WIDER Working Paper 2024/66

Estimating the value-added tax gap in Tanzania

A study of small, medium, and micro enterprises

Amina Ebrahim,[1] Sebastián Castillo,[2] Vincent Leyaro,[3] Ezekiel Swema,[4] Oswald Haule,[4] Massaga Fimbo,[4] and Ephraim Mdee[4]

November 2024

United Nations University World Institute for Development Economics Research

wider.unu.edu

Abstract: This study measures the VAT compliance gap for small and medium-sized entities in Tanzania. Specifically, the study measures the under-reporting component of the VAT compliance gap. This study uses VAT declaration and audit data to conduct a bottom-up estimation to measure the extent of VAT misreporting in small, medium, and micro enterprises. The study's objective is to examine the extent of VAT under-reporting from 2014 to 2020 and identify the behaviour of firms that contribute to the VAT gap. The study finds a VAT gap of 62%, with the largest gaps in the agriculture, transportation, and manufacturing sectors. The accommodation, professional, wholesale and retail, and manufacturing sectors drive the overall VAT gap and are thus more important in identifying VAT revenue losses. The study provides evidence of firms' strategic reporting behaviour to avoid being audited, which increases the possibility of VAT evasion. Lastly, the study provides a cost–benefit ratio to make cost-effective recommendations to redistribute resources to increase auditing in sectors with more considerable evasion and lower audit probability.

Key words: VAT compliance gap, bottom-up estimation, audit, under-reporting, Tanzania

JEL classification: H25, H26, H32

Acknowledgements: The authors gratefully acknowledge the anonymized data sharing by the Tanzania Revenue Authority. The views expressed do not represent the views of the Tanzania Revenue Authority. We are thankful for the comments and guidance from Prof. Jukka Pirttilä. We thank participants and discussants at the 2023 WIDER Development Conference in Oslo and the Finnish Centre of Excellence in Tax Systems (FIT) Research Workshop in Tampere in October 2023. Mostafa Bahbah provided superb research assistance. Sebastián Castillo acknowledges the financial backing of the Finnish Centre of Excellence in Tax Systems Research (FIT), funded by the Academy of Finland (project 346251).

Note: This study has received ethical approval by the Joint Ethical Review Board of the United Nations University (Ref No: [202104/01]) on 11 May 2021.

[1]UNU-WIDER, Helsinki, Finland email amina@wider.unu.edu; [2]University of Helsinki, Finnish Centre of Excellence in Tax Systems Research (FIT), Helsinki, Finland; [3]University of Dar es Salaam, Dar es Salaam, Tanzania; [4]Tanzania Revenue Authority, Dar es Salaam, Tanzania

This study has been prepared within the UNU-WIDER project Tax research for development (phase 3), which is part of the research area Creating the fiscal space for development. The project is part of the Domestic Revenue Mobilization programme, which is financed through specific contributions by the Norwegian Agency for Development Cooperation (Norad)

Information and requests: publications@wider.unu.edu

ISSN 1798-7237 ISBN 978-92-9267-529-5

https://doi.org/10.35188/UNU-WIDER/2024/529-5

Typescript prepared by Lesley Ellen.

United Nations University World Institute for Development Economics Research provides economic analysis and policy advice with the aim of promoting sustainable and equitable development. The Institute began operations in 1985 in Helsinki, Finland, as the first research and training centre of the United Nations University. Today it is a unique blend of think tank, research institute, and UN agency—providing a range of services from policy advice to governments as well as freely available original research.

The Institute is funded through income from an endowment fund with additional contributions to its work programme from Finland and Sweden, as well as earmarked contributions for specific projects from a variety of donors.

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introduction

Value-added tax (VAT) has emerged as a pivotal fiscal tool in developed and developing economies. It has been acknowledged as a 'breathtaking tax development' (Ebrill et al. 2001) and the most significant tax structure advancement of the latter half of the 20th century (Cnossen 1991).[1] VAT is broad-based and prevents cascading taxes (tax on tax) and over-taxation. It is deemed superior to import or turnover taxes in terms of 'production efficiency' (Keen 2012) and more effective than retail sales taxes in terms of revenue mobilization (Kopczuk and Slemrod 2006). The VAT includes compliance incentives for business-to-business transactions and can generate revenue earlier in the supply chain, even if retailers evade tax liabilities. Given these qualities, VAT has the potential to raise a substantial amount of revenue even at moderate rates and in countries in the early stages of development (World Bank 1991).

Our central aim is to accurately measure the VAT gap, which measures the evasion behaviour of firms. Using this measure, we can assess the distributional implications of VAT under-reporting, identify the primary industries where compliance is challenging, and quantify the cost-effectiveness of the audit. The main intention of this is to understand three issues related to evasion: i) the incentives for it, ii) the efficiency of monitoring, and iii) the distributional (or inequality) consequences.

