EDI WORKING PAPER SERIES WP/2022

HOW TO TARGET ENFORCEMENT AT SCALE? EVIDENCE FROM TAX AUDITS FROM SENEGAL

Pierre Bachas
World Bank & IFS
Anne Brockmeyer
IFS, UCL, World Bank & CEPR
Alipio Ferreira
TSE & IFS
Bassirou Sarr
EHESS

January 2022

Abstract

Developing economies are characterized by limited compliance with government regulation, such as taxation. Resources for enforcement are scarce, but the increasing availability of digitized data and data processing technologies have the potential to improve the targeting of enforcement. Leveraging an experiment at scale in Senegal, we compare the yield of tax audit cases selected by a risk-scoring algorithm to cases selected by tax inspectors based on a traditional discretionary procedure. The algorithm computed indicators of inconsistencies and anomalies based on available information about firms, including their own tax declarations and third party data. Discretionary methods select larger firms than the algorithm, and uncover equivalent evasion rates, thus outperforming it in absolute values of fines.

*Pierre Bachas: World Bank Research, pbachas@worldbank.org; Anne Brockmeyer: Institute for Fiscal Studies, anne.brockmeyer@ifs.org.uk; Alipio Ferreira: Toulouse School of Economics, alipio.ferreira@tse-fr.eu; Bassirou Sarr: Ecole des Hautes Etudes en Sciences Sociales (EHESS). We thank the Senegal Tax Administration (DGID) for an outstanding collaboration, in particular, Bassirou Samba Niasse, Amadou Abdoulaye Badiane, Oumar Diop Diagne, Hady Dieye, Mor Fall, Serigne Mabaye Fall and Mathiam Thioub. We thank Samba Mbaye, Assane Sylla and Medoune Sall from the CRDES, for the excellent partnership in the taxpayer survey. We also thank Denis Cogneau, Laurent Corthay, Janet Jiang, Nicola Limodio, Jan Loeprick, Markus Kitzmuller and Dan Rogger for helpful comments and discussions, and Oumy Thiandoum for excellent research assistance. Finally, we gratefully acknowledge the administrative support of the Paris School of Economics and CEPREMAP, and the financial support from the UK government through the Economic Development and Institutions Initiative (EDI) and the Centre for Tax Analysis in Developing Countries (TaxDev). The findings, interpretations, and conclusions are entirely those of the authors. They do not necessarily represent the views of the World Bank, its affiliated organizations, its Executive Directors or the governments they represent, nor the Senegal Tax Administration.

About Economic Development & Institutions

Institutions matter for growth and inclusive development. But despite increasing awareness of the importance of institutions on economic outcomes, there is little evidence on how positive institutional change can be achieved. The Economic Development and Institutions – EDI – research programme aims to fill this knowledge gap by working with some of the finest economic thinkers and social scientists across the globe.

The programme was launched in 2015 and will run until March 2022. It is made up of four parallel research activities: path-finding papers, institutional diagnostic, coordinated randomised control trials, and case studies. The programme is funded with UK aid from the UK government. For more information see http://edi.opml.co.uk.

UKaid
from the British people

1 Introduction

To ensure compliance with regulations, governments allocate scarce resources towards enforcement activities. In low-income countries, bureaucrats at enforcement agencies hold a significant degree of discretion over the inspection strategy. Discretion over inspection choice may be optimal if the bureaucrats' personal experience and soft information enable them to select inspections effectively. However, data-driven methods to select audits may reduce human errors and make enforcement more accountable. Although digitized data is increasingly available, many administrations do not take advantage of it systematically but rather on an ad-hoc basis. Can the systematic use of data in settings with low state capacity improve enforcement activities?

In this paper, we study the implementation at scale of a data-driven selection of firm tax audits in Senegal and compare it to the discretionary selection method in use. Until 2017, inspectors selected all audit cases in Senegal with a discretionary procedure. From 2018 onwards, we have started collaborating with the Senegalese tax administration to select part of the audits program via a data-driven method based on risk indicators. We compare the two selection methods across several dimensions: the characteristics of selected firms, the amount of recovered taxes in the inspections, and the quality of the inspections.

The research team compiled a large dataset of self-reported tax declarations with third-party information from customs, procurement contracts, and transacting partners. We then used the dataset to construct two sets of risk indicators: discrepancy indicators and anomaly indicators. Discrepancy indicators are intra-firm comparisons between tax declarations and third-party data. In contrast, anomalies indicators are inter-firm comparisons revealing outlying reporting behaviors. We then ranked firms within comparable groups and assigned them scores, which we used to select firms into the audit program.

We conducted the experiment at scale, intervening in approximately half of the audit program in the participating tax centers.[1] The experiment included the two types of audits in Senegal: in-person *full audits*, carried out by groups of inspectors, and desk audits, carried out individually by inspectors from their offices. Each tax unit selected half of the cases planned for the *full (in-person) audit* program, and the risk-score assigned the remaining

[1]Selected firms represented 24% of corporate tax revenue of the tax centers in the experiment. The total amount of corporate tax liability (VAT and CIT) over the years 2015-2018 for the tax centers used was around 315 billion FCFA, and the selected firms in the 2019 program accounted for 75 billion FCFA.), and implemented directly by the audit planning and intelligence division of the tax administration.

half. Moreover, each inspector selected 45% of her *desk audit* program, 45% came from the risk-score, and the remaining 10% were selected randomly. Moreover, we cross-randomized an *information treatment* for desk audits across the three selection methods. Information-treated cases received information on the most significant compliance risks detected by the risk score and detailed data from third parties regarding that taxpayer. The information treatment facilitates data access and analysis, thus potentially easing inspectors' work. We submitted the experiment, hypothesis, and specifications to the AEA registry.

We analyzed how the audit selection methods differ regarding the type of selected firm, the completion rates, and the uncovered tax evasion. On average, discretionary method cases report 50-80% larger revenues than risk score-selected firms and higher profits. Second, we find inspectors were four percentage points less likely to start risk-based cases than discretionary cases. Among started audits, risk score cases were five percentage points less likely to uncover evasion and require a payment from the taxpayer. In absolute terms or as a percentage of the firm's turnover, the required payments were statistically equivalent across the discretionary and risk-based selection methods. Finally, the information treatment yielded no significant improvement in audit yield, but it increased the probability of the audit being started.

The results are heterogeneous across the two types of audits. For example, the inspectors' reluctance to start risk-based audits was stronger for full audits than for short audits. Moreover, the results vary depending on how much data was available about the firms. The risk-based method was more likely to uncover evasion for firms for which third-party data is available (such as customs data and treasury payments) than the discretionary methods and less likely for firms without third-party information.

To construct the risk score, we applied the best international practices after ample consultation. We did not rely on machine-learning tools and fine-tuned parametrization, preferring an explicit parametric formula with indicators that the inspectors could easily understand. Moreover, limited data availability, particularly regarding historical audit results, reduced the scope for machine learning methods. The intervention only targeted the selection methods without changing career or monetary incentives for tax inspectors. Therefore, tax inspectors could choose to devote efforts asymmetrically across selection methods despite the pressure from their hierarchy to devote equal efforts. Changing the monetary incentives is not allowed legally. The institutional and technical constraints faced in this experiment are likely to represent a credible benchmark for other low-income countries considering introducing a transparent risk-based selection of audits at scale, in particular in West Africa, which often looks at Senegal for administrative innovations.

To our knowledge, this is the first study to rigorously evaluate the differences between audit selection methods for tax audits. The IMF and World Bank have long advocated risk-based algorithms for audit selection. However, we know no impact evaluation of the adoption of such algorithms.[2] In the risk-based approach that we propose, we emphasize the use of third-party information to flag potential evaders. The importance of third-party information to detect and deter evasion is emphasized in a growing literature (Pomeranz 2015, Kleven et al. 2011, Kleven et al. 2016 Naritomi 2019).[3] However, if third-party data is scarce, inspectors' private information may be valuable. In that case, the advantages of migrating towards a data-driven approach are mitigated. Our study provides a unique opportunity to exploit the benefits of adopting risk-based audit selection, a standard practice in developed and developing countries (Khwaja et al. 2011) but absent in Senegal until recently.

Our paper also contributes to answering the question about the value of discretion in audit selection (Duflo et al. 2018, Kang and Silveira 2021), particularly relevant in developing countries. In an experimental approach, we provide credible estimates of the differences of the risk-based approach relative to discretion. In contrast, Duflo et al. 2018 also propose an experimental approach but compare random audits to discretionary ones.

Finally, the study of selection methods is related to how tax administrations are run. Tax administrations have scarce resources, which they must allocate based on complex functions of policy objectives, political considerations, inspectors' incentives, and data about taxpayers. Efficient administrations are vital to building state capacity (Besley and Persson 2013), by aligning correctly the incentives of bureaucrats with those of the state (Xu 2019, Bertrand et al. 2018; Finan et al. 2017) or improving enforcement. Recent experimental evidence has shown that monetary incentives for tax inspectors improve the quality of inspections (Okunogbe and Pouliquen 2018) and increase revenues (Khan et al. 2015). In our Senegal experiment, we keep the incentives of inspectors fixed and change the selection

[2] According to Khwaja et al. 2011, for example, in the U.K., 55% of all cases are based on discretionary selection. In contrast, 35% and 10% of cases are respectively selected via a risk-scoring technique and a simple random sample. This approach is closest to the policy reform we introduce in Senegal. In other sub-Saharan African countries, Kenya uses a risk for all large taxpayers and discretionary selection for all others. Tanzania and Lesotho constitute examples on the extreme, respectively relying only on risk-scoring and random selection to audit all taxpayers.

[3] Recent papers study firms' behaviour as they get exposed to new third-party information trails and show that taxpayers substitute evasion to less verifiable margins (Carrillo et al. 2017, Slemrod et al. 2015).

methods for audits. Audit selection is also a relevant topic in fighting wastes and money diversion of government expenditures (Banerjee et al. 2020, Gerardino et al. 2020).

Overall our results point to a nuanced contribution of data-driven selection methods. On the one hand, the data-driven method allows for systematic and objective audit selection, sparing time and efforts at the selection process and execution stage. On the other hand, the data-driven method seems to select several firms for which inspectors fail to find evasion. Moreover, the risk-based method's failure to select firms with detectable tax evasion is more significant among firms with little third-party data, suggesting that the risk-score method relies heavily on broad data availability. For firms with limited data availability, such as firms that only declare one type of tax, the discretionary method performs better.

2 Institutional Setting: Senegal's Revenue Administration

2.1 Taxes in Senegal

Tax revenue represented on average 16.7% of GDP in Senegal between 2013 and 2019. These revenue collection levels are below the West African Economic and Monetary Union (WAEMU) target of 20%, and fall short of goals set in Senegal's own medium term expenditure strategy. Tax gap estimates indicate that 23% of the theoretical VAT revenue is not collected (a shortfall of 2% of GDP) and that close to 63% of theoretical receipts from income taxes are missing (approximately 7% of GDP).

