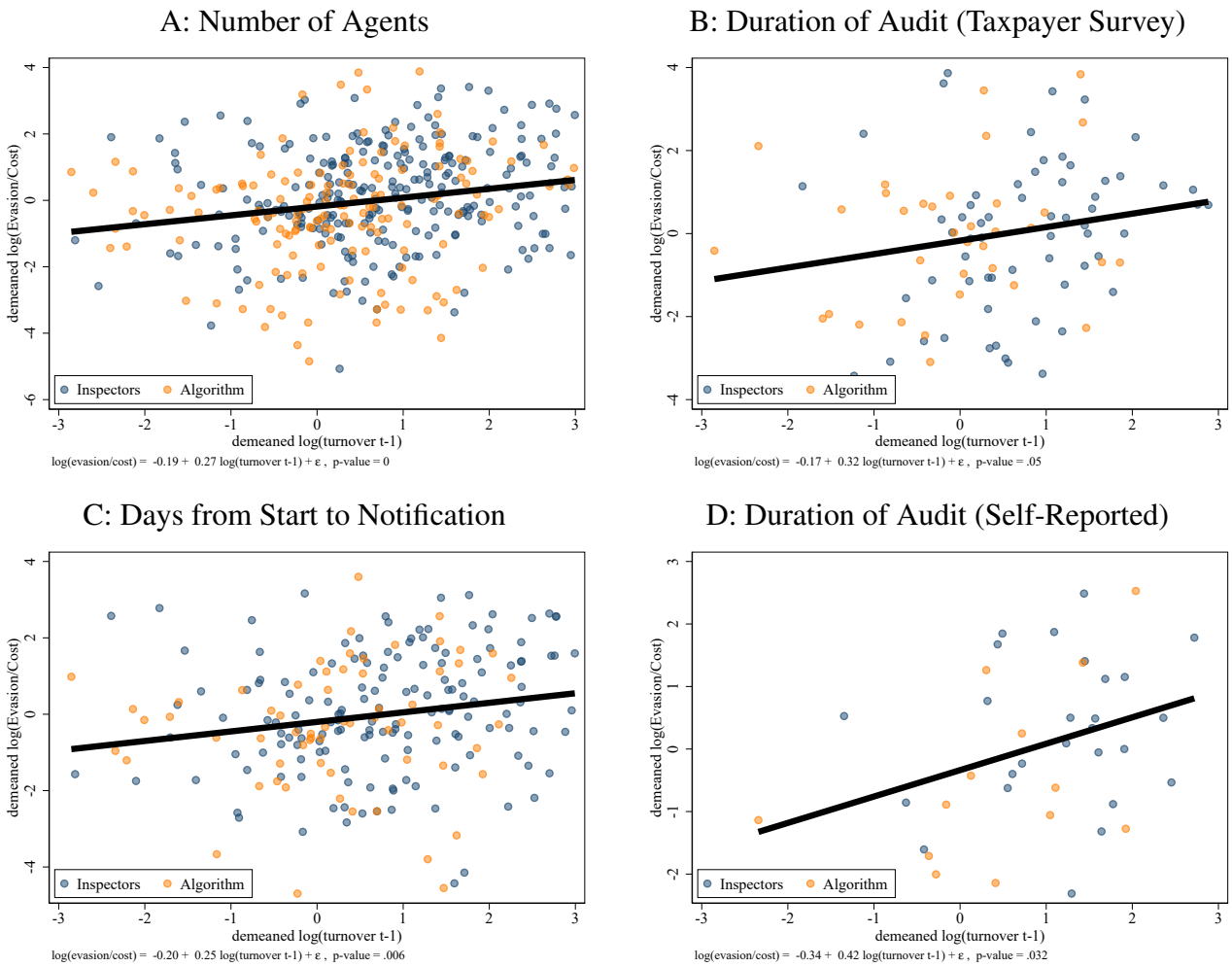
Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case, controlling for list fixed effects (year x tax office for full audits, year x inspector for desk audits). The outcomes are extracted from the audit reports. The three 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 D.3: 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 D.4: Productivity of Uncovering Evasion Using Different Definitions of Cost



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 D.7: Selection Method and Survey Results for Corruption - Probability of Answering the Question

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	-0.017	0.047	0.015	0.010	-0.028	-0.008	0.023	0.040	0.032	0.019	-0.007	0.007
	(0.076)	(0.077)	(0.054)	(0.017)	(0.031)	(0.017)	(0.047)	(0.067)	(0.040)	(0.053)	(0.052)	(0.037)
Inspectors x Overlap	-0.070	-0.163	-0.127	0.033	0.007	0.021	0.104	0.031	0.062	0.188*	-0.059	0.051
	(0.269)	(0.186)	(0.159)	(0.040)	(0.036)	(0.027)	(0.064)	(0.128)	(0.076)	(0.103)	(0.149)	(0.100)
Algorithm x Random		-0.060	-0.040		0.017	0.007		-0.124	-0.117		-0.017	-0.017
		(0.100)	(0.092)		(0.037)	(0.033)		(0.079)	(0.071)		(0.072)	(0.066)
N	199	252	451	199	252	451	199	252	451	199	252	451
R2	0.126	0.318	0.235	0.160	0.152	0.154	0.230	0.316	0.300	0.138	0.240	0.197
Mean outcome	0.65	0.62	0.63	0.97	0.96	0.97	0.87	0.76	0.81	0.88	0.86	0.87

B: Only Audited Firms

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Algorithm	-0.039	-0.114	-0.065	-0.013	-0.077	-0.046	0.033	-0.069	-0.016	-0.012	0.036	0.025
	(0.157)	(0.124)	(0.090)	(0.026)	(0.051)	(0.029)	(0.092)	(0.099)	(0.059)	(0.083)	(0.083)	(0.060)
Inspectors x Overlap	0.580***	0.026	0.297*	-0.001	0.037	0.030	0.086	-0.005	0.054	0.332***	-0.227	0.026
	(0.140)	(0.241)	(0.178)	(0.002)	(0.085)	(0.048)	(0.066)	(0.108)	(0.071)	(0.110)	(0.235)	(0.196)
Algorithm x Random		0.124	0.123		0.080	0.063		-0.014	-0.036		0.011	0.040
		(0.146)	(0.135)		(0.052)	(0.039)		(0.117)	(0.102)		(0.110)	(0.100)
N	85	110	195	85	110	195	85	110	195	85	110	195
R2	0.156	0.392	0.287	0.302	0.159	0.224	0.138	0.395	0.328	0.164	0.399	0.300
Mean outcome	0.67	0.62	0.64	0.96	0.97	0.96	0.90	0.80	0.85	0.88	0.83	0.85

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. A is for all selected firms interviewed in the survey, whereas 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 6 discussed in Section 5.3.

E The Deterrence Effects of Tax Audits

Summary

- Evasion rates (the share of turnover remaining unreported) are larger for smaller firms, but aggregate turnover, profits and tax payments are concentrated among larger firms.
- When the semi-elasticity of the evasion rate with respect to the audit probability does not vary with firm size, inspector-selected audits have a larger deterrence effect than algorithm-selected audits. With a semi-elasticity decreasing in firm size, selecting all audit cases via the risk-score algorithm could increase total declared turnover relative to the baseline, but only by 0.2% .
- We estimate from the data a semi-elasticity of -2 on average across firms: when the audit probability rises by 1 percentage point, the evasion rate falls by 2 percent. In this scenario, the deterrence effects under a pure inspector selection would raise reported turnover by 2% of baseline, compared to the current mixed selection. This is another reason why inspectors might prefer their selection.