The VAT gap is the difference between the amount of VAT imposed, based on the tax system, and the amount collected. The VAT gap can be decomposed into policy and compliance gaps. The policy gap refers to the specifications in the tax code to determine the theoretical VAT liability. The compliance gap is the difference between the potential VAT revenue and VAT declaration. The compliance gap is further decomposed into three components (Durán-Cabré et al. 2019; Gemmell and Hasseldine 2014):

1. **Under-reporting gap:** VAT filers may under-report their tax liability to avoid paying taxes. This gap is measured by the difference between the tax declaration and audit information that uncovers undeclared VAT.
2. **Filing gap:** VAT-registered businesses may not file their VAT declaration, and non-registered businesses do not file and bunch below the VAT registration threshold (Gyoshev et al. 2023). The filing gap is the difference between the number of potential and actual VAT filers in the country.
3. **Revenue gap:** VAT filers declare their tax, but the business may not pay the tax liability within the required payment period. The revenue gap is the difference between the potential and actual VAT revenue.

This study is concerned with the first component, the under-reporting gap. For simplicity, the rest of the study refers to the under-reporting tax compliance gap as the VAT gap.

In 2019 the Tanzania Revenue Authority (TRA) studied the tax gap over different taxes, including the VAT (Tanzania Revenue Authority 2019). The study estimated a VAT gap of 36.3% in 2018

[1] Approximately two-thirds of least-developed countries have a VAT (Annacondia and van der Corput 2012). About 166 countries worldwide have adopted a VAT. All 54 African countries levy a VAT (Almunia et al. 2021; Crowe Horwath International 2016).

(compliance gap equal to 63.7%), equivalent to 2.1% of gross domestic product. The TRA study used macroeconomic data and conducted an aggregated top-down analysis.

This present study uses monthly VAT declarations in Tanzania between 2014 and 2020 and auditing data for specific tax regions simultaneously. For the auditing data the study team received data from six tax regions: four in Dar es Salaam (Illala, Kinondoni, Temeke, Tegeta) and from Singida and Dodoma. Due to data limitations the study is only concerned with the sample of firms represented by small, medium, and micro enterprises (SMMEs). This sample represents almost 50% of the total number of firms that fill out VAT forms and a substantial percentage of total output and input, conforming to a representative sample of SMMEs in Tanzania.

The availability of microdata and the aims described above permit the study to take a bottom-up approach to the VAT gap estimation. The study provides a valuable complement to past estimations by the TRA (Tanzania Revenue Authority 2019). It is relevant for the fulfilment of Addis Tax Initiative (ATI) Commitment 1 outlined in the ATI Declaration 2025, which points out that ATI members 'will enhance the effectiveness of partner countries' revenue administrations in curbing non-compliant behaviour by strengthening their capacities and capabilities, including risk management frameworks'.

The rest of the study continues as follows. Section 2 discusses the context related to the study. Section 3 presents the data used. Section 4 describes the methodology. Section 5 presents the results, and Section 6 concludes.

2 Context

2.1 Historical background

Following decades of VAT adoption in other countries across the world and in sub-Saharan Africa, Tanzania introduced the tax in 1998, replacing the sales tax. Initially, businesses with annual gross sales of more than TZS40 million were required to be registered as VAT agents. The wide adoption of the VAT was driven by its neutrality in international trade and the absence of distortion in domestic production and distribution (Fjeldstad 1995). Additionally, VAT provides stable revenue based on consumption, which tends to fluctuate very little, and is collected at different stages along the value chain, resulting in a broader tax base and more revenue than a sales tax.

Tanzania adopted consumption VAT (the most common type) using a credit–invoice computation method. The method works in such a way that the tax is levied on the total value of sales at each stage of production and allows a credit for any VAT paid on inputs in production (Fjeldstad 1995; Mrema 2012). Employing a credit method, VAT payable by the trader (i.e. the firm) is the difference between the tax collected on its sales (output tax) and the tax it paid on its purchases (input tax). Consequently, the consumer carries the ultimate VAT burden, while the merchant (i.e. the firm) acts as a tax collection agent.