Similar to other developing countries, most taxes in Senegal are remitted by large and medium companies (Slemrod et al. 2001). In particular, firms remit the Value Added Tax (VAT) and income taxes (Corporate income tax, personal income tax and dividend withholding taxes), accounting for 36% and 29% percentage of total tax revenue in 2019. Firms also withhold income taxes on their employees' wages (Pay-as-You-Earn), which is often the only source of reporting on salaried income, given the incompleteness of self-reported personal income taxes. Other significant revenue sources are customs duties (15%) and specific taxes on petroleum, which we do not cover in this study.

The Corporate Income Tax (CIT) is paid annually, at a rate of 30% profits or a 0.5% of turnover, whichever is larger. The Value Added Tax (VAT) is paid on a monthly basis, at a standard rate of 18% and a reduced rate of 10% for tourism businesses and hotels. A small number of financial sector firms pay the financial services tax instead of the VAT, also at a rate of 18%. Small firms with a yearly turnover of less than 50 million CFA Francs (about 100,000 USD) are eligible for a simplified tax (*Contribution globale unique*, CGU), which replaces all other taxes. The CGU is levied on turnover, at rates varying from 1% to 8%, where rates vary across sectors and increase in turnover. As already mentioned, the Pay-As-You-Earn taxes are withheld personal income tax on employees' wages with a formal employment contract.

2.2 The tax enforcement agency

The Direction Générale des Impôts et des Domaines (DGID) is the administrative body in charge of domestic tax collection and enforcement, and reports to the Ministry of Finance. Figure 1 displays DGID's organizational chart. The large taxpayer directorate oversees firms whose turnover is greater than equal to 3 billion CFA francs (approximately 5.3 million USD) and has four units, which are specialized by economic sectors.[4] The medium taxpayer directorate oversees firms with less than with turnover between 100 million CFA francs and 3 billion CFA francs, and has two units. A third unit is in charge of the regulated liberal professions such as lawyers, notaries and medical practitioners. The remaining taxpayers, mostly small and medium enterprises (SMEs), are assigned to one of 19 regional tax offices.

There are two principal types of audits: desk audits and full audits.[5] Desk audits (or short audits) are carried out by individual inspectors from within the tax authority's premises, using the firm's tax returns and, eventually, third-party data. Taxpayers are unaware of these audits unless inspectors make information requests, for example, when data is missing or seems inconsistent. Full audits are carried out by a team of inspectors at the taxpayer's premises. Full audits are announced at least five days before the audit starting date with an information request notice to the taxpayer. Tax inspectors may collect information for several weeks at the taxpayer's premises and continue requesting information for up to 12 months.[6]

[4] Unit 1 is in charge of the mining and energy sectors. Unit 2 deals with financial services and the telecommunications industry. Unit 3 covers real estate and firms. Unit 4 is a generalist one with broad competence covering all other sectors.

[5] There are also surprise audits which can take place either based on information that DGID receives either internally or from whistle-blowers. Surprise audits are similar to full audits, except that they are unannounced, as their name indicates.

[6] For firms with a turnover of less than 1 billion CFA francs (about 2 million USD), full audits can only last up to four months. These maximum limits are general rules. There may be extensions in cases with highly suspicious activity or when there is a delay in the transmittal of the requested information to auditors.

The selection method of tax audits in Senegal is essentially discretionary. Inspectors follow some rules of thumb, such as avoiding recently audited firms and firms with low turnover. However, there are no objective rules or formulas to add or drop firms from their selected program. Our study intervenes precisely at this stage of the tax administration's operation by including a machine-based selection in part of the audit program of the tax authority, both for short and full audits.

Figure 2 illustrates the steps in the audit process. After reviewing a case, inspectors list the detected irregularities and penalties and send them to the taxpayer in an "initial notice". They can also request additional information from the taxpayer. Upon receiving the initial notice, taxpayers have 30 days to respond to the inspector's findings.[7] The inspector examines the response has 60 days to prepare and send a "confirmation notice", again with the detected irregularities and penalties. The inspector then creates a revenue order for the tax collection unit, which requires the taxpayer to make a payment within ten business days. Taxpayers can appeal at the Minister of Finance or a judicial court, and the appeal may suspend the payment process temporarily.

3 Data

Our study draws on three sets of administrative data sources and two surveys. The three sets of administrative data are the tax declarations filed by taxpayers, third-party data on transactions, and audit outcomes. We discuss details of the matching process and match rates in Appendix C. We complement the administrative datasets with a taxpayer survey and a tax inspector survey, which were designed by the research team and were not available to the Senegalese tax authorities.

Tax Declarations. Table 6, Panel A, provides an overview of the available tax declarations. Our primary sources of information are the tax declarations on Corporate Income Tax, Value Added Tax, and the Pay-As-You-Earn tax (withheld progressive personal income tax), covering the period of 2014-2019. The CIT data covers about 4 thousand firms per year, and the VAT data around 8 thousand firms.[8] Finally, we match these data with monthly Pay-As-You-Earn data, which allows us to calculate the number of employees and the aggregate wage bill for each firm.

[7]If the taxpayer fails to respond, it means for legal purposes that they agree with the inspector's findings.

[8]Many more firms declare VAT than CIT because self-employed individuals and unincorporated firms file VAT but not CIT.

Third-Party Data. Table 6, Panel B, describes the third-party data, that is, information about transactions of companies which we obtain from third parties. The third-party datasets are the import-export transactions (customs data), payments from state institutions to firms (procurement data), and in recent years VAT annexes documenting transactions between firms.[9] These datasets are at the transaction level, and we aggregate them at the firm-year level to merge with the tax data. As the last two columns in Table 6 indicate, a non-negligible share of firms captured in the third-party data fail to file taxes in the corresponding year. The share of taxpayers for whom third-party data is available hovers around 28%, with the share increasing over time and in firm size.

Audits data. We collect selected audit programs and audit results data for fiscal years 2018, 2019, and 2020. The selected audit programs are partly produced by the risk-scoring algorithm, and the rest by the inspectors themselves. The audit results contain information on key audit process steps: audit announcement, notification, confirmation, and payment request. The audit results data contains several ad hoc audits which were carried out despite not being initially programmed. The data contain the inspector's name, taxes verified in the audit, infractions detected, evaded amounts, applicable penalties, and the dates of each step. We use this information to compute our outcomes, such as audit yield and evasion rates. Moreover, we asked inspectors to fill in spreadsheets with qualitative information about each audit case, such as the perceived difficulty of the audit, whether the taxpayer was uncooperative, the business activities were complex, or information was unavailable.

Tax Inspector Survey. Prior to our intervention, we conducted a detailed survey among all participating tax inspectors, capturing information about their demographics, employment history, perceptions of the audit function, methods for audit selection, and use of different sources of information. The survey data contain 97 inspectors, which covers most inspectors involved in audits in 2018-2020.

Taxpayer Survey. We surveyed approximately 750 firms in the Dakar region, most of which had been audited shortly before. We conducted the taxpayer survey in two waves, from October to December 2020 and March to May 2021. The survey allowed us to elicit taxpayers' perspectives on tax inspections, audit risk, and their opinions on the tax authority.

[9] VAT annexes have become increasingly available in recent years, following efforts by the tax administration to digitize information and require that taxpayers file their VAT annexes electronically.

4 Audit Selection and Experimental Design

4.1 Discretionary Selection

Until 2018, all audit cases in Senegal were selected exclusively with a discretionary procedure. At the beginning of the year, the Director-general of the tax authority requests each unit to propose the annual program of firm audits. Each unit suggests a set of full audits and desk audits, the latter suggested by the inspectors that will conduct them individually.[10]

Tax inspectors use a standardized form to motivate the full audit selection. The form contains information on the identity of the selected firm, past audit history, and a summary of relevant indicators such as tax turnover and profit margin. Once the tax unit's manager approves the form, a selection committee in the Director-general's office finalizes the list of firms for the full audit program. The committee accepts most proposed cases, though the committee may request additional information, reject proposals, or add their proposals based, for example, on denunciations. The committee then returns the names of approved audits to tax units.[11]The selection of short audits also takes place at the beginning of the year, but the procedure is simpler than for full audits. Individual inspectors propose cases to their tax unit's director without any particular guideline.

4.2 Risk-Score Method

In the past decade, the Senegalese tax administration has invested in digitizing its tax data, widening the availability of information about its taxpayers and creating the opportunity to select audits selection in a data-driven way. The cooperation between the researchers and the tax authority started in 2017, first by mapping available data sources and indicators that could be useful to assess compliance risk. We designed a risk-scoring tool based on a set of indicators, drawing on work by the World Bank (tax administration projects in Pakistan and Turkey), SKAT in Denmark, and the IMF's recommendations to Senegal.

We designed an algorithm based on intuitive indicators, which we discussed and explained to the tax authority staff. We preferred this method rather than a machine-learning tool, which would yield a less transparent selection. There are two reasons for preferring an

[10]Since desk audits are selected individually, different inspectors might select the same taxpayer; in practice, this is rare as inspectors specialize by economic sectors or geographical areas. When this happens, the manager presumably rules which inspector is in charge of the case.

[11]This description is based on interviews with members of the committee.

indicator-based parametric algorithm to a nonparametric machine-learning algorithm. First, we needed a simple and transparent tool that would easily convey the identified compliance risks associated with a firm to tax inspectors. Second, the available data on historical digitized audit results was sparse, limiting the scope for model training and prediction of tax evasion.[12] Our proposed risk-score tool is a transparent risk assessment based on international best-practice, designed in cooperation and dialogue with the tax authority taking into account their capacity constraints. The constraints faced by DGID are likely to bind in many low-income countries, especially in West Africa, which often looks at Senegal for administrative innovations.

Table 1 summarizes the seven critical steps in the design of the risk score algorithm. Step (1) corresponded to the construction of a database covering all tax declarations across years and merged with third-party reported sources, as discussed in section 3. Steps (2) and (3) determined the risk indicators based on intra-firm discrepancies across data sources and inter-firm anomalies based on comparisons with similar firms. Step (4) defined the peer-group comparison clusters, defined by economic activity and tax center. Step (5) assigned a numerical value to each risk indicator, depending on the size of the inconsistency or anomaly (with higher scores for larger discrepancies). Step (6) assigned weights to each indicator, reflecting our judgment about their relative importance. Finally, step (7) aggregated the weighted indicators over the past four fiscal years to form a single risk score.

As already mentioned, the risk score relies on two types of risk indicators: discrepancies and anomalies. Discrepancies are intra-firm indicators, which flag taxpayers with inconsistent information across different datasets. For example, a discrepancy arises if the self-reported turnover is inferior to what we can expect from reading customs data, state procurement, and transacting partners. In contrast, anomalies indicators are inter-firm indicators, which compare a firm to a group of similar peers. An example is a firm with an abnormally low margin of profits relative to its peers. Firms were given a higher risk score for all indicators depending on how severe the irregularity seemed to be. In the last iteration of the algorithm, we included four discrepancy indicators and six anomaly indicators to construct the risk score. We over-weighted the discrepancies compared to anomalies to reflect the higher confidence that discrepancies reflect non-compliance, while anomalies might only reflect temporary economic problems or poor management.

[12] In the early stages of the design, we implemented a random-forest algorithm to predict evasion, which predicted historical evasion with similar degrees of accuracy as the parametric indicators, but which was far less easy to manipulate and interpret.