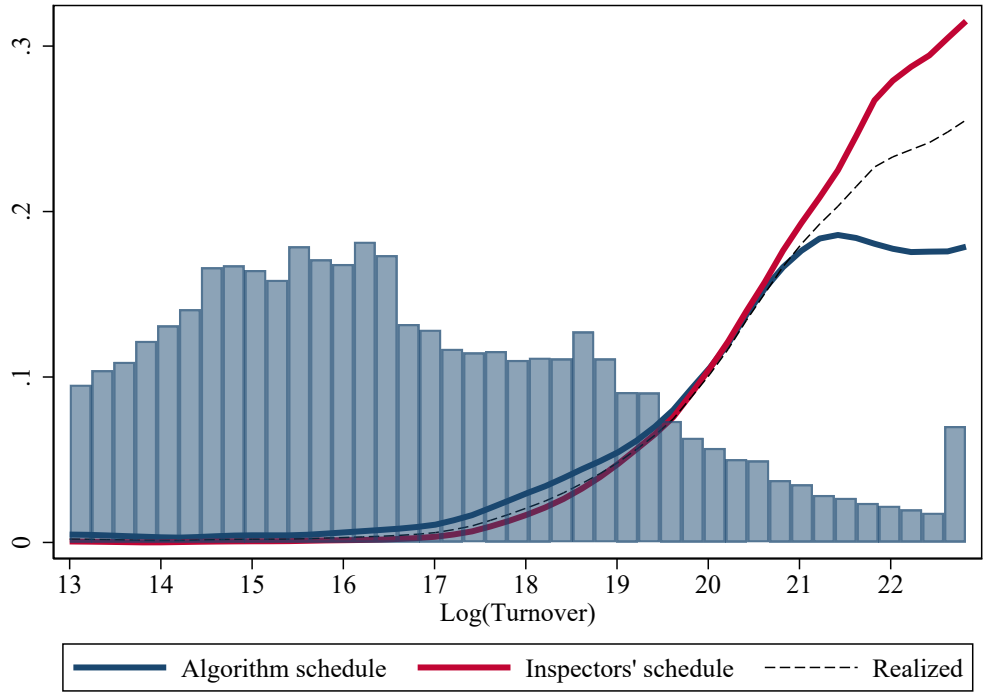
Objective If the goal of audits was to maximize overall tax revenue collection in a given year, the tax administration might take into account the two main effects of audits: the impact on tax evasion uncovered, and the deterrence effect due to the fear of an audit. The former effect may induce higher compliance of audited firms in the future ("ex-post compliance"), and the latter may improve the degree of truthful declarations, both among audited and non-audited firms ("ex-ante compliance"), as discussed in Paradisi and Sartori (2024). The deterrence effect depends on the perceived audit probability, which is notoriously difficult to measure (Bergolo et al., 2023). The section examines the following question: how large would the deterrence effect have to be for the risk-score selection to improve revenue collection compared to the discretionary selection?

As shown in Table 3, panel B, risk-score selected firms differ from those selected by inspectors along several margins (Section 3.5). We focus on firm size, which is arguably the most important and visible margin: risk-score selected firms are on average 1.78 log points smaller in terms of turnover. Figure E.1 plots the probability of a firm being selected for an audit as a function of its size (turnover), under the risk-score selection and under the inspector selection. As expected, both curves rise with firm size: firms in the bottom half of the distribution have a close to zero probability of audit under either schedule, but the risk-score starts selecting firms in the third quartile, whereas the inspector selection only starts selecting firms in the top quartile. The two curves cross at the 93rd percentile of reported turnover. For the largest firms, the risk-score audit probability is just under 20%, while the discretionary audit probability rises to 30%. Thus, if all audits in Senegal were risk-score-selected, keeping the number of audits constant, upper-mid-sized firms would face a higher audit probability than under discretionary selection, whereas the largest firms would face a lower audit probability.

How would a different schedule of audit probabilities by firm-size impact deterrence? To try to answer this question, we conduct a simulation based on simple assumptions about firm behavior and estimates from the data. First, we assume that the audit probability by firm size is known to firms, so that the perceived audit probability is equal to the actual audit probability. Moreover, we assume that the evasion rate (i.e., misreporting of turnover) is a function of the audit probability. By estimating the sensitivity of evasion rates to the audit probability, we can compute counterfactual levels of aggregate turnover declarations under the assumption that the entire audit program is selected by the risk score, and under the assumption that the entire program is selected by inspectors.

To estimate the impact of the audit probability on misreporting, we leverage the fact that firms with similar reported turnover may be allocated to different tax centers, e.g. because they are located in different regions, and thus be exposed to different audit probabilities. This quasi-random variation in the audit probability conditional on firm size allows us to estimate the effect of the audit probability on the evasion rate. We then use these estimated effects to compute the counterfactual.

Figure E.1: Audit Probabilities by Firm Size



Notes: This figure shows the probability of a full audit as a function of turnover using the full sample of firms with tax declarations. The realized probabilities, based on the 2018-2020 audits, are shown in the dashed line. The blue line shows the audit rate for the risk-score selection and the red line show the audit rate for the inspector selection. The blue bars show the distribution of firms by size, winsorized at the top 5%.

Empirical Analysis The data used for this exercise is the same as for the experiment in the paper: administrative data on tax declarations and audit reports. The analysis proceeds in three steps. First,

we compute the audit probability conditional on firm size and tax office, non-parametrically. For that estimation, we use only the executed full audits in 2018-2020. Let $P(\text{audit}_i)$ be a dummy variable indicating whether firm i has been audited or not. We estimate the following equation using a kernel and local linear regression:

$$P(\text{audit}_i) = f(\text{turnover}_i, \text{tax office}_i) + \xi_i \tag{E.1}$$

This first steps allows us to then generate a predicted audit probability $\hat{p}_i$ for all firms (both audited and non-audited), given a firm's size and tax office. $\hat{p}_i$ is a continuous variable ranging from 0 to 1.

Second, we estimate the impact of the audit probability on evasion rates. The evasion rate is defined as evasion/(evasion + turnover), and "evasion" is the amount of evaded taxes divided by 30% to reflect the average corporate income tax rate levied on profits. This means we assume that firms under-report profits by under-reporting turnover.

We regress evasion rates on audit probabilities to estimate the effect of changes in audit probability on compliance. However, despite being computed within tax offices and conditional on size, these audit probabilities could still be correlated with unobserved characteristics of firms that affect compliance. To mitigate endogeneity concerns, we estimate the impact of audit probability on evasion rates in a two-stage least squares model, where the first stage uses dummies of tax office as instruments for the audit probability. The idea is that a firm in the large taxpayer unit can be of the same size as another firm in the medium taxpayer unit but face a higher audit probability. The second stage estimates the impact of the audit probability on the log evasion rate, obtaining a semi-elasticity: the percentage reduction in the evasion rate in response to one percentage point increase in the audit probability. We also estimate a model allowing the semi-elasticity of the evasion rate to depend on log turnover by adding an interaction term between the audit probability and log turnover.

Equation E.2 formally represents this step, where we use the sample of audited firms in 2018-2020:

$$\begin{aligned} \hat{p}_i &= \alpha_0 + \alpha_1 \log \text{turnover}_i + \text{Office}_i + \epsilon_i \\ \log(\text{evasion rate}_i) &= \beta_0 + \beta_1 \log \text{turnover}_i + \beta_2 \hat{p}_i + \beta_3 \hat{p}_i \times \log \text{turnover}_i + \varepsilon_i \end{aligned} \tag{E.2}$$

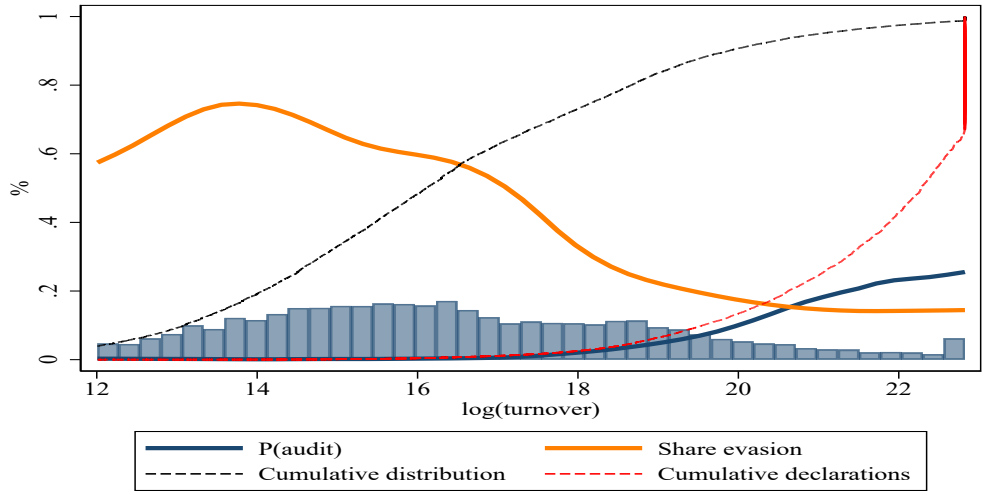
Deterrence estimates Figure E.2 shows the evasion rate as a function of firm size. Evasion rates fall substantially with firm size: audited firms with the median turnover in Senegal underreport more than 60% of their turnover, versus 25% at the 75$^{\text{th}}$ percentile and 17% at the 90$^{\text{th}}$ percentile. The dark dashed line shows the cumulative distribution of firms based on declared turnover, and the red dashed line shows the accumulated declared turnover at each level of turnover. Over 30% of the total declared turnover is concentrated in the top 5% largest firms.