The credit–invoice VAT system is advantageous in a tax jurisdiction with many traders (i.e. firms) with poor record-keeping capability. However, this system requires a significant audit trail, leading to administrative complexity and high costs. Tanzania's VAT compliance strategy is self-assessment based, where each VAT-registered trader declares the output tax, input tax, and tax payable. There is a compliance risk in the over-declaration of input tax or under-declaration of output tax, which becomes very high when there is a high level of informality and low usage of tax invoices and receipts, possibly undermining VAT collection (Sokolovska and Sokolovskyi 2015). VAT non-compliance mainly appears as traders overclaiming input tax, failing to file receipts, and

deliberately falsifying invoices and receipts, and collusion between traders and buyers (Fjeldstad et al. 2020; Wilks et al. 2019). In the context of high levels of non-compliance and weak tax administrative capacity, many eligible taxpayers may not register. They may fail to file returns and pay taxes even if they do. Sophisticated cross-checking and computerization must be employed in enforcing VAT compliance (Fjeldstad 1995). As such, Tanzania utilizes several ICT systems, including the ITAX system, the electronic filing of VAT returns, and electronic fiscal devices.

2.2 VAT system in Tanzania

VAT applies to all goods and services supplied or imported in Tanzania. The standard VAT rate in mainland Tanzania is 18%. However, the exports of some goods and services are zero-rated. Firms with a yearly taxable turnover exceeding TZS100 million in mainland Tanzania must register for VAT. The TRA can register investors who have not started producing taxable income and expect a negative relation but wish to be VAT-registered to reclaim the tax paid on start-up costs. A business that only sells exempt goods does not need to register for VAT and cannot recover the VAT paid on inputs. Registration is also compulsory for professional service providers (such as lawyers and accountants) and government-affiliated entities that conduct economic activities. Registered firms have to submit VAT returns monthly and pay the tax owed by or on the 20th of the subsequent month.

VAT and other duties are due on imported goods during importation. In contrast the VAT amount can be postponed for capital goods subject to agreement with the TRA. When it comes to imported goods, registered firms account for VAT through a reverse charge mechanism. However, this method is only feasible for firms with 10% or more exempt goods.

Despite all efforts VAT performance in Tanzania is low compared to other countries in the region. Cnossen (2019) presents VAT performance across African countries. The average VAT C-efficiency in Africa is 0.37, which is lower than the global average (0.47).[2] The C-efficiency is 0.21 in Tanzania, implying that only 21% of VAT is collected compared to the potential collections if all consumptions were taxed at the standard rate. The low performance of VAT collection results from tax exemptions, zero-rated items, and tax evasion. The low C-efficiencies should also be attributed to the policy gap, i.e. the numerous exemptions and zero rates on domestic products (Cnossen 2019).

A limited number of empirical studies have examined the difference between VAT paid and potential VAT in developing countries. In a recent study in Zambia using administrative tax data, Adu-Ababio et al. (2023) found a company income tax (CIT) and VAT tax gap of between 45% and 57%. For Tanzania, a report from the TRA in 2019 using a top-down approach studied the VAT gap between fiscal years 2016 and 2018 and found a tax gap of 36% for 2018 (and similar for 2016 and 2017). These results show the magnitude of the problem, reinforcing the necessity of studying the VAT. The effectiveness of the monitoring is relevant when resources are scarce and public spending is rising. Besides estimating the VAT gap, this is the central argument for the usefulness of the study's distributional effect.

[2] VAT C-efficiency refers to the ratio of VAT revenues to the potential VAT revenues, for instance the product of the VAT rate and the final consumption.

3 Data

3.1 Description

The VAT monthly return data is derived from the TRA's iTax system, which contains information on domestic VAT declarations excluding the revenue from taxes obtained at customs. The data contains 33 variables from 2015 to 2021 and includes 38,077 taxpayers identified by an anonymized taxpayer identification number (TIN). The variables are taken from the ITX240.02.B form[3] and are merged with the VAT registration data (form ITX245.02.E[4]) using the TIN. The two datasets provide detailed information on taxpayer sales and purchases along with background information such as tax region, business activity (ISIC 2-digit), and industry classification (ISIC 4-digit). This data allows us to aggregate the declarations to obtain the total VAT declared per year.

The audit data was collected from the annual tax audit reports prepared by tax offices for submission to the audit head office in Dar es Salaam. The dataset includes audit information from tax offices in Dar es Salaam (Illala, Kinondoni, Temeke, Tegeta), Singida, and Dodoma. While there are 33 tax offices in Tanzania, these six represent the most prominent (see Table A1 and Table A2 in the appendix). Variables include the start and end dates of the audit, the period audited (earlier years), the amount collected under each tax type (CIT, personal income tax (PIT), pay-as-you-earn (PAYE), VAT, Skills Development Levy (SDL), directors assessment, withholding tax, stamp duty, road licence, excise duty, and other), audit unit (small or medium taxpayer units), audit type (issue or comprehensive), and anonymized identifiers for audit officers. The information about the audited period and when the auditing was conducted enables us to determine the time (days) spent on each audit as the time between the beginning and end of the audit process. The audit data includes 2,901 unique taxpayers and 5,107 audits from 2018 to 2022. It also includes the anonymized TIN for audited firms, allowing it to be merged with the VAT declaration data and firm information. This process is conducted using anonymized TIN, following the procedure of the TRA.