4.3 Study Design

The intervention changed the selection method for audits at the tax authority by selecting part of the audit program with a data-driven risk-based algorithm. A small portion of the program for short audits was also selected entirely randomly. Figure 3 illustrates the timeline of the design and case selection. The selection of the audit program proceeded in three main steps. First, inspectors selected cases at their discretion and submitted them to their hierarchy. Second, we ranked firms based on these computed risk scores within each tax center and selected a pre-agreed number for audits (usually the same number of firms selected via the discretionary method). [13] In the case of desk audits we also selected some firms at random. Third, a committee within the tax administration reviewed all the selections, excluding some firms that had been recently audited and a few state-owned companies. The approved lists were sent to the tax centers and individual inspectors.[14]

In summary, 50% of full audits were selected by the discretionary method (that is, by inspectors) and 50% by the risk score algorithm. In contrast, 40% of short audits were selected by the discretionary method, 40% by the algorithm, and 20% at random. The exact number of cases varies by tax center as displayed in Tables 2 and 3.

We also proposed sequencing of audits to induce inspectors to carry out the algorithm and discretionary cases in an alternated manner.[15] Inspectors knew which cases were algorithm-selected and which ones were selected by the discretionary method (either themselves or their hierarchy).[16] We also shared with inspectors a methodological note containing the indicators used in the algorithm. We presented the algorithm at a workshop organized by the intelligence unit of the tax authority in Dakar.

[13] In case of overlap between algorithm and discretionary methods, we assign the next highest score until we meet the pre-agreed number of algorithm cases.

[14] The Director General's office informed inspectors about the experiment, urging them to follow guidelines in carrying out audits at the proposed sequencing and reporting audit results rigorously. This complements presentations by the intelligence unit of DGID and the research team to each center.

[15] We randomly ordered each tax inspector's list of cases. This ensures that the order of audits is uncorrelated with audit quality. However, we were unable to ensure discipline in following the designated sequence for the workload. For instance, inspectors could choose to prioritize cases they select themselves and which they believe could leave to higher yield. Nonetheless, the Director General signed a guideline urging staff to follow the sequence set in their assignments.

[16] We marked the non-discretionary cases as "New methods" in the inspectors' spreadsheets. Therefore they could not distinguish between cases selected by the risk-score and cases selected at random.

Finally, we included a supplementary information treatment for desk audits. For a random sample of desk audits, regardless of their selection method, we provided inspectors with a readable version of the combined datasets of the selected firm (information treatment 1). For another random sample, we provided inspectors with the list of three main risk indicators flagged for the selected firm. By providing the data in a spreadsheet, we ease the efforts that they would have used to work on a case, compared to control cases.

5 Results

We study how the selection methods in three dimensions: i) the characteristics of selected firms, such as their size and profitability, ii) the probability that inspectors analyze the selected case, and iii) the probability that the inspectors detect irregularities during the audit. To examine the effect of the selection method on audit implementation and results, we estimate the following model:

$$y_{iot} = \beta_0 + \beta_1 Algorithm_{iot} + \beta_2 Random_{iot} + \beta_3 AdHoc_{iot} + \delta_t + \gamma_o + \varepsilon_{iot} \tag{1}$$

Where y_{iot} is the outcome of an audit for case i, registered in tax office o and selected for audit in year t, and ε_{iot} is a conditional mean zero error term. $Algorithm$ is an indicator function that is equal to 1 if the firm was selected for audit by the algorithm, and 0 otherwise. We define in a similar way $Random$ for randomly selected cases (which applies only for desk audits), and $AdHoc$ for audits that were carried out despite not being initially programmed. Finally, δ_t are year fixed effects, and γ_o tax office fixed effects. We estimate the model by ordinary least squares.

The experiment intervened at the selection process, and the intervention level is a case i in the audits program, which is filled by a firm according to a selection method. This is not an intervention to assess behavioral changes in firms, but to assess the ability of methods to find firms with certain characteristics. The behavior of a firm is fixed, and the counterfactual observation for an algorithm-selected firms is the discretionary selection that would have taken place in its place.

Using the potential outcomes notation, we can write the outcome y as a function of the selected method: $y(A)$ for algorithm and $y(D)$ for discretionary, with y_{iot} being the observed outcome for case i. The main object of interest is the average effect on the audits program of selecting firms via the algorithm relative to selecting them in a discretionary manner. That is, $ATE = \mathbb{E}[y(A) - y(D)]$. Since we selected the same number of cases for the algorithm as the number of discretionary cases (in most instances), we can guarantee that the selected cases by each method are the most preferred cases according to that method. This ensures that $\mathbb{E}[y(A)|Algorithm = 1] = \mathbb{E}[y(A)|Algorithm = 0]$ and $\mathbb{E}[y(D)|Algorithm = 1] = \mathbb{E}[y(D)|Algorithm = 0]$. We can identify ATE by computing averages from the observed sample, or estimating β_1 by ordinary least squares.

We also run some specifications containing controls for the presence third party data (at the firm level), and inspector fixed effects. We run most specifications separately for full audits and desk audits. Whenever they are pooled together, we add a dummy variable indicating whether the audit is a full audit.

5.1 Differences in characteristics of algorithm and discretionary cases

The algorithm cases are markedly different from the discretionary case concerning the size of selected firms. Table 7 summarizes pre-audit firm characteristics across different selection methods: randomly selected firms, risk-score selected firms, and tax authority selected firms.[17] The table shows that the randomly selected firms present similar averages as the population of firms for their declared turnover, profits, profit rates, and tax liability over the period 2015-2018. Even though the number of randomly selected firms was low (around 10% of the audits program), it is a somewhat representative group of the firms registered at the tax centers under analysis.

Firms selected by the algorithm and by DGID, on the other hand, are very different from the population's average (or the randomly selected firms'). Selected firms in both cases are larger in terms of their declared turnover, profits, and tax liability. Moreover, as can be seen in table 7, DGID selected firms with substantially larger declared turnover than the risk score algorithm (53% larger on average), larger profits (though not significant), and larger profit rates. The firms in the two selection methods present a similar amount of tax liability. The risk score is larger for the algorithm-selected firms, which is unsurprising since the risk-score selection explicitly picks the firms with the largest values of this variable.

The main difference is the declared turnover by firms. The discretionary method is more likely to select firms with large declared sales, even though this typically means selecting

[17] Very few firms were selected both by the tax authority and the algorithm (remember that the discretionary selection happened before the discretionary selection). In the comparisons of table 7, we disregard these firms, and we include a dummy to control for these cases in the regressions tables.

firms already paying high levels of taxes. On the other hand, the algorithm considers low turnover declarations as a risk factor and attributes a higher probability of selecting those firms.

5.2 Effect of selection method on audit execution

We here analyze the probability of inspectors starting case i. We observe that inspectors started a case when they filled out at least one key information regarding the audit process, such as an information request to the taxpayer. Inspectors know the selection method of each case and possibly take that into account in deciding to start a case. If discretionary cases are easier to conduct or have a higher expected return, inspectors may be reluctant to open algorithm-selected cases.

Table 5 examines this hypothesis. The outcome is a dummy variable that takes value 1 if inspectors opened the case and 0 otherwise. Each column shows a linear probability model, and the coefficient on $Algorithm$ shows the difference between the average probability of starting an algorithm case versus a discretionary case. On average, algorithm-selected audits are 10-14% less likely to be opened. The algorithm cases that are more likely to be started are those that have a larger turnover. The negative result on the probability of starting algorithm cases is stronger for full audits than short audits, where the difference between the probabilities is small.

One reason inspectors preferred their cases is the firm's size: the algorithm cases were on average smaller than the discretionary cases, as measured by the firm's mean declared turnover over the previous four years. Inspectors have a marked preference for larger firms: they select larger firms than the average of their tax center, and they start more often audits of larger firms among those selected. However, even controlling for firm size, it is clear that the probability of an algorithm-selected firm having an audit started is lower, on average, for almost all levels of firm size.

5.3 Effect on audit outcomes

We next compare cases by looking at the audit outcomes, in particular, whether inspectors detected irregularity. We code the outcome as a binary variable that takes value 1 if the inspectors sent a "confirmation notice" with a positive required adjustment or penalty.

The analysis is conditional on the audits that were started, a decision taken by inspectors. Still, among started cases, many do not finish in positive required adjustments or penalties, meaning that despite the inspectors' analysis, there was no detected irregularity. Here we show how the probability of finding an irregularity changes for algorithm versus discretionary cases, conditional on audits being started.

Table 10 shows that the neither the audit selection method, nor the firms size, nor other audit case attributes significantly predict the audit return in absolute value, as measured by the notifications sent to taxpayers. This results is robust across all specifications, and is similar across all tax centers (Figure 6).

Table 11 shows slightly different results for the audit yield as a share of turnover, a proxy of the firm's evasion rate. Mechanically, turnover is negatively correlated with the outcome. The algorithm selection dummy is not positive, suggesting algorithm-selected cases exhibited a higher audit return, although this result is not statistically significant. The effect becomes significant, however, in the tax offices in charge of liberal professional, among whom we would indeed expect evasion to be high and the predictive power of the algorithm to be strong.

6 Conclusion

In this paper, we have studied whether a data-driven algorithm can help improve the targeting of enforcement, focusing on the context of tax audits. Collaborating with the Senegalese tax administration DGID in an intervention at scale, we compare the implementation and return of audit cases which were selected by a risk-scoring algorithm to cases selected by tax inspectors based on a traditional discretionary procedure. We also test whether providing inspectors with easily analyzable information and with risk flags about the selected cases improves audit outcomes. Our analysis relies on partial outcome data, so that the results are still preliminary. We find that algorithm-selected audits are slightly less likely to have been implementation, and that inspector-selected audits focus on larger firms and detect the same evasion rate, thus yielding a higher return in absolute value. We do not find any evidence that the provision of information on the case or of risk flags affects audit outcomes. We aim to soon update our empirical analyses with manually digitized data on audit outcomes, which we consider more complete and of higher quality.

References

Banerjee, A., Duflo, E., Imbert, C., Mathew, S., and Pande, R. (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.

Bertrand, M., Burgess, R., Chawla, A., and Xu, G. (2018). The glittering prizes: Career incentives and bureaucrat performance. Technical report, mimeo.

Besley, T. and Persson, T. (2013). Taxation and development. In *Handbook of public economics*, volume 5, pages 51–110. Elsevier.

Carrillo, P., Pomeranz, D., and Singhal, M. (2017). Dodging the taxman: Firm misreporting and limits to tax enforcement. *American Economic Journal: Applied Economics*, 9(2):144–64.

Duflo, E., Greenstone, M., Pande, R., and Ryan, N. (2018). The value of regulatory discretion: Estimates from environmental inspections in india. *Econometrica*, 86(6):2123–2160.

Finan, F., Olken, B. A., and Pande, R. (2017). The personnel economics of the developing state. In *Handbook of Economic Field Experiments*, volume 2, pages 467–514. Elsevier.

Gerardino, M. P., Litschig, S., and Pomeranz, D. (2020). Distortion by audit.

Kang, K. and Silveira, B. S. (2021). Understanding disparities in punishment: Regulator preferences and expertise. *Journal of Political Economy*, 129(10):000–000.