The results from estimating Equation E.2 are shown in Table E.1: the semi-elasticity of evasion is

negative, as expected, at -1.95 (column 4). A 1 p.p. higher audit probability is associated with a reduction in the evasion rate of around 2 percent. The positive coefficient on the interaction term between the audit probability and log turnover (column 2) means that for larger firms, the semi-elasticity is lower, such that they react less to a higher audit probability.

To conduct counterfactuals, we need to impute evasion rates conditional on size for non-audited firms. We can then use the estimated elasticity to calculate the change in declared turnover for each firm as a result of shifts in the audit probability (the deterrence effect). We conduct the imputation by assuming that size-specific evasion rates in the general population of firms are identical to those in the sample of audited firms. This likely leads us to over-estimate evasion rates.[39]

Figure E.2: Evasion Rates, Audit Probability, and Distributions



Notes: This figure shows the estimated evasion rates (orange line) as a function of log(turnover) using the audited firms in 2018-2020, the probability of a full audit (blue line) as a function of turnover using the full sample of firms with tax declarations, the distribution of firms by size (blue bars), the cumulative distribution of firms by size (black dashed line), and the cumulative declared turnover (red dashed line).

[39] Given that the firms used for the estimation were selected for audits, their evasion rates are likely higher than the evasion rates in the wider population of firms (at a given firm size). For simplicity, we assume that the evasion rates in the audit sample apply to all firms of a given size. In alternative calculations, we could adjust the evasion rates by firm size either in a reduced-form manner by using the risk-score of firms (controlling for size), or the machine learning model predicting evasion, applied to the entire sample of firms. This adjustment substantially changes the overall evasion in the population, but the results presented here remain qualitatively similar. These results are available upon request.

Table E.1: Evasion Rate and Audit Probability

	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
Audited probability	-24.41***	37.04**	-5.48***	-2.25*
	(7.31)	(18.80)	(1.52)	(1.30)
Audited probability X Log(Turnover)	1.04***	1.51*		
	(0.39)	(0.81)		
Log(Turnover)	-0.39***	-0.08***	-0.15***	-0.10***
	(0.09)	(0.01)	(0.03)	(0.01)
N	385.00	329.00	385.00	329.00
R2	0.19		0.18	

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows the results of a regression of the evasion rate on the audit probability, following Equation E.2. Columns 1 and 3 show OLS regressions, while columns 2 an 4 show 2SLS estimations with tax office dummies as instruments for the audit probability.

Counterfactual reported turnover under alternative audit schedules Finally, we compute the counterfactual reported total turnover under the risk-score schedule and under the inspectors' selection schedule (using the schedules shown in Figure E.1). Using the risk-score schedule for the entire audit program would raise audit probabilities for upper-mid-sized firms but lower them for the largest firms, compared to the realized schedule. The opposite is true when using the inspectors' schedule. Figure E.3 shows how much more or less total turnover would be declared under each schedule for different values of the semi-elasticities of the evasion rate to the audit probability. The estimated elasticity from our data is shown as a vertical dashed line.

Panel A assumes a constant semi-elasticity (Table E.1, column 4). Panel B assumes the elasticity falls (in absolute value) with firm size (Table E.1, column 2). The solid lines show the percentage point difference in the total turnover declared due to the deterrence effect, relative to the current realized schedule (a mix of risk-score and inspector selection). The red line corresponds to the scenario where all executed audits followed the inspector schedule. The blue line corresponds to the scenario where all executed audits followed the risk-score schedule.

In both figures, for low values of semi-elasticities, we see that the turnover reported due to deterrence when following the inspectors' schedule over-performs the realized values. This is due to the very large concentration of turnover at the top of the firm-size distribution. In the case of a constant semi-elasticity (Panel A), the advantage persists for all values of the elasticity. In contrast, in the case of varying semi-elasticities (Panel B), when the slope of the semi-elasticity on log turnover is above 1.7, the risk-score schedule performs better. However, even in those cases, the aggregate gains from following the risk-score schedule would only amount to about 0.2% more than the realized declarations.

The figures also show dashed blue and red lines showing the impact on firms below the top 10th decile of turnover. In those cases, the risk-score schedule always over-performs relative to the realized values and to the inspectors' schedule. This result reflects the fact that the risk-score schedule exhibits higher audit probabilities for these firms outside the top decile. However, the reduced turnover in the top decile dominate any increase in turnover from the bottom nine deciles.

Using our estimated semi-elasticity of -2.05 as benchmark and applying it to all firms regardless of size, we find that the inspectors' schedule would have yielded 2% more total declared turnover in Senegal. These results suggest another reason for why tax inspectors might prefer their own selection: the deterrence effect of the inspectors' schedule is likely larger, leading to higher overall tax revenues.

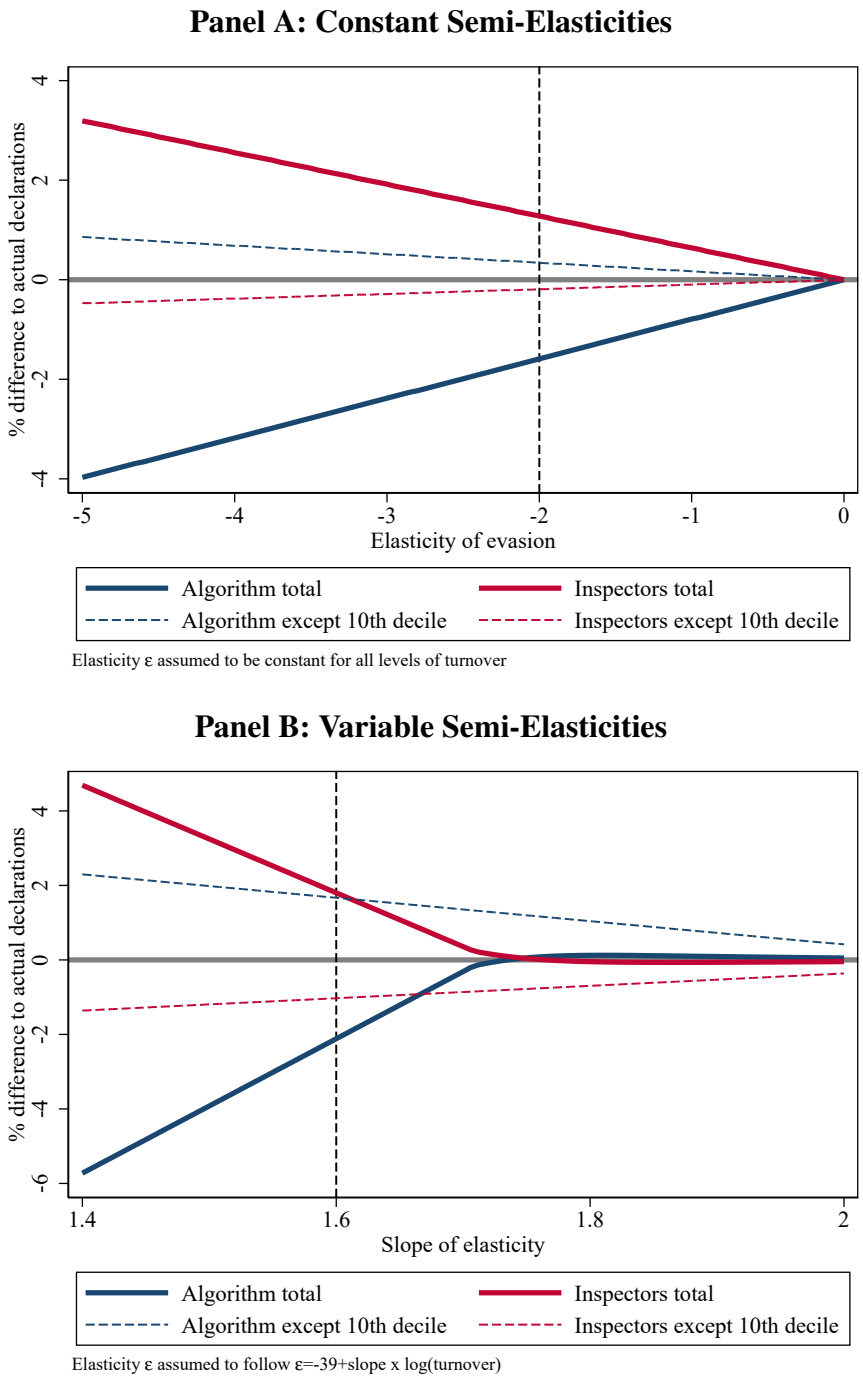
Relation to the Literature Finally, one may ask what is a reasonable range of (semi) elasticities of evasion with respect to audit probabilities. Unfortunately, the literature does not provide a clear answer to this question. Studies on the direct deterrence effects find moderate but significant and lasting effects of actual audits on compliance (Advani et al., 2023; Alm, 2019). Alm (2019) states that *"A more consistent result across studies is that a higher audit rate leads to more compliance, with an estimated reported income-audit rate elasticity that varies within the narrow range of 0.1 to 0.2; this impact appears to be small and non-linear, so that the deterrent effect of a higher audit rate eventually diminishes."*