The primary data source is the product of merging audit information, VAT declarations, and firm information. Through this the study creates a unique database (a panel) to follow firms from the return year 2014 to 2020. This period was chosen for two reasons: (i) to have more information about the declarations of firms and (ii) to have a considerable number of auditing processes. Information before 2014 and after 2020 is not considered as there were few firm declarations before 2014 and few auditing processes after 2020.[5] Firms with no VAT registration number (0.35% of the sample) are also excluded.

3.2 VAT and audit data description

Auditing of tax is delegated to each tax region to conduct and decide each step of the method. It is common practice for each tax region to prepare an annual audit plan and submit it to the head office. This plan establishes the characteristics, objectives, and audit steps for the corresponding fiscal year. Tax regions are required to meet specific collection targets each year. The head office reviews and approves the audit for implementation in that particular year. Every month the tax

[3] See https://www.tra.go.tz/domestic%20tax%20forms/ITX%20240.02.B%20-%20VAT%20monthly%20return.pdf.

[4] See https://www.tra.go.tz/index.php/forms/151-domestic-revenue-forms.

[5] Audits are potentially still ongoing for the period 2021, 2022 and 2023.

regions report the implementation status of the plan to the head office by showing the number of audits completed and the amount of tax collected.

The firms to be audited are selected based on a risk assessment at the tax region level. The risk assessment is based on turnover trends, taxpayer payments, and other factors. The audit cases included in the audit plan are typically intended for comprehensive audit analysis. Therefore they have different types of taxes, such as corporate tax, PIT, VAT, and PAYE. The plan includes other tax types, such as the SDL and withholding tax. Following this, a control verification exercise is conducted for a particular type of tax. This exercise typically entails verifying tax compliance by requesting short-term information from taxpayers consisting of data for one month or a few months. Taxpayers may be audited in consecutive years.

Some definitions are formalized to simplify the explanation of the data and the subsequent explanation of the result. First, a return year *t* is the period between 1 July of year *t-1* and 30 June of year *t*. For example, the return year 2020 is from 1 July 2019 to 30 June 2020.

Second, an audited firm is a firm audited between 2014 and 2020. Hence a firm audited for 2018 will be recognized in the analysis as an audited firm from 2014 onwards. Third, an audited process means conducting and completing the audit for the corresponding return year. This means that a firm could be audited in 2020 to inspect the return year 2015 so that the audit process corresponds to 2015.

Finally, as it is possible to identify the tax collected in each auditing process, a compliant firm shows zero VAT collected in the corresponding auditing process. In other words a compliant firm is a firm that was audited but for which no undeclared VAT was found during the auditing process, regardless of whether other non-paid taxes, such as CIT or PIT, could be found.

Table 1: Audit status across tax years

	Number of firms				Audit process		
Return year	Audited firms	Unaudited firms	Total	% Audited	Completed audits	VAT compliant	% VAT compliance
2014	1,507	6,342	7,849	19.2%	310	31	10.0%
2015	1,671	6,923	8,594	19.4%	586	54	9.2%
2016	1,824	7,488	9,312	19.6%	821	66	8.0%
2017	1,949	8,060	10,009	19.5%	735	68	9.3%
2018	2,052	8,794	10,846	18.9%	787	68	8.6%
2019	2,120	9,765	11,885	17.8%	839	86	10.3%
2020	2,106	11,054	13,160	16.0%	904	148	16.4%

Note: the tax year runs from 1 July to 30 June. The VAT compliance percentage is the number of VAT-compliant firms over the number of completed audits.

Source: authors' calculations based on TRA VAT and audit data.

Table 1 gives a summary of the definitions mentioned above across return years. The first four columns show information about the number of firms per year. Around 18.6% of the firms belong to the group of audited firms and follow a decreasing pattern across years. However, the number of audited firms increases across the years, indicating that the growth of this group is slower than in the unaudited firm group. The subsequent columns show the auditing process information: the number of completed audits and VAT-compliant firms per year. Around 10% of the completed audits show firms that were VAT compliant, with an oscillating pattern similar to a U-shape. Interestingly, the number of completed audits does not increase over the years; instead there is a decrease in 2018 and a return to an increase in 2019.

Table 2 shows the total number of audited and unaudited firms, and the audits per tax region and economic activity, each with the corresponding percentage from 2014 to 2020. The region with the most firms is Ilala, followed by Kinondoni. Both regions also show a high audit rate, but Temeke has the highest rate. Regarding audits, in most regions the rate (audits/number of audited firms) is high, greater than 38% in four out of six regions. Concerning the economic sector, most of the audited firms are in the accommodation, arts, health, admin, professional, and wholesale and retail sectors. The sectors with the highest numbers of audited firms are the transportation and storage and manufacturing sectors. The percentage of completed audits is high across activities, averaging 39%.