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

Khwaja, M. S., Awasthi, R., and Loeprick, J. (2011). *Risk-based tax audits: approaches and country experiences*. The World Bank.

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., and Saez, E. (2011). Unwilling or unable to cheat? evidence from a tax audit experiment in denmark. *Econometrica*, 79(3):651–692.

Kleven, H. J., Kreiner, C. T., and Saez, E. (2016). Why can modern governments tax so much? an agency model of firms as fiscal intermediaries. *Economica*, 83(330):219–246.

Naritomi, J. (2019). Consumers as tax auditors. *American Economic Review*, 109(9):3031–72.

Okunogbe, O. M. and Pouliquen, V. (2018). Technology, taxation, and corruption: evidence from the introduction of electronic tax filing. *World Bank Policy Research Working Paper*, (8452).

Pomeranz, D. (2015). No taxation without information: Deterrence and self-enforcement in the value added tax. *American Economic Review*, 105(8).

Slemrod, J., Blumenthal, M., and Christian, C. (2001). Taxpayer response to an increased probability of audit: evidence from a controlled experiment in minnesota. *Journal of public economics*, 79(3):455–483.

Slemrod, J., Collins, B., Hoopes, J., Reck, D., and Sebastiani, M. (2015). Does credit-card information reporting improve small-business tax compliance? Technical report, National Bureau of Economic Research.

Xu, G. (2019). The colonial origins of fiscal capacity: Evidence from patronage governors. *Journal of Comparative Economics*.

A Figures

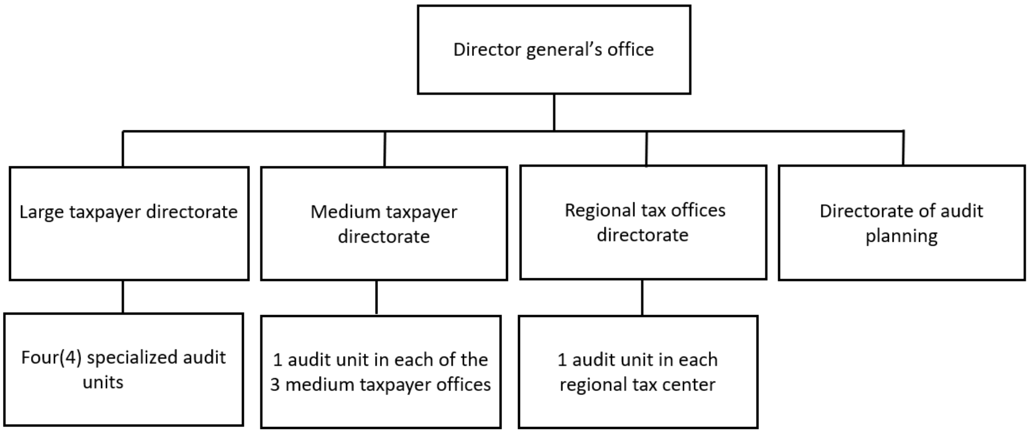


Figure 1: DGID's organizational chart

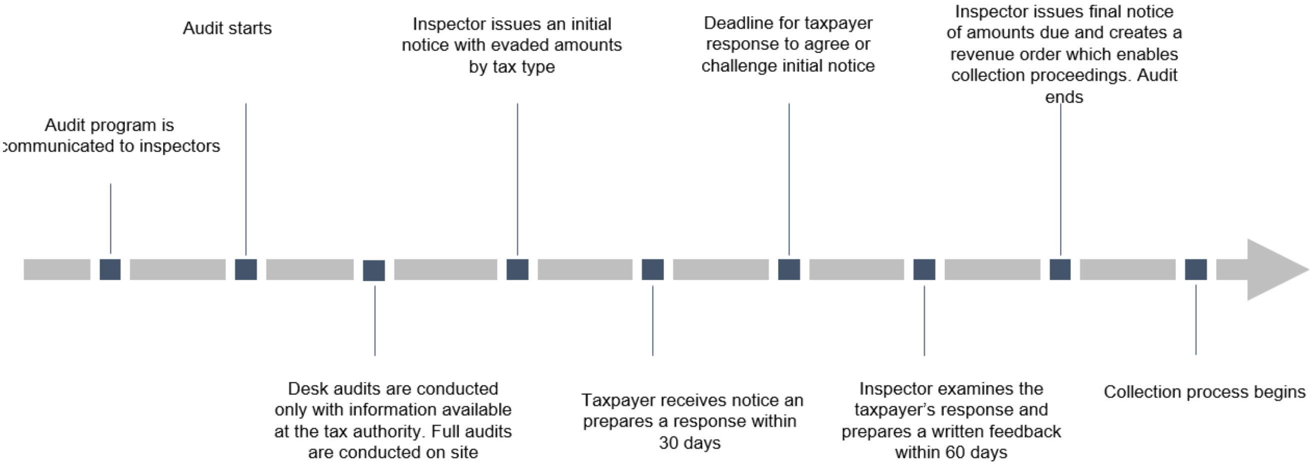


Figure 2: Audit process

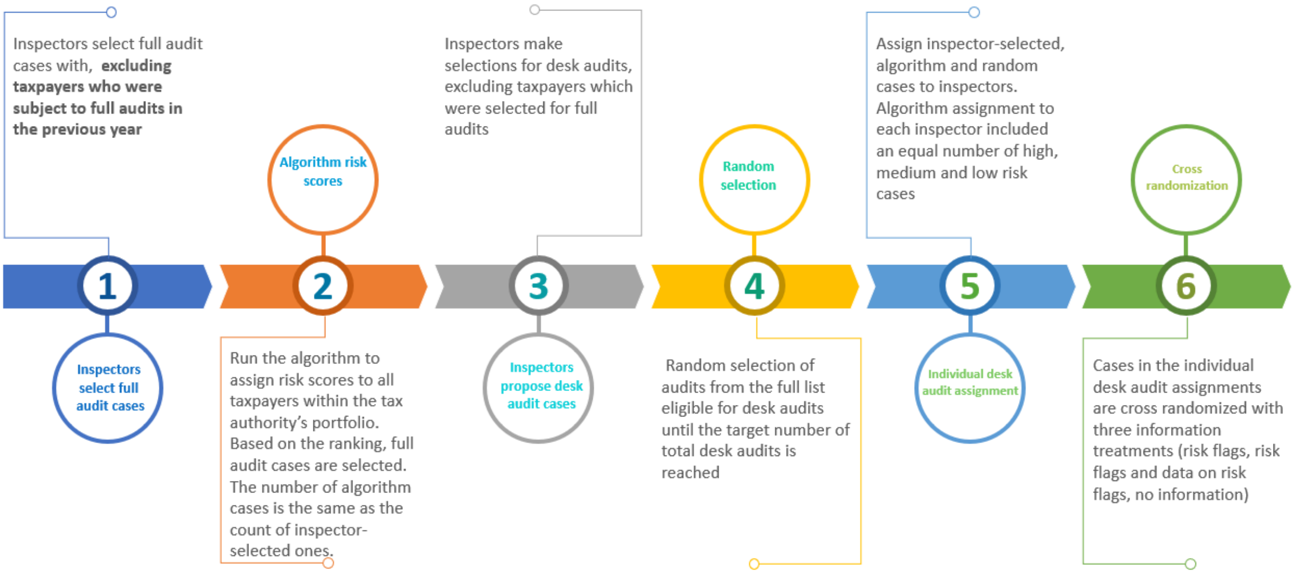


Figure 3: Program design and audit selection timeline

B Audit procedure

Table 1: Steps of risk-score design

Step	Description
(1) Prepare database	The tax declarations of each taxpayer are merged across type of taxes (VAT, CIT, Payroll) and across years. Data from third parties is then added (customs, procurement, transaction network).
(2) Choose indicators: discrepancies	Discrepancies are situations in which a self-reported tax liability can be considered as misreported or incomplete, by cross checking several data sources together.
(3) Choose indicators: anomalies	Anomalies correspond to abnormal reporting behavior, compared to peers. Anomalies suggest that firms should be monitored, but do not indicate tax evasion behavior with certainty.
(4) Define comparison clusters	Clusters regroup firms in the same economic sector and of comparable size. Peer comparisons are done within clusters
(5) Assign values to indicators	The magnitude of the inconsistency is used to assign a value, ranging from one to ten (using deciles). For anomalies firms within the top decile of a particular indicator receive a value of one.
(6) Assign weights to indicators	Weights are assigned to each indicator reflecting beliefs about their relative importance.
(7) Aggregate indicators and years	The weighted risk indicators are first aggregated across indicators in each year. Then the yearly scores are summed up to form a total risk score covering the past four years of tax declarations. More recent years are slightly over-weighted.

C Program execution

The following sections provide an analysis of the 2019 audit reports, executed in the scope of an experiment in partnership with the Senegalese Internal Revenue Services (DGID in the French acronym, henceforth designated IRS). The experiment consisted in altering the selection method of the audits program of 2019 in some fiscal centers. Part of the audits program was chosen according to the IRS' discretionary method, and part was chosen according to an algorithm, following explicit rules. The tax authority was then asked to carry out the audits on the selected firms. At the end of the year, only part of the initially planned audits had been carried out. The purpose of the analysis is to establish whether the use of the algorithm improved the ability of the tax authority to select firms for audit, especially in terms of verified tax evasion.

The audits program of 2019 consisted of 1298 firms in seven different tax centers: the two centers for middle-sized enterprises (called CME 1 and CME 2 in the French acronym), the center for liberal professionals (CPR) and four location-specific centers for small and medium enterprises, all of them in the region of Dakar, Senegal's capital (the four centers were Dakar Plateau, Grand Dakar, Ngor Almadies and Pikine Guediawaye). Part of the 1298 firms were not initially in the list of selected firms, prepared in the beginning of 2019, but were added at the IRS' discretion during the course of the year. We added them as firms selected by the IRS in our analysis.

Table **??** summarizes the execution of the 2019 progam. Out of the 1298 selected firms, 1068 were chosen to be subject to "short audits" (also called CP in the Senegalese IRS' jargon), and the remaining 230 were supposed to be subject to "full audits" (VG in the IRS' jargon). The execution rate was around 50%, meaning that for half the firms in the list there is no indication that the inspectors audited them. For the remaining half, only 37% of them ended in a request for adjustment and eventual payment of a fine.