Recent work, that studies changes to communicated audit probabilities, finds unclear evidence of indirect deterrence. Bergolo et al. (2023) find no significant effect of modifying audit probability signals in tax compliance letters, which they argue stems from a behavioral bias called probability neglect. They discuss that even for the few other papers (Dwenger et al., 2016) that have found a positive association between reporting rates and communicated audit probabilities, the implied deterrence elasticities are very small.[40] Further, the literature is also not definitive on whether evasion elasticities with the audit rate are size-dependent: some paper find that smaller firms are more elastic (Pomeranz, 2015; Naritomi, 2019), but other find similar elasticities across firm size (Bergolo et al., 2023; Best et al., 2021).

These results imply that the deterrence elasticity we estimate with the instrumental variable in this appendix is on the higher end of the literature's estimates. Similarly, the substantial size-dependence in this elasticity might not be a general feature. Yet, even with large elasticities and large size-dependence, our estimated deterrence effects do not favor the risk-score algorithm selection.

[40] One exception is DAgosto et al. (2018) indicating that if the audit rate in a taxpayer's province increases by 1%, tax compliance increases by 1.4%. These difference might reflect the micro vs macro approach.

Figure E.3: Counterfactual Turnover Declarations Relative to Realization



Notes: This figure shows the results of simulation exercises varying the semi-elasticity of evasion rate with respect to percentage point variations in the probability of audit. Evasion rate is defined as the share of underreported turnover in tax declarations. In the figures, we show how aggregate turnover would change relative to the realized aggregate turnover in 2020 in two scenarios: the entire audit program is determined by the inspectors' schedule (red line) or by the algorithm's schedule (blue line). The dashed lines show the relative change in total turnover accumulated up to the 9th decile of turnover. Panel A shows how the total turnover (and accumulated up to the 9th decile) would differ from the realized turnover for different values of the semi-elasticity of evasion rate, assuming that all firms have the same semi-elasticity. Panel B shows the same exercise assuming that the semi-elasticity follows the formula $\varepsilon = -41 + \text{slope} \times \log turnover$ and we vary the slope on the X-axis. In both figures, the vertical dashed lines indicate the estimated values based on Equation E.2.

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	-0.24***	-0.07**	0.03	-0.08*	-1.04***	-0.26
	(0.05)	(0.03)	(0.04)	(0.04)	(0.28)	(0.21)
Risk score	0.02	-0.02	-0.01	0.01	0.34***	0.38***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.13)	(0.13)
Alg. x Risk score	0.01	0.04**	0.01	0.02	-0.28*	-0.43***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.15)	(0.13)
Lagged log(turnover)	0.02***	0.01***	0.01**	0.01***	0.02	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)	(0.01)
N	944	2731	508	992	457	728
R2	0.33	0.33	0.13	0.44	0.34	0.43
Mean outcome	0.53	0.37	0.89	0.73	19.26	17.73

This table is identical to Table 7, 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***	-0.08***	0.10*	-0.07*	-0.74**	-0.26
	(0.06)	(0.03)	(0.05)	(0.04)	(0.33)	(0.23)
Unw. Risk score	0.06**	0.00	0.01	0.03	0.31*	0.07
	(0.03)	(0.02)	(0.02)	(0.02)	(0.17)	(0.17)
Alg. x Unw. Risk score	-0.04	0.04	-0.05	-0.00	-0.34	0.03
	(0.04)	(0.02)	(0.03)	(0.03)	(0.22)	(0.19)
N	944	2731	508	992	457	728
R2	0.27	0.32	0.12	0.43	0.33	0.42
Mean outcome	0.53	0.37	0.89	0.73	19.26	17.73

This table is identical to Table 7, 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	-0.27***	-0.27***	-0.07***	-0.08**	-0.02	-0.02	-0.05	-0.02	-1.06***	-1.26***	0.07	-0.03
	(0.05)	(0.06)	(0.03)	(0.03)	(0.04)	(0.05)	(0.04)	(0.05)	(0.35)	(0.38)	(0.24)	(0.29)
VAT anomalies risk score		0.01		0.00		0.01		-0.00		0.34		0.22
		(0.03)		(0.02)		(0.02)		(0.02)		(0.21)		(0.17)
CIT anomalies risk score		-0.04		0.02		-0.01		0.00		0.13		-0.07
		(0.03)		(0.02)		(0.03)		(0.02)		(0.20)		(0.17)
Inconsistencies risk score	0.03	0.04	0.01	0.01	0.02	0.02	0.03	0.03	0.27	0.24	-0.14	-0.15
	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.20)	(0.20)	(0.14)	(0.15)
Anomalies risk score	0.04*	0.06	0.02	0.00	0.03*	0.03	-0.02	-0.02	0.17	-0.09	-0.02	-0.11
	(0.02)	(0.04)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.12)	(0.21)	(0.13)	(0.22)
Algorithm x Inconsistencies	-0.00	-0.01	0.01	0.01	-0.03	-0.03	-0.02	-0.01	-0.43*	-0.41	0.19	0.22
	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.25)	(0.25)	(0.18)	(0.18)
Algorithm x Anomalies	-0.05	-0.05	-0.02	-0.01	-0.06*	-0.06	0.02	-0.03	-0.05	0.27	0.05	0.20
	(0.03)	(0.04)	(0.02)	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)	(0.18)	(0.26)	(0.15)	(0.27)
Algorithm x VAT anomalies		-0.05		-0.01		0.00		0.08*		-0.49*		-0.40*
		(0.04)		(0.02)		(0.03)		(0.04)		(0.27)		(0.23)
Algorithm x CIT anomalies		0.06*		-0.00		0.00		-0.01		-0.02		0.23
		(0.03)		(0.02)		(0.03)		(0.03)		(0.24)		(0.20)
N	944	944	2731	2731	508	508	992	992	457	457	728	728
R2	0.27	0.27	0.32	0.32	0.12	0.12	0.43	0.44	0.34	0.34	0.42	0.43
Mean outcome	0.53	0.53	0.37	0.37	0.89	0.89	0.73	0.73	19.26	19.26	17.73	17.73