Table 2: Number of firms by tax region and economic activity

	Audited	%	Unaudited	%	Total	Completed audits	%
Tax region							
Dodoma	305	10.3 %	2,645	89.7 %	2,950	68	22.3 %
Ilala	4,520	16.9 %	22,190	83.1 %	26,710	1,152	25.5 %
Kinondoni	4,297	16.3 %	22,046	83.7 %	26,343	1,907	44.4 %
Singida	103	16.1 %	535	83.9 %	638	40	38.8 %
Tegeta	501	9.3 %	4,895	90.7 %	5,396	200	39.9 %
Temeke	3,503	36.4 %	6,115	63.6 %	9,618	1,615	46.1 %
Economic activity							
Agriculture	69	15.4 %	379	84.6 %	448	28	40.6 %
Mining	258	20.8 %	983	79.2 %	1,241	104	40.3 %
FIRE	891	22.7 %	3,029	77.3 %	3,920	348	39.1 %
Construction	1,255	16.6 %	6,328	83.4 %	7,583	475	37.8 %
Wholesale & retail	3,263	17.7 %	15,136	82.3 %	18,399	1,138	34.9 %
Manufacturing	1,750	24.3 %	5,447	75.7 %	7,197	744	42.5 %
Information & communication	571	17.5 %	2,685	82.5 %	3,256	188	32.9 %
Transportation & storage	1,652	31.2 %	3,638	68.8 %	5,290	642	38.9 %
Accommodation, arts, health, admin, professional	3,301	14.9 %	18,919	85.1 %	22,220	1,217	36.9 %
Public admin	154	19.3 %	645	80.7 %	799	69	44.8 %
Household, extraterritorial, non-business	65	5.0 %	1,237	95.0 %	1,302	29	44.6 %

Note: the tax year runs from 1 July to 30 June. The audits completed relate to the number of firms audited in the corresponding year. FIRE means the financial, insurance and real estate sectors.

Source: authors' calculations based on TRA VAT and audit data.

Table 3 shows the summary of VAT declarations without credits claimed for audited and unaudited firms in tax regions and economic sectors. It also outlines the evaded amounts. Ilala and Kinondoni have the highest declared VAT values and total amounts evaded. Some firms report negative VAT due to more extensive input than output VAT. This negative VAT occurrence is most prevalent in Singida.

Some of the details mentioned earlier still hold from a sectoral perspective—the difference in signs between audited and non-audited firms. The sectors with larger VAT payments are the accommodation, arts, health, admin, professional, and wholesale and retail sectors. Additionally, some sectors, such as agriculture, transportation, and storage, show different means between audited and unaudited firms. This suggests differential behaviour that drives auditing.

Table 3: VAT declared by tax region and economic activity

	Audited			Unaudited			Evaded		
Tax region	Mean	sd	Total	Mean	sd	Total	Mean	sd	Total
Dodoma	4.6	45.2	15,097	0.7	32.6	16,579	1.9	2.8	1,476
Ilala	8.4	141.2	439,341	1.5	116.9	358,904	7.8	22.8	105,613
Kinondoni	8.4	105.6	413,350	1.6	72.1	358,429	6.4	23.5	141,639
Singida	-0.2	16.5	-267	0.3	8.0	1,665	0.3	0.6	136
Tegeta	12.8	76.8	73,054	1.3	22.0	66,741	6.4	29.1	14,849
Temeke	7.9	94.4	318,325	1.4	165.7	89,751	2.6	6.8	49,921
Economic activity									
Agriculture	-5.3	66.3	-3,926	3.3	52.9	11,679	1.7	3.2	553
Mining	2.3	145.4	6,579	1.9	80.9	19,808	5.2	16.8	6,306
FIRE	4.3	80.6	45,101	2.4	36.3	81,249	4.4	15.0	18,165
Construction	6.3	219.1	90,635	1.0	149.1	68,777	6.6	20.2	36,313
Wholesale & retail	6.6	97.0	247,710	0.6	19.0	98,809	5.6	22.4	75,233
Manufacturing	8.9	98.0	179,183	1.3	29.0	71,858	3.3	8.7	29,135
Information & communication	16.4	127.8	106,787	2.7	27.3	76,763	6.8	15.3	14,919
Transportation & storage	5.8	68.7	111,064	-0.4	251.3	-14,249	5.1	19.6	38,359
Accommodation, arts, health, admin, professional	12.0	102.3	458,025	2.4	103.5	466,350	5.9	22.4	83,745
Public admin	8.5	104.6	14,974	1.6	40.5	10,677	5.3	8.3	4,208
Household, extraterritorial, non-business	3.7	133.8	2,763	0.0	41.1	347	19.8	49.2	6,695

Note: the monetary amounts are expressed in TZS millions. FIRE means the financial, insurance and real estate sectors.