Table 2: Summary execution full audits

		Selected					Started				
		All	Algorithm	Random	Discretionary	Ad hoc	All	Algorithm	Random	Discretionary	Ad hoc
All years	All	947	451	0	533	0	887	198	0	304	419
	LTU	386	157	0	257	0	345	74	0	145	149
	MTU	301	154	0	151	0	437	104	0	123	216
	Liberal	76	46	0	30	0	68	10	0	18	41
	SME	184	94	0	95	0	37	10	0	18	13
2018	All	294	159	0	149	0	382	102	0	124	176
	LTU	163	89	0	86	0	172	48	0	70	68
	MTU	106	55	0	53	0	178	46	0	45	91
	Liberal	20	12	0	8	0	22	4	0	7	12
	SME	5	3	0	2	0	10	4	0	2	5
2019	All	327	144	0	198	0	328	65	0	115	160
	LTU	122	37	0	96	0	144	24	0	60	69
	MTU	94	48	0	48	0	134	31	0	38	67
	Liberal	26	14	0	12	0	35	6	0	8	21
	SME	85	45	0	42	0	15	4	0	9	3
2020	All	326	148	0	186	0	177	31	0	65	83
	LTU	101	31	0	75	0	29	2	0	15	12
	MTU	101	51	0	50	0	125	27	0	40	58
	Liberal	30	20	0	10	0	11	0	0	3	8
	SME	94	46	0	51	0	12	2	0	7	5

Obs: This table contain the number of firms selected for audit and the number of audits that were started by the tax authority. Discretionary audits are the audits chosen by the tax authority. Algorithm audits are the ones chosen by the risk-based algorithm. Random audits are selected at random within the tax centers. Ad hoc audits are audits that were not in the initial program but were carried out.

Table 3: Summary execution short audits

		Selected					Started				
		All	Algorithm	Random	Discretionary	Ad hoc	All	Algorithm	Random	Discretionary	Ad hoc
All years	All	3526	1708	364	1658	24	2829	300	119	396	2071
	LTU	531	242	59	313	4	297	47	23	53	189
	MTU	792	365	137	321	14	1761	188	69	254	1285
	Liberal	523	247	77	218	3	458	31	8	45	379
	SME	1680	854	91	806	3	313	34	19	44	218
2018	All	798	318	202	303	21	1293	138	78	207	899
	LTU	208	83	59	78	4	165	33	23	33	85
	MTU	351	140	85	132	13	817	83	45	137	568
	Liberal	224	91	53	89	2	222	18	5	31	172
	SME	15	4	5	4	2	89	4	5	6	74
2019	All	985	468	162	390	3	813	111	41	139	539
	LTU	2	0	0	2	0	73	0	0	2	71
	MTU	350	174	52	145	1	506	81	24	95	322
	Liberal	158	76	24	60	1	96	4	3	6	83
	SME	475	218	86	183	1	138	26	14	36	63
2020	All	1743	922	0	965	0	723	51	0	50	633
	LTU	321	159	0	233	0	59	14	0	18	33
	MTU	91	51	0	44	0	438	24	0	22	395
	Liberal	141	80	0	69	0	140	9	0	8	124
	SME	1190	632	0	619	0	86	4	0	2	81

Obs: This table contain the number of firms selected for audit and the number of audits that were started by the tax authority. Discretionary audits are the audits chosen by the tax authority. Algorithm audits are the ones chosen by the risk-based algorithm. Random audits are selected at random within the tax centers. Ad hoc audits are audits that were not in the initial program but were carried out.

Table 4: Tax audit selection methods in selected countries

ntry	Discretionary selection	Risk analysis	Random selection
ya	Yes ; For all except large taxpayers	Yes ; Only for large taxpayers	No
egal	Yes	Yes, Introduced in FY 2018	Introduced in FY 2018
babwe	Yes; Inspectors rated on selection.	Yes; based on turnover variances	No
otho	No	No	Yes ; Randomly by ma
zania	Abandonned in 2007	Yes	
ted Kingdom	Yes; For 55% of audit cases	Yes; Risk scoring	Yes ; Simple random s
tzerland	Yes for all cases	No	Yes, periodically for so
ted States	No	Yes	
nce	Yes; For intelligence gathering	Yes; statistical techniques, data-mining	No
garia	Yes ; According to set criteria	Yes; Central risk analysis	No
key	No	Yes; Analysis by tax type	Yes ; to collect unbias

ources; Khwaja et al. 2011 and Authors' survey of select country tax officials.

C.1 Firms' characteristics

Table 5: Number of firms by data source

		2014	2015	2016	2017	2018	2019
Self reported	VAT	0	6138	6359	6486	5883	5842
	CIT	0	3823	3970	4245	4159	0
	CGU	0	16	34	63	76	62
	WIT	0	4503	4574	5101	5329	5344
	TAF	0	19	18	19	18	16
Third party	Imports	0	1500	1556	1483	1450	0
	Exports	0	446	463	441	429	0
	Treasury	0	547	547	428	444	0
	VAT annexes	0	0	0	0	0	0
Audits data	Fiches de suivi	0	0	0	0	0	1286
	Saisie	0	0	0	0	0	0

Note: Number of firms for which data was available, according to each data source. There are three main sources of data: self-reported tax declarations (Value Added Tax, Corporate Income Tax, simplified regime CGU, Withtheld Income Tax, financial services tax TAF), third party data (exports, imports, treasury payments and VAT annexes concerning inter-firm transactions) and the data produced by the tax inspectors regarding the audit program of 2019. The data includes the following tax centers in Senegal: medium taxpayers 1, medium taxpayers 2, liberal professionals, Dakar Plateau, Grand Dakar, Pikine Guediawaye, Ngor Almadies.

Table 6: Number of firms by data source

	Mean population	Mean random selection	Difference	p-value	Mean IRS selection	Mean algorithm selection	Difference	p-value
Turnover (mean 2015-2018)	162	172	-10	.87	463	290	173	0
Mean profit (mean 2015-2018)	1	3	-1	.85	6	-2	7	.23
Profit rate (2015-2018) Mean Payroll (2015-2018)	9	11	-3	.48	24	15	9	.01
Tax liability (total 2015-2018)	33	30	3	.76	94	84	10	.48
Risk score	0	156	-156	0	175	1618	-1443	0
Turnover 2018	233	326	-93	.4	597	400	197	.01
Profit 2018	15	8	7	.87	34	70	-36	.46
Number of employees 2018	414	12	403	.3	229	215	14	.93
N	11386	154	.	.	600	574	.	.

Note: Number of firms for which data was available, according to each data source. There are three main sources of data: self-reported tax declarations (Value Added Tax, Corporate Income Tax, simplified regime CGU, Withtheld Income Tax, financial services tax TAF), third party data (exports, imports, treasury payments and VAT annexes concerning inter-firm transactions) and the data produced by the tax inspectors regarding the audit program of 2019. The data includes the following tax centers in Senegal: medium taxpayers 1, medium taxpayers 2, liberal professionals, Dakar Plateau, Grand Dakar, Pikine Guediawaye, Ngor Almadies.

C.2 Outcomes

To analyze the data, we propose six outcomes: the probability that the audit started, the probability that there was an adjustment (conditional on audits having started), the amount of the first notification (the initial quantity of suspected evasion communicated to the taxpayer), the confirmed amount of evasion, the evasion rate as a percentage of the total tax liability, and the evasion rate as a percentage of mean turnover. A first comparison of the outcomes across short audits and full audits, and across selection methods, can be observe in table **??** below.

Table 7: Mean characteristics firms - All firms

	Mean Random	Mean IRS selection	Mean algorithm selection	Difference	p-value
1 probability being started	.58	.63	.49	.14	0
2 audit ending in adjustment	.31	.57	.31	.27	0
3 log (initial notice)	17.17	17.88	17.33	.55	.01
4 log (final notice)	16.59	17.22	16.75	.47	.02
5 evasion as % liability	.71	.68	.68	.01	.9
6 evasion as % of mean turnover	.4	.3	.4	-.1	.01
7 days spent on case	4.55	40.77	19.08	21.69	0
8 log turnover 2019	11.02	14.36	14.12	.24	.85

Note: Mean characteristics of firms in selection and in the population. Total tax liability includes only self declared tax liability in VAT, CIT, PAYE and CGU for firms. The data includes the following tax centers in Senegal: medium taxpayers 1, medium taxpayers 2, liberal professionals, Dakar Plateau, Grand Dakar, Pikine Guediawaye, Ngor Almadies. Values of turnover, tax liability and profits are expressed in Millions FCFA. Profit rate is in percentage of turnover, computed as the mean profit divided by the mean turnover. Number of employees refers to the number of employees in the PAYE declarations.

The definition of the outcomes is as follows:

- $i(Auditstarted)$: indicator function that takes value 1 if the audit contained any indication that the inspector worked on it. This variable takes value 1 whenever the audit report of the firm contains the indication of some evasion quantity, some qualitative variable, or even an indication of the date in which the audit was started. For many cases, the audit is started but not finished.
- $i(Adjustment > 0)$: indicator function containing some quantity of uncovered evasion. It can be the final amount the firm is asked to pay or the initially notified amount (which happens more often).
- $log(Notification)$: log of the value of the notified amount of evaded taxes. That is the amount of evasion that is assessed by the inspectors after the inspection. This amount is then negotiated with the firm, which provides some explanation about the problems, and is typically reduced in the confirmation stage.
- $log(Evasion)$: log of the assessed evasion of the firm. In this stage, we use the value of the *confirmed amount* of evaded taxes or the *final requested payment*. Whenever the two values are not the same (which happens very rarely) we take the max between them. We complement missing information with the value of notification (the outcome before) adjusted by the mean deduction from notification and confirmation at the tax office level and for each particular audit type (full or short audits). For example, in the Liberal Professions office, we observe that on average the confirmation is 57% the value of the initial notification (when both quantities are filled in) for full audits, so when we only have the value of notification (for full audits in that particular office) we complement the evasion variable by multiplying it by 57%.

C.3 Description of the firms and outcomes by firm size

In the 2019 wave of the experiment, we proposed firms to be audited in tax centers in the Dakar area. The tax centers included small to medium enterprises. Based on their self reported yearly turnovers, we can plot the distribution of firm size in each of the tax centers below.

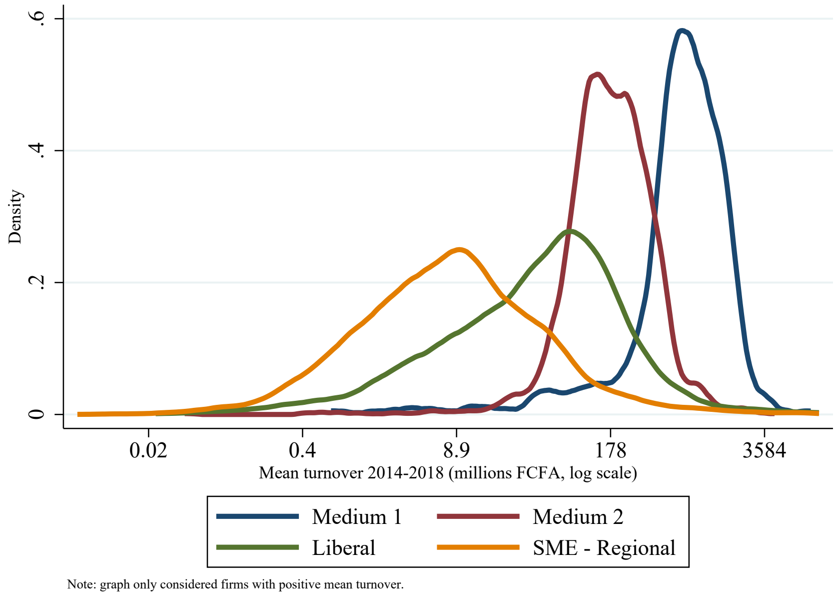
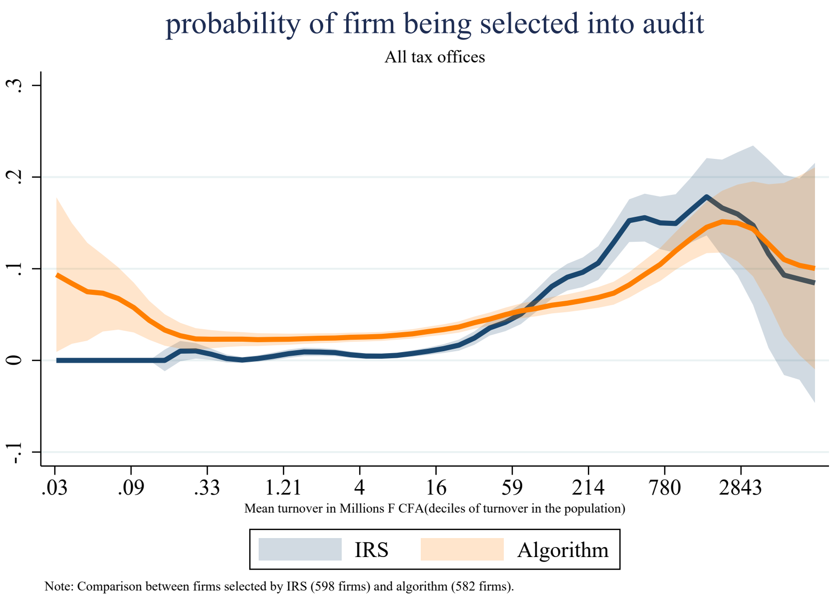