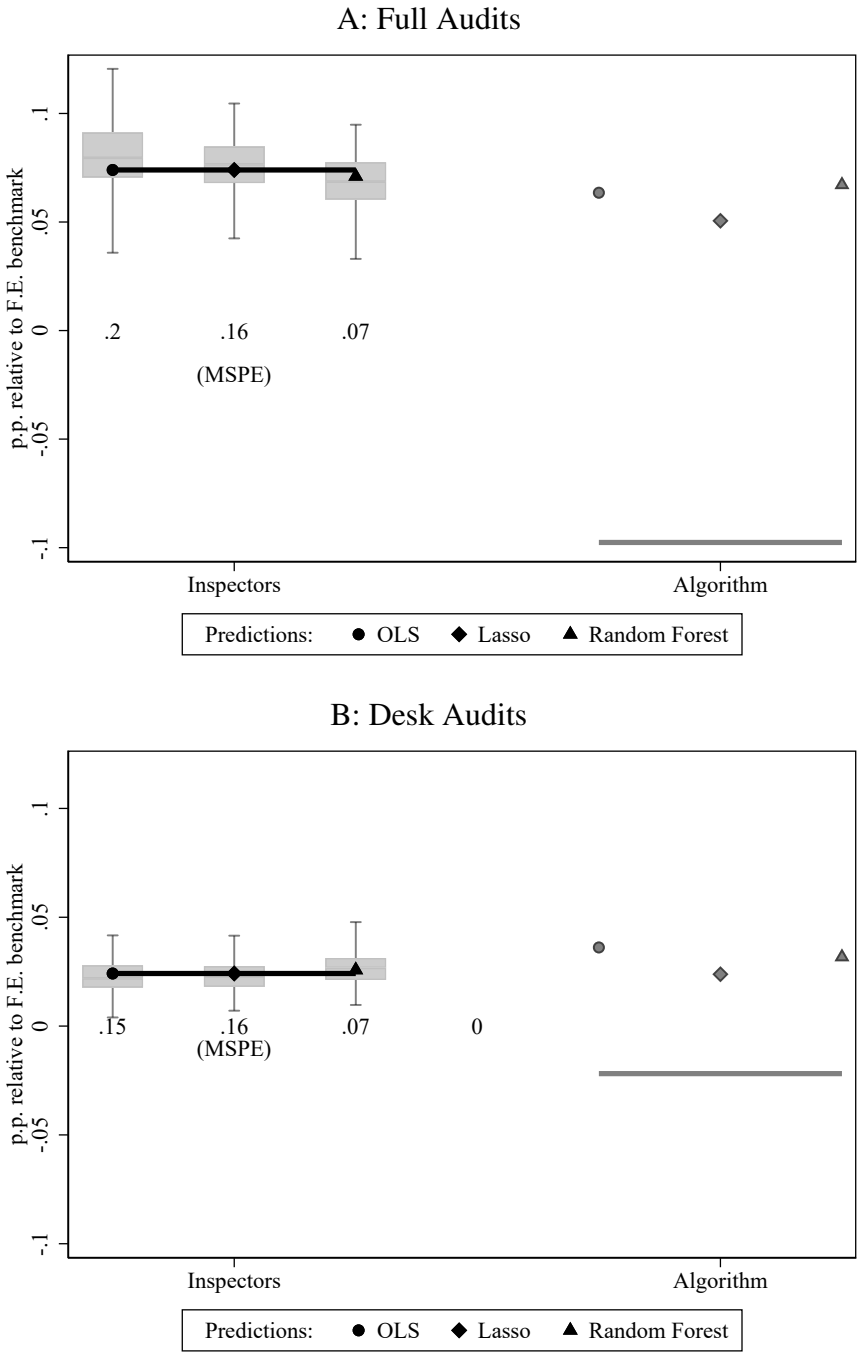
This table is similar to Table 7, but shows the association between audit outcomes and components of the risk score. This table is discussed in Section 6.1.

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*	-0.19	-0.05	-0.08*	-0.19
	(0.04)	(0.05)	(0.22)	(0.04)	(0.05)	(0.22)
Algorithm x Random	-0.00	-0.01	-0.29	-0.00	-0.01	-0.29
	(0.03)	(0.04)	(0.19)	(0.03)	(0.04)	(0.19)
Information	0.02	-0.04	-0.24			
	(0.04)	(0.04)	(0.20)			
Algorithm x Information	0.01	0.07	0.28			
	(0.04)	(0.06)	(0.26)			
Info. (indicators)				0.03	-0.05	-0.09
				(0.04)	(0.05)	(0.23)
Info. (indicators+data)				0.01	-0.04	-0.37
				(0.04)	(0.05)	(0.24)
Alg. x Info. (Indicators)				0.02	0.07	0.04
				(0.05)	(0.07)	(0.31)
Alg. x Info. (Indicators+data)				0.00	0.07	0.49
				(0.05)	(0.06)	(0.31)
N	2136	892	628	2136	892	628
R2	0.32	0.41	0.42	0.32	0.41	0.43
Mean outcome	0.42	0.70	17.68	0.42	0.70	17.68

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.

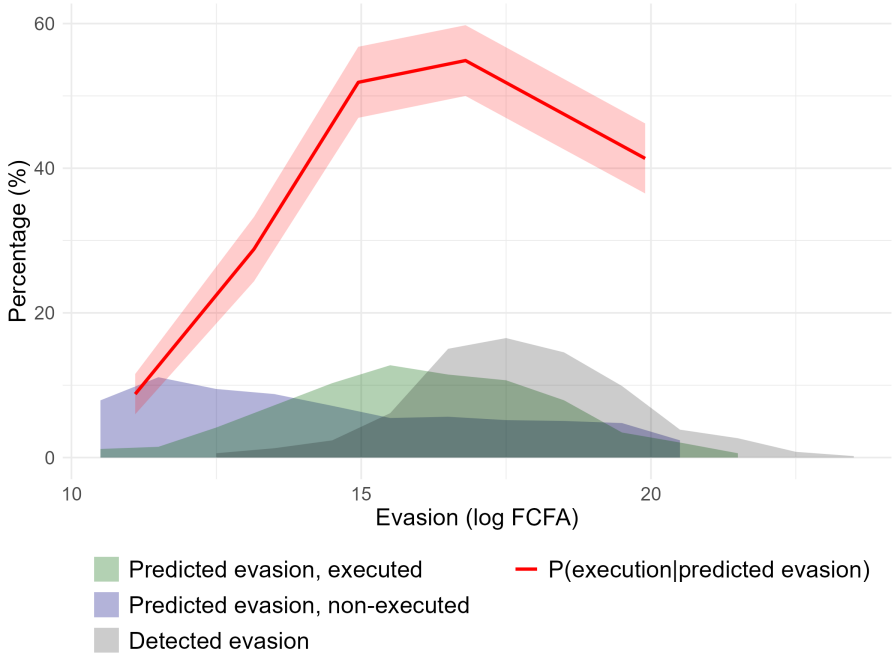
Figure F.1: Predicting Audit Execution Using Observable Characteristics



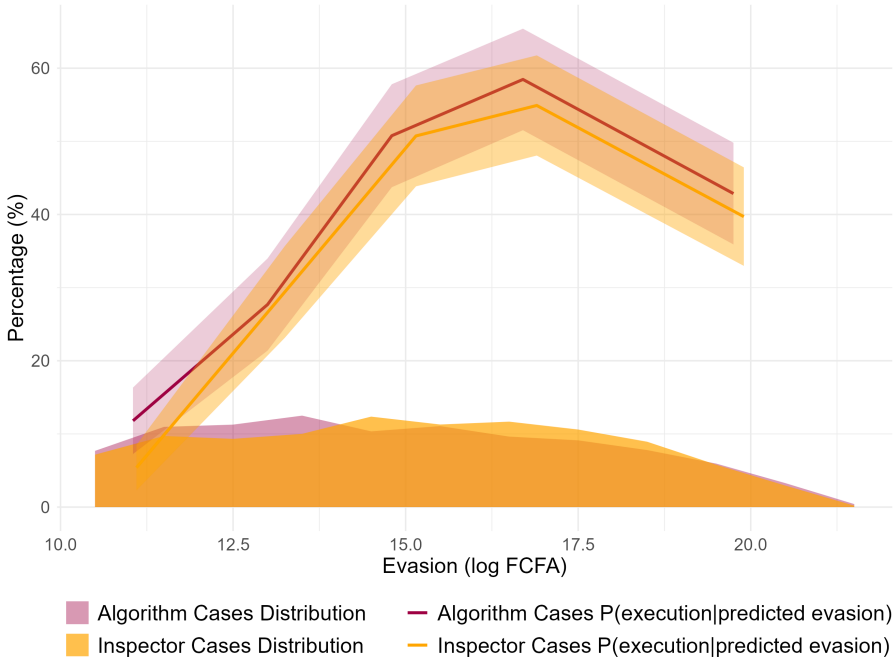
Notes: This figure shows the predicted mean execution of audits based on observable characteristics of the firms. Several models are estimated using only inspector-selected firms, and the model is used to predict the execution rate of algorithm cases based on the firms' characteristics. The difference 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.