Source: authors' calculations based on TRA VAT and audit data.

Finally, Table A1 and Table A2 in the appendix show the evolution of critical variables for our study. Table A1 summarizes the tax regions, and Table A2 compares audited and unaudited firms in audited tax regions. The audited tax region represents a meaningful percentage of the total amount on variables such as VAT payable. However, it is essential to see that audited firms follow a similar pattern to non-audited ones, suggesting that they are similar to non-audited ones. This similarity is crucial to the methodology described in the next section.

4 VAT compliance gap methodology

4.1 Tax gap and efficiency measures

The study considers two classifications for the analysis of the VAT gap and the effectiveness of the audit process. The first is the VAT gap, which has two components. The first component relates to evasion or non-paid taxes, termed 'VAT recovered' in Equation (1). This variable is the amount discovered through auditing as non-paid VAT. The second component is the VAT declaration, which provides the information on VAT declared for the corresponding period. The VAT gap, as recently calculated in ratio terms by Best et al. (2021), is as follows:

$$VAT\ GAP = \frac{\sum VAT\ Recovered}{\sum VAT\ Recovered + \sum VAT\ Declared} \quad (1)$$

where the VAT recovered is the extra VAT collected after an audit. Equation (1) shows the total VAT gap for a particular period, for example, a return year. Hence, it is necessary to sum each variable across firms for the required period. This equation is similar to the C-efficiency rate (Ebrill et al. 2001). In both cases the main point is to capture the part of the potential revenue that is not collected due to under-reporting or evasion.

The second measure refers to the cost of auditing and the efficiency of the entire process. As an explicit cost variable for the auditing process is unavailable, we use the time spent on auditing. This assumes that the hourly salary of an auditor is constant, thereby giving the difference in cost that comes from the days the auditor needs to work. Hence, any difference is given by the need for the audit process to be more exhaustive for some sub-groups, meaning that there is a difference in the number of days taken to conduct the audit. The numbers of VAT declarations in the audited periods are used to normalize the days spent on auditing. This measure gives an average cost and captures the number of days that the TRA spent auditing a firm per declaration made in the same period. We call this ratio the cost of auditing ratio, and the formulation is as follows:

$$\frac{Days\ taken\ to\ conduct\ the\ audit}{Number\ of\ declarations\ in\ the\ auditing\ process} \tag{2}$$

This ratio captures the cost, measured by auditing days, of auditing one monthly VAT declaration. For example, a rate of 9 indicates that auditing the firm took nine days per VAT declaration made in the same period.

Lastly, the study defines an overall cost-effectiveness measure. The cost-effectiveness ratio is obtained by dividing the VAT gap by the cost of auditing. The efficiency measure is as follows:

$$\frac{VAT\ gap}{Cost\ of\ auditing} \tag{3}$$

This ratio captures the amount of the VAT gap discovered per audit cost. For example, a ratio of $28/10 = 2.8$ means that 2.8 percentage points of the VAT gap are discovered per ten days of auditing cost. Recall that the auditing cost is normalized for the number of VAT declarations produced. In the example, for a monthly VAT declaration, ten days of auditing (the cost) is needed to uncover 2.8 percentage points of the VAT gap (benefit).

4.2 Bottom-up approach

The bottom-up approach is a micro approach for investigating tax gaps exacerbated by evasion and avoidance. It provides better diagnostic information about the differences in and quantitative significance of the VAT gap. For further insights on this approach, see Almunia et al. (2021), Pomeranz (2015), and Slemrod et al. (2001). The bottom-up approach is more commonly used by revenue authorities which conduct random audits. In both random and risk-based audits, tax gap estimates are obtained by inferring the audit results to the total population under certain conditions (Barra et al. 2023)

In the case of Tanzania, audits are risk-based in that detected evasion from audits does not represent evasion for all taxpayers. Where audits are risk-based the VAT gap can be estimated using techniques that infer taxpayer characteristics based on the audit sample's observations. Several prediction methods, each with their own advantages and disadvantages, are described in more detail by Barra et al. (2023). Based on the data at hand and the context, this study pursues an approach which uses a machine learning method.

The bottom-up approach uses information compiled after auditing to estimate the tax gap. The methodology uses the VAT recovered (the extra VAT collected after an audit) and the VAT declared in the return form for the corresponding period to estimate the potential amount of under-reported VAT for all VAT-registered firms. All VAT-registered firms have information on VAT declared, but only audited VAT-registered firms have VAT recovered amounts. By using predictive methods such as regression analysis and machine learning, among others, it is possible to estimate the potential amount of VAT recovered.