Figure 4

In every tax office, firms with larger declared turnover have a higher probability of being audited, in particular for IRS selected cases. The algorithm also gives explicitly more weight to firms with more declared turnover. Even though the algorithm explicitly gives more weight to firms with larger turnover, its selection is less concentrated at large firms than the inspector selection. The following figure shows how the two selections differ in terms of (self-declared) firms size. Firms with mean declared turnover lower than 16 Million FCFA (roughly 25 thousand euros) per year have virtually no chance of being selected for audit by the tax authority, while the algorithm assigns them positive probability of audit. In particular, firms with extremely low declarations had almost 10% chances of being selected by the algorithm, while no chance of being selected by the tax inspectors.



Obs: Non parametric regression of outcome on mean turnover (no controls), using Epanechnikov kernel, bandwidth computed according to the rule-of-tumb method.

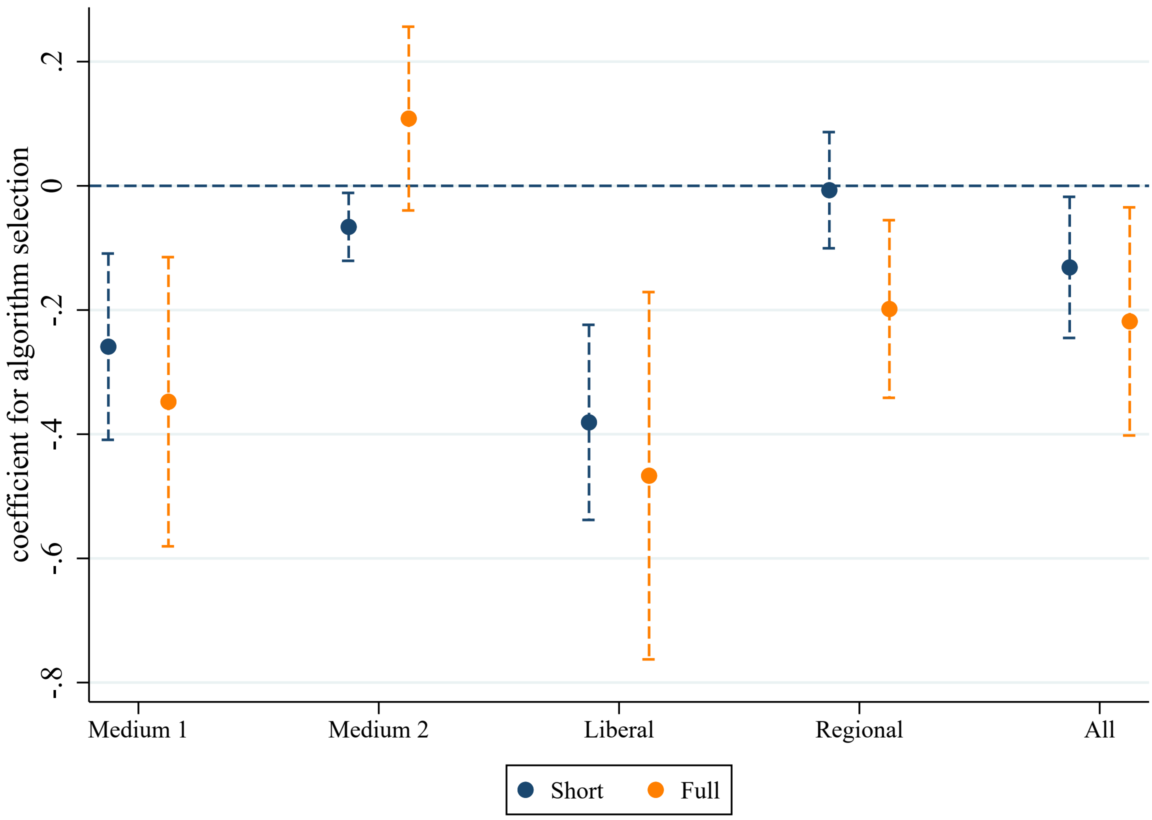
C.4 Impact of selection on outcomes

Table 8: Effect of algorithm selection on probability of audit being started

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	All audits	Full audits	Full audits	Full audits	Full audits	Full audits	Short audits	Short audits	Short audits	Short audits	Short audits	Short audits	Short audits
Algorithm selection	-0.0402***	-0.117***	-0.143***	-0.0980**	-0.171***		-0.0162	-0.0985***	-0.00847	-0.00913		0.00400	-0.0322**
	(0.0101)	(0.0286)	(0.0500)	(0.0383)	(0.0358)		(0.0103)	(0.0362)	(0.0115)	(0.0107)		(0.00932)	(0.0135)
Algorithm X Exporting				-0.0394					0.00955				
				(0.0515)					(0.0232)				
Algorithm X TP data					0.157***					0.0208			
					(0.0516)					(0.0285)			
Algorithm X LTU						-0.121***					-0.00912		
						(0.0447)					(0.0281)		
Algorithm X Medium						-0.0649					0.0856***		
						(0.0447)					(0.0304)		
Algorithm X Lib.						-0.396***					-0.108***		
						(0.0713)					(0.0256)		
Algorithm X SME						-0.0567					-0.0347***		
						(0.0562)					(0.00700)		
Random	-0.0620**						-0.0163	-0.171***	0.0642**	0.0625**	-0.00768	0.0424	-0.0295
	(0.0274)						(0.0278)	(0.0503)	(0.0272)	(0.0272)	(0.0286)	(0.0404)	(0.0286)
Ad hoc	0.554***	0.277***		0.319***	0.308***	0.275***	0.631***		0.661***	0.648***	0.643***	0.523***	0.639***
	(0.0109)	(0.0244)		(0.0245)	(0.0245)	(0.0250)	(0.0121)		(0.0121)	(0.0126)	(0.0136)	(0.0206)	(0.0121)
Full audit	0.130***												
	(0.0124)												
Exporting firm				0.0788***					0.0120				
				(0.0265)					(0.0120)				
TP data					0.00384					0.0527***			
					(0.0255)					(0.0101)			
Information treatment													0.0351**
													(0.0144)
Sample	All centers	All centers	Only MTU	All centers	All centers	All centers	All centers	Only MTU	All centers	All centers	All centers	All centers	All centers
Tax Center fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Activity group fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector fixed effects	No	No	No	No	No	No	No	No	No	No	No	Yes	No
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	7440	1361	302	1361	1361	1361	6079	868	6079	6079	6079	5036	6079
R2	0.617	0.325	0.0464	0.297	0.301	0.333	0.683	0.0552	0.666	0.668	0.686	0.810	0.683
Mean outcome	0.48	0.64	0.73	0.64	0.64	0.64	0.44	0.54	0.44	0.44	0.44	0.42	0.44

Note: OLS regression of probability of audit being started on the selection method. Different specifications controlling for the type of audit, the firm's mean turnover (with the information available over years 2015-2018), and dummies for the 6 tax centers (medium enterprises 1, medium enterprises 2, liberal professions, Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies). Standard errors are shown parentheses, and were computed clustered at the tax center level.

Figure 5: Effect of algorithm selection on probability of audit being started, by tax center



Obs: Coefficients of the regression of the outcome on the algorithm selection, controlling for mean firm turnover, by tax office and type of audit. The last two coefficients (All) represent the coefficients of same regression as the last two columns of the corresponding regression table.

Table 9: Effect of algorithm selection probability of adjustment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	All audits	Full audits	Full audits	Full audits	Full audits	Full audits	Short audits	Short audits	Short audits	Short audits	Short audits	Short audits	Short audits
Algorithm selection	-0.0536**	-0.0362	-0.0749	-0.0615	-0.114**		-0.0819**	-0.131***	-0.0877**	-0.112***		-0.0475	-0.0828**
	(0.0264)	(0.0400)	(0.0627)	(0.0574)	(0.0564)		(0.0353)	(0.0457)	(0.0412)	(0.0414)		(0.0526)	(0.0363)
Algorithm X Exporting				0.0855					-0.0131				
				(0.0734)					(0.0639)				
Algorithm X TP data					0.213***					0.0730			
					(0.0722)					(0.0618)			
Algorithm X LTU						0.00903					-0.0241		
						(0.0534)					(0.0676)		
Algorithm X Medium						-0.0267					-0.0974**		
						(0.0572)					(0.0423)		
Algorithm X Lib.						-0.307**					0.0149		
						(0.156)					(0.0753)		
Algorithm X SME						-0.229					-0.247**		
						(0.183)					(0.103)		
Random	-0.111**						-0.150***	-0.261***	-0.139***	-0.138***	-0.149***	-0.163*	-0.151***
	(0.0496)						(0.0516)	(0.0674)	(0.0527)	(0.0526)	(0.0517)	(0.0835)	(0.0529)
Ad hoc	-0.147***	-0.0483		-0.0289	-0.0390	-0.0488	-0.209***		-0.215***	-0.223***	-0.209***	-0.183***	-0.207***
	(0.0204)	(0.0338)		(0.0344)	(0.0347)	(0.0339)	(0.0259)		(0.0258)	(0.0258)	(0.0260)	(0.0460)	(0.0281)
Full audit	0.0461**												
	(0.0200)												
Exporting firm				0.0793**					0.0583**				
				(0.0370)					(0.0261)				
TP data					-0.0240					0.0242			
					(0.0367)					(0.0203)			
Information treatment													0.00389
													(0.0335)
Sample	All centers	All centers	Only MTU	All centers	All centers	All centers	All centers	Only MTU	All centers	All centers	All centers	All centers	All centers
Tax Center fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Activity group fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector fixed effects	No	No	No	No	No	No	No	No	No	No	No	Yes	No
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	3589	880	221	880	880	880	2709	476	2709	2709	2709	2108	2709
R2	0.0835	0.0749	0.0410	0.0532	0.0517	0.0800	0.0803	0.0554	0.0746	0.0740	0.0816	0.166	0.0803
Mean outcome	0.58	0.68	0.66	0.68	0.68	0.68	0.55	0.68	0.55	0.55	0.55	0.49	0.55