Figure F.2: Predicted Evasion and Audit Execution, Full Audits

A: Predicted Evasion Amount and Realized Execution Rate, All Cases

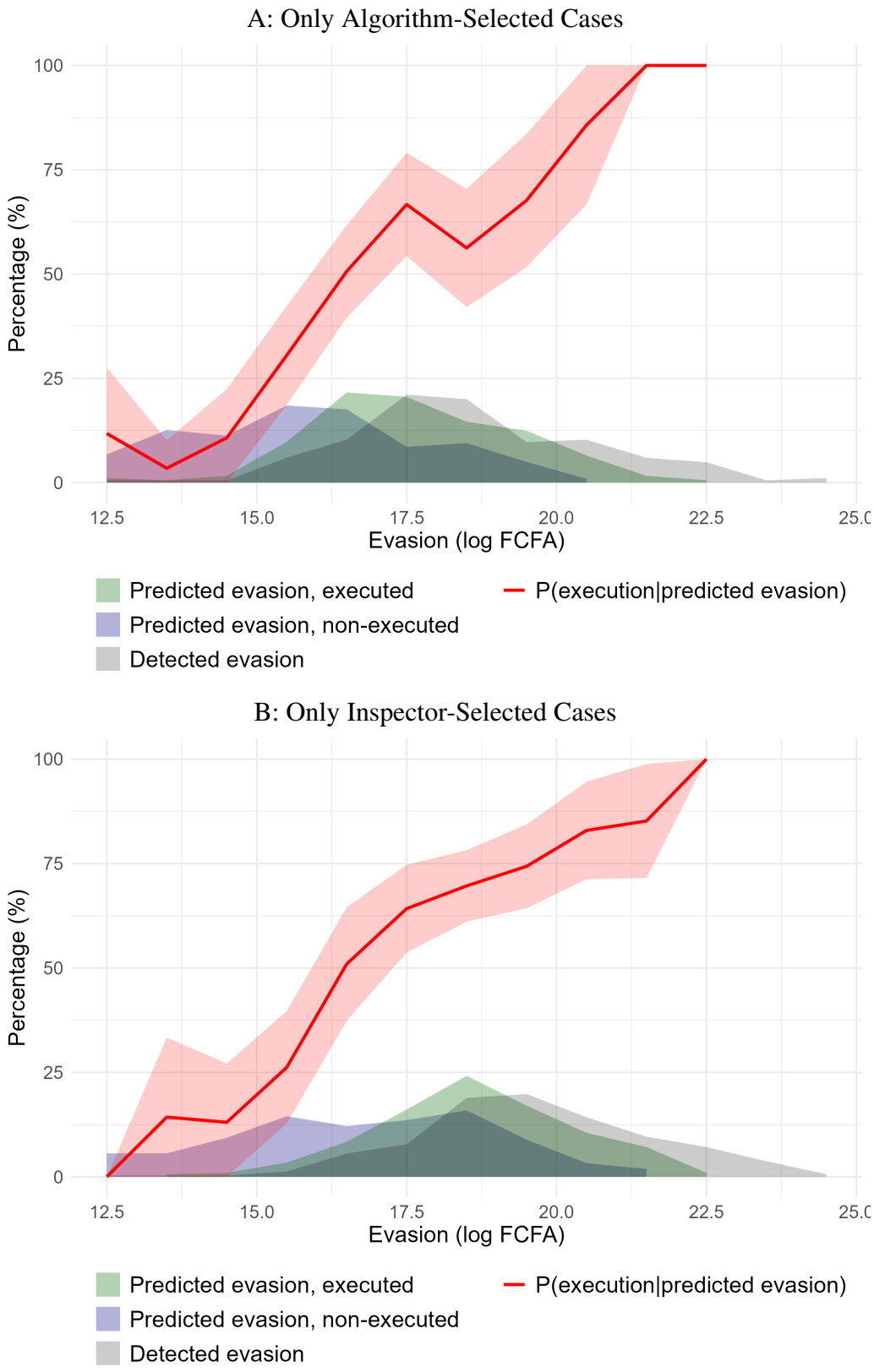


B: Realized Execution Rate for Algorithm Cases and Inspector Cases



Notes: A plots the distribution of detected evasion for all full audits (log of FCFA amounts, red), as well as the results from a random forest model predicting detected evasion based on firms' observable characteristics. The blue distribution is predicted evasion for cases that were part of the annual audit program but not actually audited. The green distribution is predicted evasion for cases that were audited. The black solid line plots the realized audit execution rate by level of predicted evasion. B shows the distribution of predicted evasion for all algorithm-selected cases (orange) and for all inspector-selected cases (purple). The solid lines (orange and purple) plot the realized audit execution rates by level of predicted evasion. This figure is for full audits only. This figure is similar to Figure 4, but focuses on desk audits.

Figure F.3: Predicted and Realized Detected Evasion Amount



Notes: This figure is similar to Figure 4 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.

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

	Algorithm (1)	Inspector (2)
L1 turnover	-0.16**	-0.21***
	(0.08)	(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	-0.03	0.15
	(0.20)	(0.16)
L1 profit rate sq	-0.09	-0.07
	(0.11)	(0.13)
L1 profit rate cu	-0.08	-0.18
	(0.24)	(0.22)
L1 productivity	0.05	0.05
	(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	0.19	0.65***
	(0.30)	(0.22)
Distance min	-0.10	-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	68	58	73	45	64	47
CME1	86	84	86	64	65	69
CME2	81	79	80	75	74	81
CPR	73	57	88	44	44	51
DSF	53	50	63	61	68	62

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: 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.