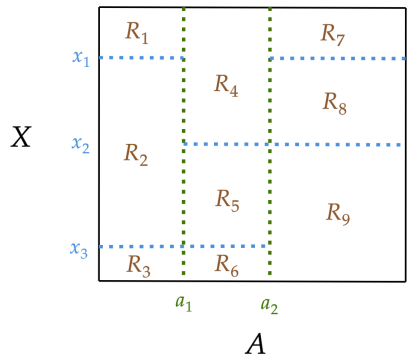
The bottom-up approach predicts the VAT amount recovered for each firm by using the available information. Adding together the VAT recovered amounts for all firms gives the total VAT recovered and the corresponding tax gap estimation. As the estimations are at the micro level, it is possible to obtain estimations at the sector, region, municipal, and macro levels. However, the accuracy of the estimation heavily relies on the prediction method used. For this reason it is necessary to go further in this step and to use methodologies that ensure highly predictive accuracy.

4.3 Machine learning estimation

This study follows a machine learning (ML) approach to predict the amount of non-paid VAT in unaudited firms and unaudited periods. The ML approach has high predictive accuracy in comparison to other methodologies used in economics (see Athey and Imbens (2019) and Mullainathan and Spiess (2017) for a more detailed explanation). This methodology uses an algorithm to improve the prediction of a variable based on information that the user provides. With this information the algorithm (machine) iterates until it finds the best combination of variables for an accurate prediction (learn). In this case the information is the VAT declaration and audit data, and the algorithm finds the best variable combination to predict the non-paid VAT.

The study follows the random forest algorithm (Breiman 2001) and implements the package described in Schonlau and Zou (2020). This algorithm uses multiple decision trees to predict a variable using available information. The method uses a sample and bootstrapping for each tree to estimate a prediction. This means that it randomly selects data points from the dataset to create a new sample, with the probability of some points being selected multiple times while others might not. For each tree the method uses one sample to train the tree. This process of bootstrapped sampling and training trees is repeated many times. Each time, the method looks at the combination of variables that produces the lowest prediction error. This error is measured by RMSE (root mean square error), which tells us how far off our predictions are from the actual values. Simultaneously, the method decides the optimal number of variables for each iteration. In this sense the method learns the relevant variables to predict the required variable more accurately.

Figure 1: Random forest illustration



Source: authors' elaboration.

Figure 1 presents an example of this methodology. Let us suppose that we have a target variable we want to predict and two covariates, X and A. The method splits each covariate n times, creating different spaces and denominating them by R in the figure. By dividing covariate X into three and covariate A into two, we create nine subsamples, each representing a combination of X and A. These subsamples, or regions, are labelled as R_1 through R_9. For each of the nine subsamples we calculate the average value of the target variable and, later, the average of those estimations is obtained. It is possible to iterate this using bootstrapping and obtain resamples. As only part of the covariate is used in each iteration, we ensure that the samples are not too similar. This helps to reduce the correlation between the samples and makes the model more robust and less likely to overfit the data. Finally, we use cross-validation, using a separate sample of data, to test the model's predictions. This helps us to determine how well the model performs on new, unseen data. By testing on separate data, it is possible to choose the best combination of splits and iterations that minimize the prediction error. It is important to note that, as we are iterating and using resamples, the model is more flexible than a linear one as it tests different alternatives or combinations of variables. The machine learns the relevance of this interaction and decides the weight of the variable. In this sense the method chooses which variables help to produce an accurate prediction.

The ML strategy used can be summarized in the following three stages:

Stage one: Choosing and training the ML model

In this stage the ML methodology chooses two critical parameters: the number of iterations in each forest branch and the number of independent variables randomly selected at each split.

1. Splitting the data: Dealing with the audited data only. The sample is divided into two parts: the training sample, which consists of 80% of the data, and the testing sample, which consists of the remaining 20%.
2. Tuning the number of iterations: In this step we determine the best number of times the random forest model should repeat its process (iterations) to make accurate predictions. The process involves running the model with different iteration counts (from 10 to 500) and checking how well it predicts the evasion.
3. Tuning the number of variables to use in each split: In this step, the ML model determines the best number of variables to select randomly at each split in the random forest model. The process involves running the model with different numbers of variables and checking how well it predicts the evasion.

In steps 2. and 3. the model measures two types of errors: out-of-bag and validation RMSE errors. These errors indicate prediction accuracy using data not included in the model's training process. Those two critical parameters are chosen according to the minimum prediction errors.

Stage two: Testing the model

Here the methodology uses the critical parameters obtained from the first stage and tests the ML model by comparing the performance of the random forest model and the regression model using the RMSE.