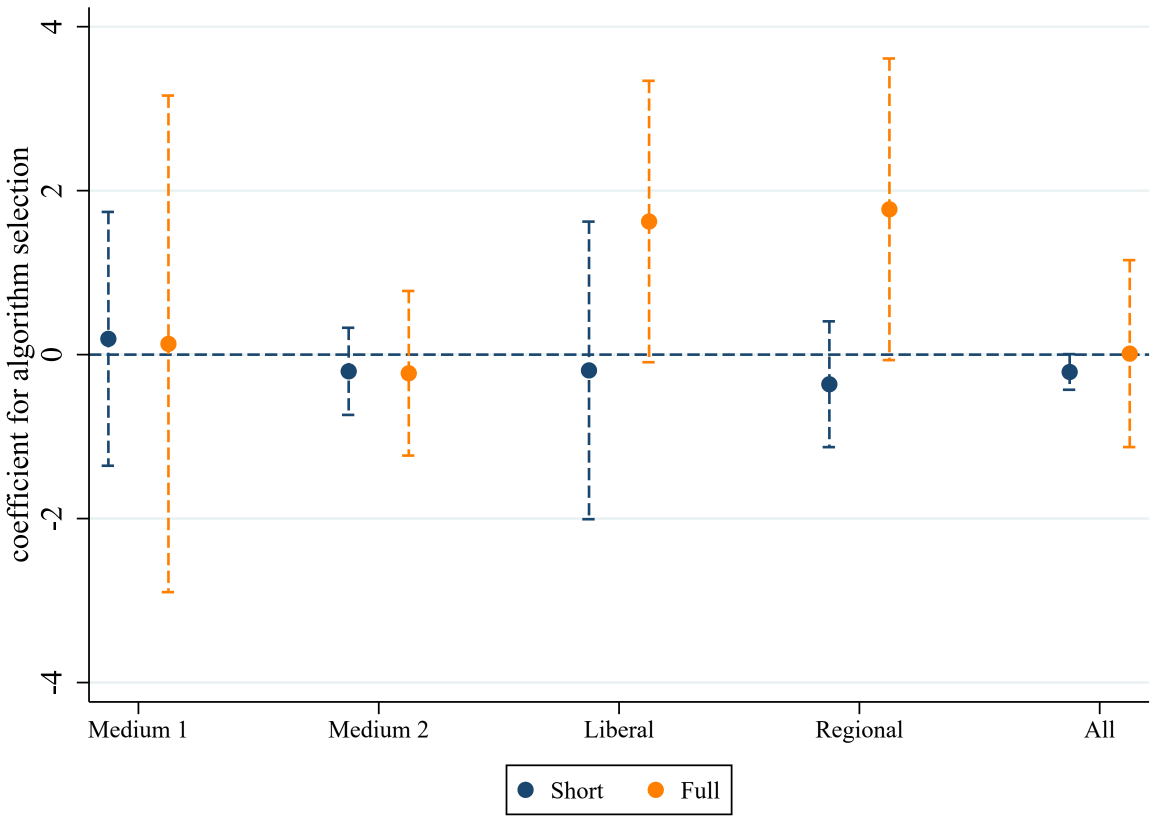
ote: OLS regression of log (initial notice) on the selection method. Different specifications controlling for the type of audit, the firm's rnover (with the information available over years 2015-2018), and dummies for the 6 tax centers (medium enterprises 1, medium enter liberal professions, Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies). Standard errors are shown parentheses, and mputed clustered at the center level.

Table 10: Effect of algorithm selection on log (initial notice)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	All audits	Full audits	Full audits	Full audits	Full audits	Full audits	Short audits	Short audits	Short audits	Short audits	Short audits	Short audits	Short audits
Algorithm selection	-0.0511	-0.382**	-0.861***	-0.611**	-0.504*		0.138	0.0200	0.121	0.00717		0.250	0.147
	(0.123)	(0.183)	(0.242)	(0.276)	(0.268)		(0.156)	(0.184)	(0.183)	(0.185)		(0.181)	(0.157)
Algorithm X Exporting				0.407					0.0601				
				(0.345)					(0.268)				
Algorithm X TP data					0.274					0.391			
					(0.342)					(0.264)			
Algorithm X LTU						-0.112					0.183		
						(0.274)					(0.296)		
Algorithm X Medium						-1.080***					0.104		
						(0.234)					(0.170)		
Algorithm X Lib.						1.536**					0.503		
						(0.660)					(0.971)		
Algorithm X SME						2.304***					0.106		
						(0.644)					(0.608)		
Random	0.00579						0.00710	-0.145	-0.0102	0.0111	0.00700	0.231	0.0218
	(0.192)						(0.193)	(0.235)	(0.192)	(0.194)	(0.193)	(0.280)	(0.198)
Ad hoc	0.0666	0.122		0.153	0.129	0.0739	0.125		0.112	0.0852	0.126	0.281*	0.104
	(0.0910)	(0.153)		(0.154)	(0.154)	(0.149)	(0.111)		(0.109)	(0.110)	(0.111)	(0.153)	(0.127)
Full audit	1.003***												
	(0.0901)												
Exporting firm				0.218					0.325***				
				(0.164)					(0.104)				
TP data					-0.0677					0.0562			
					(0.158)					(0.0910)			
Information treatment													-0.0521
													(0.139)
Sample	All centers	All centers	Only MTU	All centers	All centers	All centers	All centers	Only MTU	All centers	All centers	All centers	All centers	All centers
Tax Center fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Activity group fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector fixed effects	No	No	No	No	No	No	No	No	No	No	No	Yes	No
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	2397	644	164	644	644	644	1753	357	1753	1753	1753	1410	1753
R2	0.258	0.140	0.0838	0.147	0.139	0.179	0.175	0.00359	0.179	0.175	0.176	0.404	0.175
Mean outcome	18.41	19.59	18.63	19.59	19.59	19.59	17.97	17.93	17.97	17.97	17.97	17.85	17.97

ote: OLS regression of log (initial notice) on the selection method. Different specifications controlling for the type of audit, the firm's rnover (with the information available over years 2015-2018), and dummies for the 6 tax centers (medium enterprises 1, medium enter liberal professions, Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies). Standard errors are shown parentheses, and mputed clustered at the center level.

Figure 6: Effect of algorithm selection on log(initial notice), by tax center

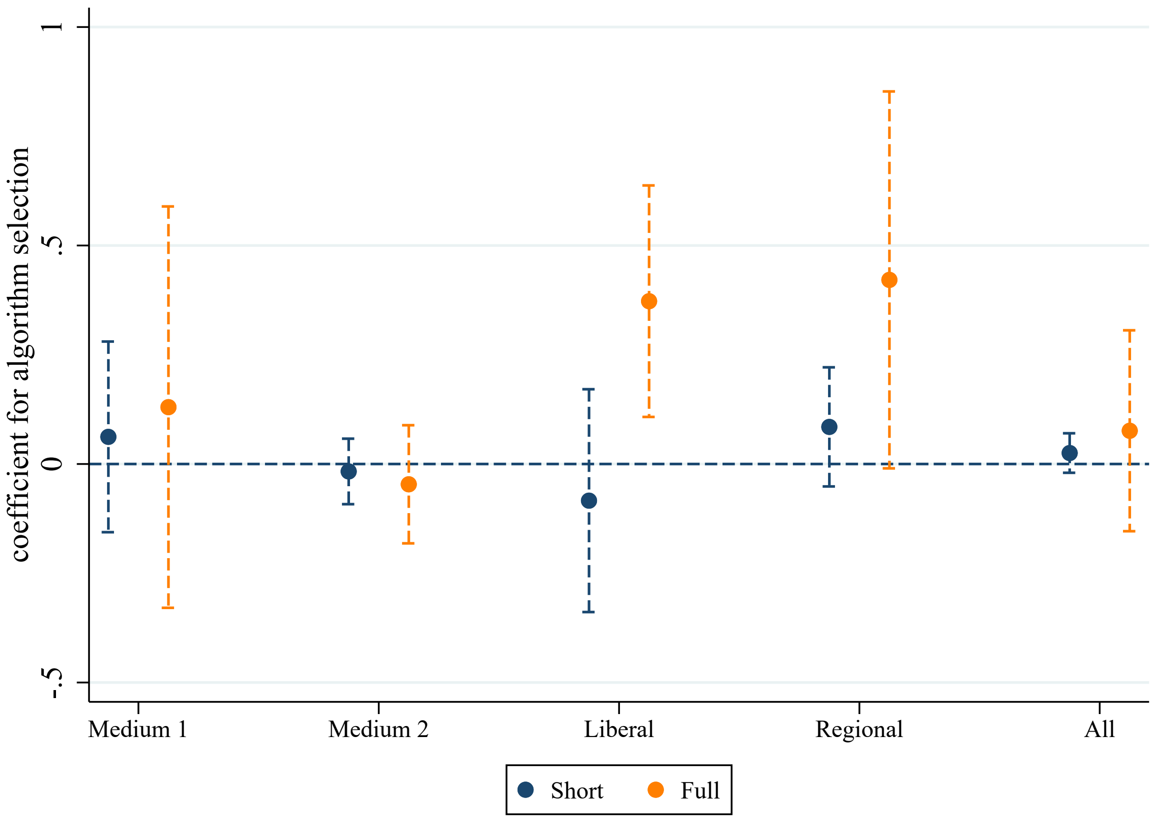


Obs: Coefficients of the regression of the outcome on the algorithm selection, controlling for mean firm turnover, by tax office and type of audit. The last two coefficients (All) represent the coefficients of same regression as the last two columns of the corresponding regression table.