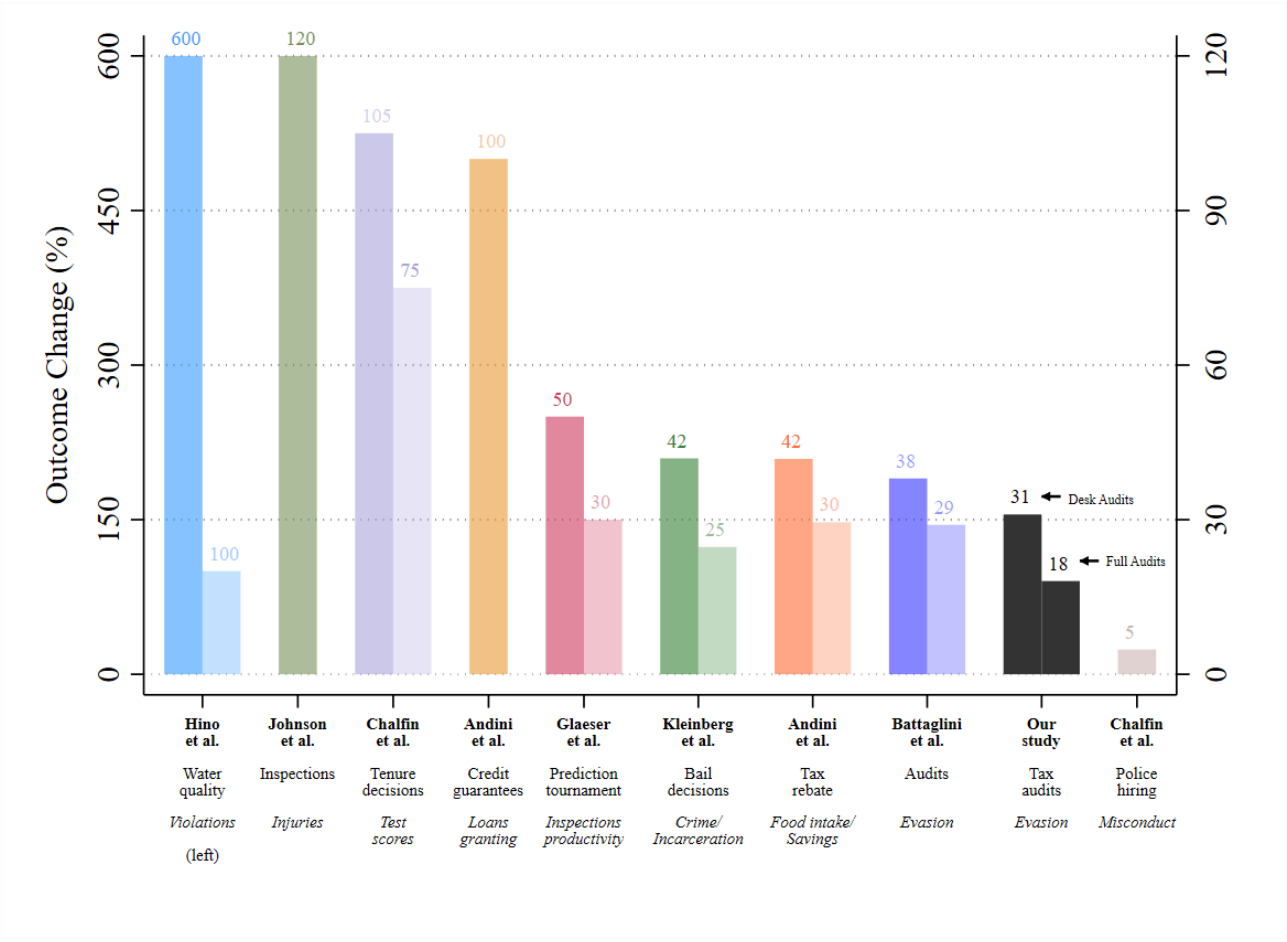
Notes: This table summarizes the literature on how machine learning algorithms can enhance or substitute for the work of government officials. This table is mentioned in Section 1.1.

Table G.2: Optimization Gains from Machine-Learning Algorithm Trained on Audit Outcome Data, Results Disaggregated by Tax Office, for Full Audits

Realized Revenue Log(mean)	Predicted Revenue Log(mean)	Δ Revenue vs Predicted w/ RF Selection Among Program Cases	Overlap Between Optimized and Realized Audit Program
(1)	(2)	(3)	(4)
A: Large Taxpayer Office			
21.92	21.82	+ 19.17%	60%
B: Medium Taxpayer Office			
19.87	19.87	+ 10.75%	77%
C: Liberal Professions Office			
20.66	20.46	+ 10.3%	62%
D: Small Taxpayer Office			
18.38	18.30	+ 91.15%	30%

Notes: This table is similar to Table 9, A, but shows the results disaggregated by tax office.

Figure G.1: Comparison of Optimized Results With Literature



Notes: This figure depicts the estimated improvements in outcomes in several papers that use machine learning to automate decisions in several settings. The outcomes and study designs of each of these other papers are explained in Appendix Table G.1. The results of our study are shown in black bars, and refer to the estimated increases in detected evasion according to the optimized execution of cases for full and desk audits, as discussed in Table 9. This figure is discussed in Section 6.4.

Table G.3: Performance of Random Forest Prediction

A: Full Audits				
	A1: Predicting Detection | Execution		A2: Predicting Evasion Amount | Detection	
	Precision	Recall	MSPE	R^2
	(1)	(2)	(3)	(4)
Overall OOS	0.88	1	2.18	0.35
Overall OOB	0.90	1	3.33	0.25
2018	0.92	1	2.47	0.31
2019	0.81	1	1.93	0.43
2020	0.94	1	2.27	0.09

B: Desk Audits				
	B1: Predicting Detection | Execution		B2: Predicting Evasion Amount | Detection	
	Precision	Recall	MSPE	R^2
	(1)	(2)	(3)	(4)
Overall OOS	0.80	0.93	2.38	0.24
Overall OOB	0.79	0.94	2.51	0.24
2018	0.91	0.95	2.94	0.31
2019	0.66	0.89	2.10	0.18
2020	1.00	0.98	2.07	0.05

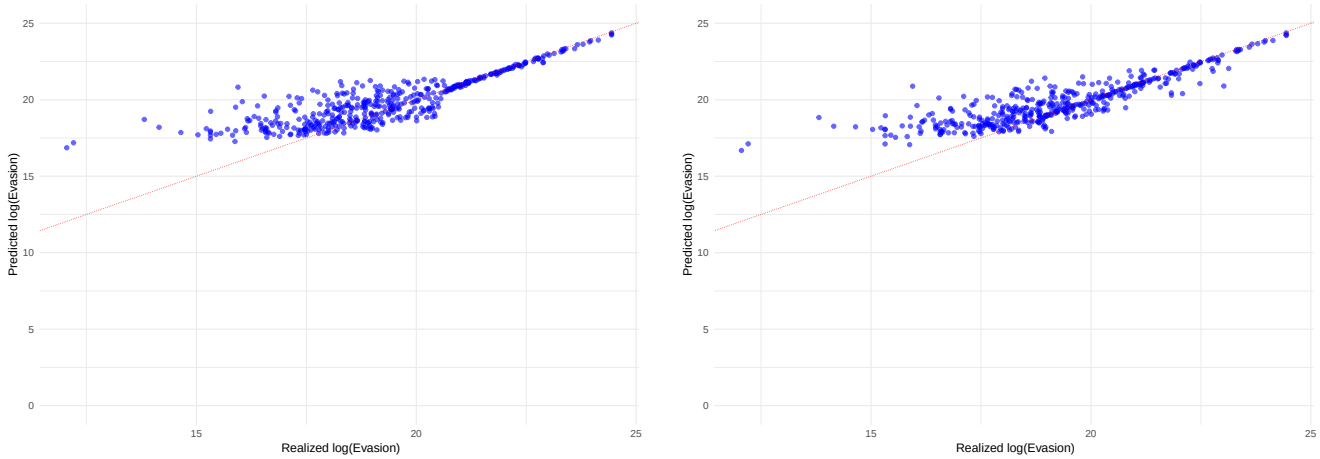
Notes: This table documents the performance of our random forest algorithm for predicting detection of evasion conditional on audit execution (panels A1 and B1) and predicting the amount of evasion conditional on detection (panels A2 and B2). The performance metrics are either calculated out-of-sample (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.2 further evaluates the accuracy of predicted evasion compared to realized evasion. This table is mentioned in Section 7.