Table 11: Effect of algorithm selection on evasion as % of mean turnover

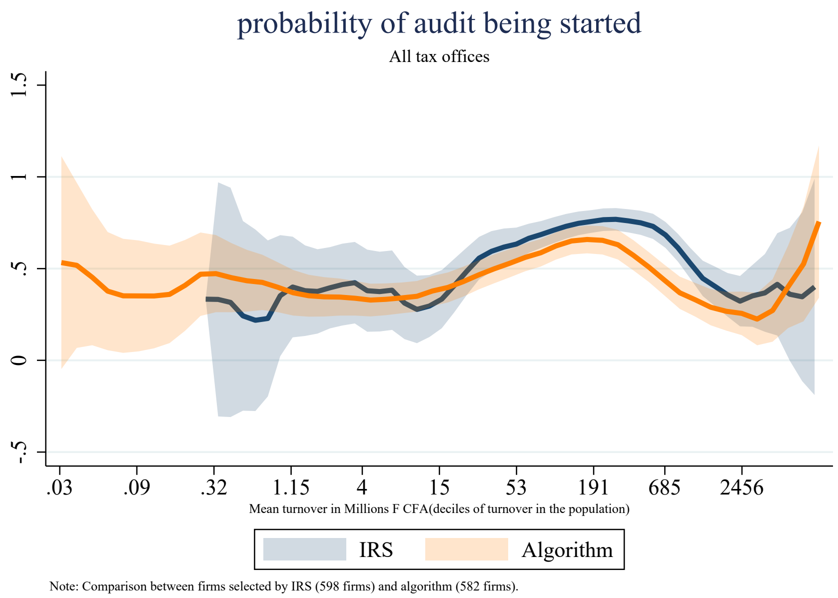
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	All audits	Full audits	Full audits	Full audits	Full audits	Full audits	Short audits	Short audits	Short audits	Short audits	Short audits	Short audits	Short audits
Algorithm selection	0.00281	0.0339	0.00257	0.0236	0.0156		-0.0254	-0.0120	-0.0101	-0.0208		-0.0346	-0.0252
	(0.0166)	(0.0289)	(0.0378)	(0.0418)	(0.0444)		(0.0187)	(0.0247)	(0.0229)	(0.0235)		(0.0316)	(0.0191)
Algorithm X Exporting				0.0470					-0.0551*				
				(0.0551)					(0.0314)				
Algorithm X TP data					0.0823					-0.0217			
					(0.0556)					(0.0298)			
Algorithm X LTU						0.0242					-0.102***		
						(0.0419)					(0.0347)		
Algorithm X Medium						0.0220					0.00137		
						(0.0365)					(0.0232)		
Algorithm X Lib.						0.0939					-0.0453		
						(0.183)					(0.0300)		
Algorithm X SME						0.559***					-0.0877		
						(0.0707)					(0.106)		
Random	0.0128						-0.00338	-0.00486	0.00521	0.00423	-0.00767	0.0254	-0.00321
	(0.0316)						(0.0312)	(0.0348)	(0.0309)	(0.0308)	(0.0313)	(0.0518)	(0.0324)
Ad hoc	0.0125	0.0298		0.0327	0.0343	0.0297	-0.00825		-0.0325**	-0.0341**	-0.00831	-0.0602**	-0.00844
	(0.0142)	(0.0259)		(0.0263)	(0.0261)	(0.0258)	(0.0166)		(0.0157)	(0.0156)	(0.0167)	(0.0296)	(0.0172)
Full audit	0.0751***												
	(0.0149)												
Exporting firm				-0.00354					-0.00503				
				(0.0276)					(0.0167)				
TP data					-0.0842***					-0.0378**			
					(0.0282)					(0.0172)			
Information treatment													-0.000601
													(0.0187)
Sample	All centers	All centers	Only MTU	All centers	All centers	All centers	All centers	Only MTU	All centers	All centers	All centers	All centers	All centers
Tax Center fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Activity group fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector fixed effects	No	No	No	No	No	No	No	No	No	No	No	Yes	No
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1616	542	148	542	542	542	1074	324	1074	1074	1074	761	1074
R2	0.0807	0.0751	0.140	0.0345	0.0485	0.0849	0.0737	0.0806	0.0449	0.0489	0.0787	0.334	0.0737
Mean outcome	0.15	0.21	0.18	0.21	0.21	0.21	0.12	0.15	0.12	0.12	0.12	0.12	0.12

ote: OLS regression of evasion as % of mean turnover on the selection method. Different specifications controlling for the type of audi m's mean turnover (with the information available over years 2015-2018), and dummies for the 6 tax centers (medium enterprises 1, m terprises 2, liberal professions, Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies). Standard errors are shown parent d were computed clustered at the tax center level.



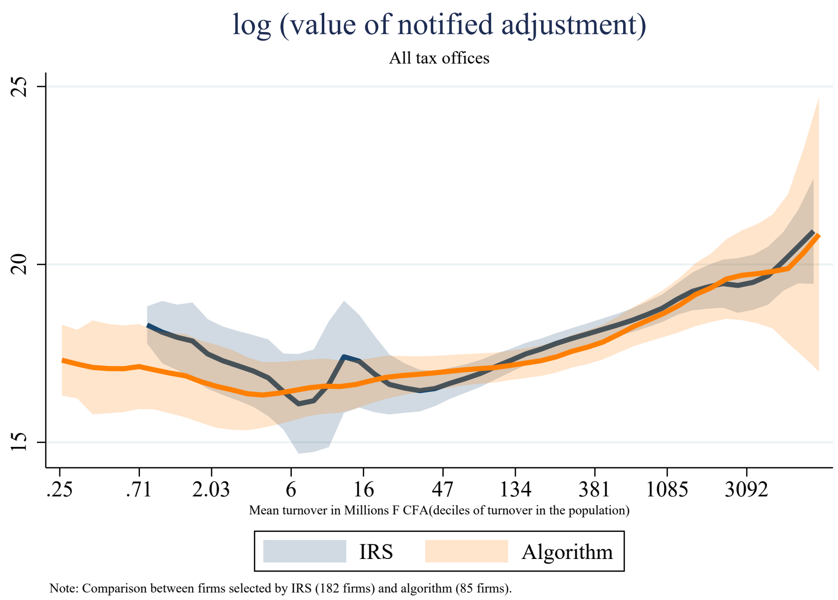
Obs: Coefficients of the regression of the outcome on the algorithm selection, controlling for mean firm turnover, by tax office and type of audit. The last two coefficients (All) represent the coefficients of same regression as the last two columns of the corresponding regression table.

Figure 7: Probability of audit being started



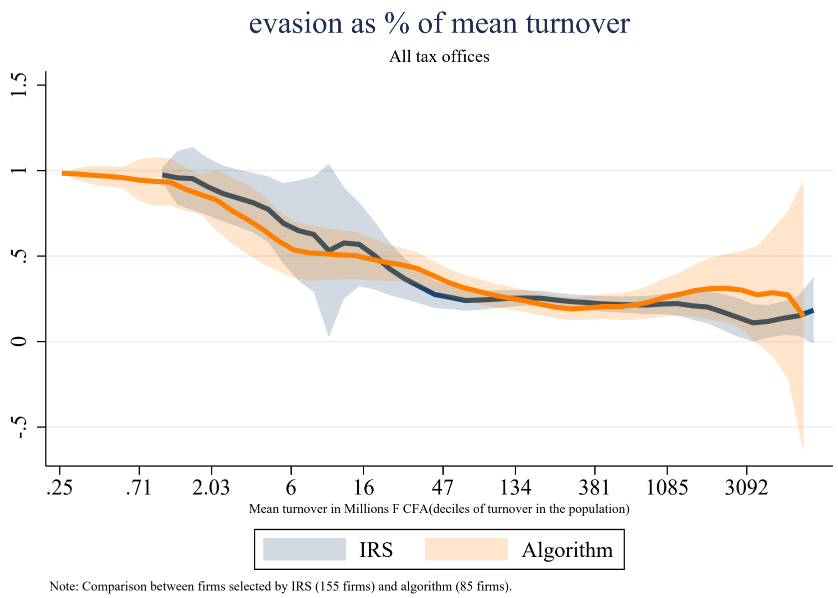
Obs: Non parametric regression of outcome on mean turnover (no controls), using Epanechnikov kernel, bandwidth computed according to the rule-of-tumb method.

Figure 8: log(Initial notice)



Obs: Non parametric regression of outcome on mean turnover (no controls), using Epanechnikov kernel, bandwidth computed according to the rule-of-tumb method.

Figure 9: Evasion as a % of mean turnover



Obs: Non parametric regression of outcome on mean turnover (no controls), using Epanechnikov kernel, bandwidth computed according to the rule-of-tumb method.

C.5 Evaluation of risk score

Appendix C Risk Scoring of Tax Evasion

C.1 Motivation

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

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

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

Table XX summarizes the seven key steps in the design of the risk-score. Step (1) corresponds to the construction of a database covering all tax declarations across years and merged with third-party reported sources. Steps (2) and (3) determine specific risk indicators, based on discrepancies across sources or behavioral outliers, examples of which are discussed below. Step (4) defines the peer-group comparison: these clusters regroup firms by economic activity and either size or geographical zones, depending on the structure of each tax center. Step (5) assigns a numerical value to each risk indicator, depending on the size of the deviation (higher scores when larger discrepancies), while step (6) assigns weights to each indicator reflecting beliefs about their relative importance. Finally, step (7) aggregate the weighted indicators in each of the past four fiscal year, and then sums up the yearly scores to form a total risk score.

Table C1: Steps of risk-score design

Step	**Description**
(1) Prepare merged dataset	The tax declarations of each taxpayer are merged across type of taxes (VAT, CIT, Payroll) and across years. Data from third parties is then added (customs, procurement, transaction network).
(2) Choose indicators: discrepancies	Discrepancies are situations in which a self-reported tax liability can be considered as misreported or incomplete, by cross checking several data sources together.
(3) Choose indicators: anomalies	Anomalies correspond to abnormal reporting behavior, compared to peers. Anomalies suggest that firms should be monitored, but do not indicate tax evasion behavior with certainty.
(4) Define comparison clusters	Clusters regroup firms in the same economic sector and of comparable size. Peer comparisons are done within clusters
(5) Assign values to indicators	The magnitude of the inconsistency is used to assign a value, ranging from one to ten (using deciles). For anomalies firms within the top decile of a particular indicator receive a value of one.
(6) Assign weights to indicators	Weights are assigned to each indicator reflecting beliefs about their relative importance.
(7) Aggregate indicators and years	The weighted risk indicators are first aggregated across indicators in each year. Then the yearly scores are summed up to form a total risk score covering the past four years of tax declarations. More recent years are slightly over-weighted.

C.2 Choosing indicators and weights

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

Table C1 summarizes the steps that we took to conceptualize the algorithm. We tried out several possible indicators that could suggest under-declaration of tax liability. We discarded most based on some analysis of data availability or statistical relevance. In the end, we discarded indicators that required information that was available for a reduced set of firms and indicators that did not seem to have any correlation with evasion, as per past evasion data. We tested these indicators on data from historical audits data. We performed out of sample regressions with LASSO and OLS and computed the out of sample mean squared prediction errors to compare different models. This allowed us to assert that the ranking normalization performed well with respect to alternatives (meaning that it presented a lower prediction error).

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

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

We designed the risk-scoring scheme using best practices, drawing on policy documents from the World Bank (tax administration projects in Pakistan and Turkey), SKAT in Denmark, and the IMF's recommendations to DGID. We provide a high-level description of this process to preserve confidentiality around audit selection processes. We compute risk scores using information sets/tax returns submitted to DGID on corporate income taxes, VAT, personal income tax withholding remittance, as well external data from customs (imports/exports) and public procurement contracts, for the period 2013-2016 [18]. The score relies on two types of risk indicators: discrepancies and anomalies. Discrepancy indicators flag taxpayers whose self-reported information according to their tax returns differs from information in datasets obtained from customs or the government budget department in charge of paying state procurement. For instance, a discrepancy indicator is logged when taxpayers' reported turnover over multiple years is lower than its aggregate costs, that its imports plus its wage bill over the same period. Anomaly indicators use industry/sector benchmarking to flag firms with unusual behavior relative to their peers. An example would be a firm in petroleum retail with low profit rate compared to its peers, which might be associated with evasion. Discrepancies and anomalies are aggregated to produce a risk-score for each taxpayer.

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