Figure G.2: Within-Sample Model Fit: Comparison of Predicted and Realized Evasion

A: Full Audits

A1: Top Quartile Weights

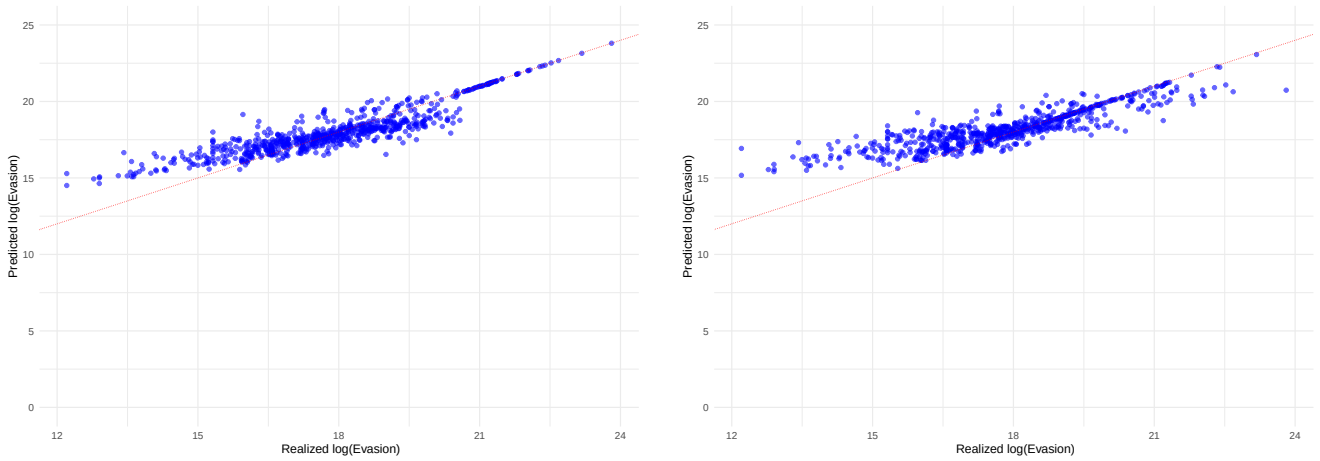
A2: Top Quartile Weights Within Office



B: Desk Audits

B1: Top Quartile Weights

B2: Top Quartile Weights Within Office

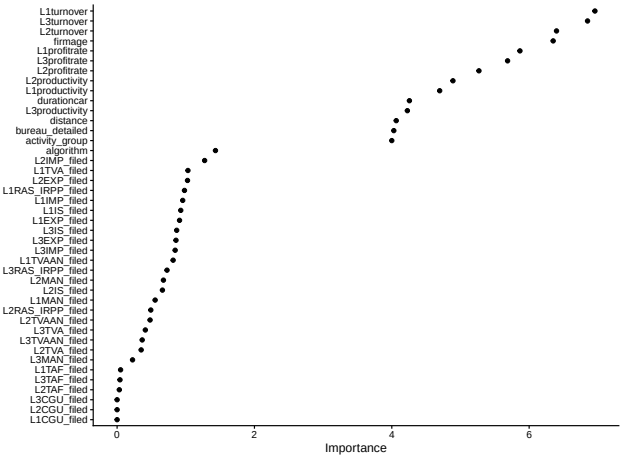


Notes: This figure shows scatterplots of the predicted and realized evasion, for full audits (A) and for desk audits (B), conditional on predicted evasion being non-zero. In each panel, we show results for a model where we increase the weight ten-fold on all cases in the top quartile of realized evasion (left column), or on all cases in the top quartile of realized evasion within each tax office (right column). Table G.3 shows other performance metrics of the prediction. This figure is mentioned in Section 7.

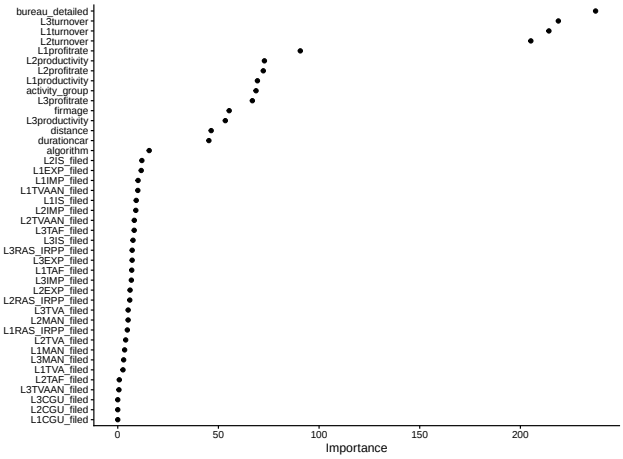
Figure G.3: Importance Ranking of Predictors in Random Forest Algorithm

A: Full Audits

A1: Predicting Detection

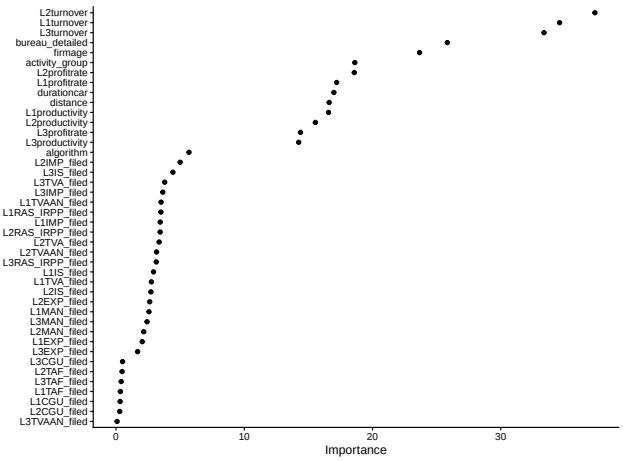


A2: Predicting Evasion Amount | Detection

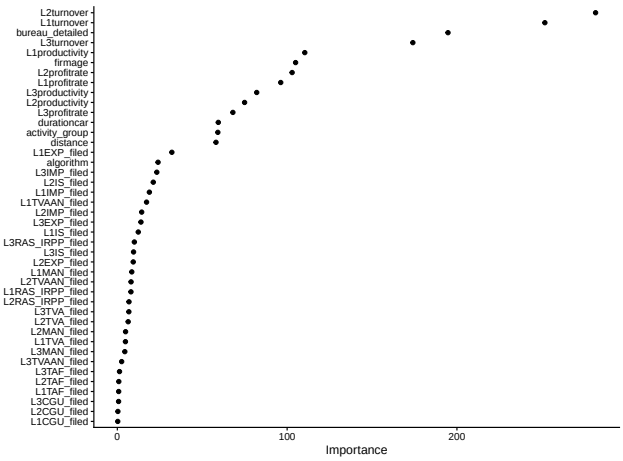


B: Desk Audits

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 (A) and desk audits (B). This figure is mentioned in Section 7.

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. OA44

Alm, James (2019) "What motivates tax compliance?" *Journal of economic surveys*, 33 (2), 353–388. OA44

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. OA53

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. OA53

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. OA53

Bergolo, Marcelo, Rodrigo Ceni, Guillermo Cruces, Matias Giaccobasso, and Ricardo Perez-Truglia (2023) "Tax audits as scarecrows: Evidence from a large-scale field experiment," *American Economic Journal: Economic Policy*, 15 (1), 110–153. OA39, OA44

Best, Michael, Jawad Shah, and Mazhar Waseem (2021) "Detection without deterrence: Long-run effects of tax audit on firm behavior," *Working Paper*. OA44

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. OA53

DAgosto, Elena, Marco Manzo, Stefano Pisani, and Filippo Maria DArcangelo (2018) "The Effect of Audit Activity on Tax Declaration: Evidence on Small Businesses in Italy," *Public Finance Review*, 46 (1), 29–57. OA44

Dwenger, Nadja, Henrik Kleven, Imran Rasul, and Johannes Rincke (2016) "Extrinsic and intrinsic motivations for tax compliance: Evidence from a field experiment in Germany," *American Economic Journal: Economic Policy*, 8 (3), 203–232. OA44

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. OA53

Hino, M., E. Benami, and N. Brooks (2018) "Machine Learning for Environmental Monitoring," *Nature Sustainability*, 1, 583–588. OA53

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. OA53

Khwaja, Muwer Sultan, Rajul Awasthi, and Jan Loeprick (2011) *Risk-based tax audits: approaches and country experiences*: The World Bank. OA3

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein (2018) "Discrimination in the Age of Algorithms," *Journal of Legal Analysis*, 10, 113–174. OA53

Naritomi, Joana (2019) "Consumers as tax auditors," *American Economic Review*, 109 (9), 3031–72. OA44

Paradisi, Matteo and Elia Sartori (2024) "Using Prediction Models to Design Tax Enforcement: Incentives vs Targeting," *CEPR Working Paper*. OA39

Pomeranz, Dina (2015) "No Taxation Without Information: Deterrence and Self-Enforcement in the Value Added Tax," *American Economic Review*, 105 (8). OA44