Stage three: Estimating the VAT gap using the trained random forest model

1. Apply the random forest model to predict undeclared VAT amounts for non-audited and audited firms when no audit was carried out using the whole data sample and the predetermined parameters from the first stage.
2. The actual and predicted evasion rates are used to calculate the VAT gap, as presented in Section 5.

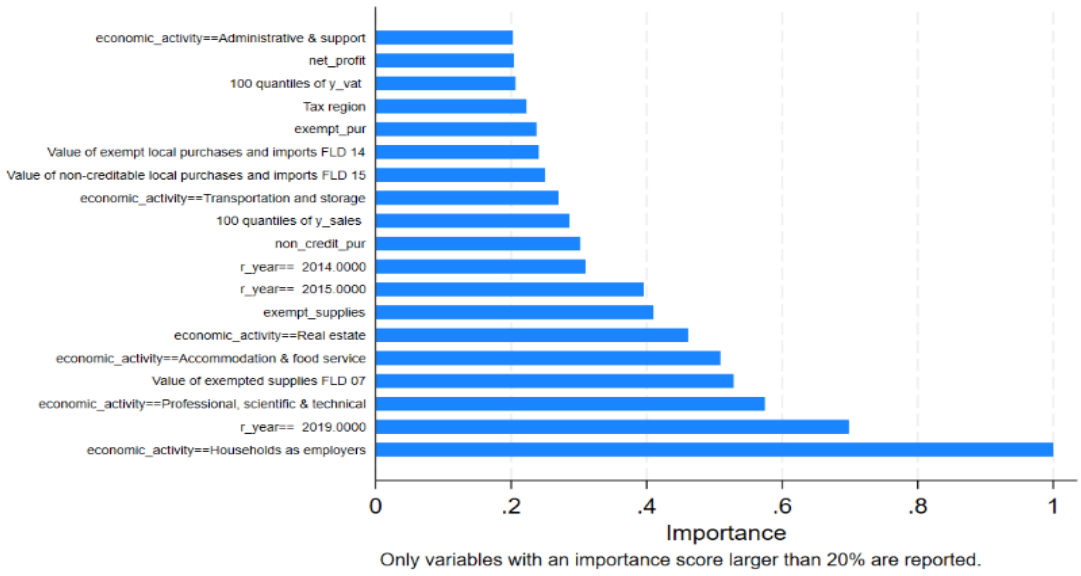
The model we estimate can be summarized as follows:

$$Evasion_{ikt} = f(sales\ variables) + f(purchases\ variables) + f(firm\ characteristic\ and\ data)$$

In the estimation the variable to predict is $Evasion_{ikt}$, which is the undeclared VAT amount for firm *i*, in tax region *k*, and in time *t*. As the audited period and the return year are known, it is possible to obtain the average amount evaded for this period by dividing the non-paid VAT found by the number of tax declarations submitted by firm *i* in this period. Regarding the variables used to predict evasion, these include all the variables available in the VAT declaration (purchases and sales of taxable, exempted, and zero-rated products); a set of firm variables (tax region, city, economic activity (2-digit ISIC), business activity (4-digit ISIC), month and return year of the declaration); and a set of variables for the firm's behaviour (proxy of profits (total sales–total purchases), percentile of yearly VAT declaration, and percentile of yearly sales).

Given the data, the study can approximate the monthly undeclared VAT using the information provided by the auditing process. The total amount of VAT discovered is divided by the number of VAT declarations that the firms made during the period inspected by the audit. Through that the study obtains the average amount of non-paid VAT during the period, giving a prediction for the monthly declaration and enriching the analyses. Predictions are checked against the actual values. The ML predictions outperform all other methods.

Figure 2: Relevance index in ML estimation



Source: authors' calculations based on TRA VAT and audit data.

Figure 2 shows the most relevant variables accompanied by the weight that the method gave them. This can be interpreted as a relevance index. The most relevant variable is households as employers

as an economic activity, followed by the return year 2019. Interestingly, the value of exempted supplies is highly relevant, although they are not part of the taxes supplied.

Finally, Table A3 and Figure A2 in the appendix demonstrate the accuracy of the prediction. Table A3 shows that, when comparing the R-square and the RMSE between the training (where the ML is estimated) and the testing (where the ML is tested) sample, the ML produces a more similar prediction to the actual value.[6] Figure A2 plots the prediction and the actual values for the audited observations, providing visual evidence of the accuracy of the estimation.

In short, the bottom-up approach provides a fine-grained analysis that is useful for making tax policy and is recommended by Barra et al. (2023). As only a small sample of firms are audited and selected based on their risk profile, prediction is required to measure the potential VAT that can be recovered through an audit process. This is required for all unaudited firms and audited firms in unaudited periods. Having set out the methodology and measures employed, the study next describes the analysis.

5 Analysis and estimation

5.1 VAT gap estimation

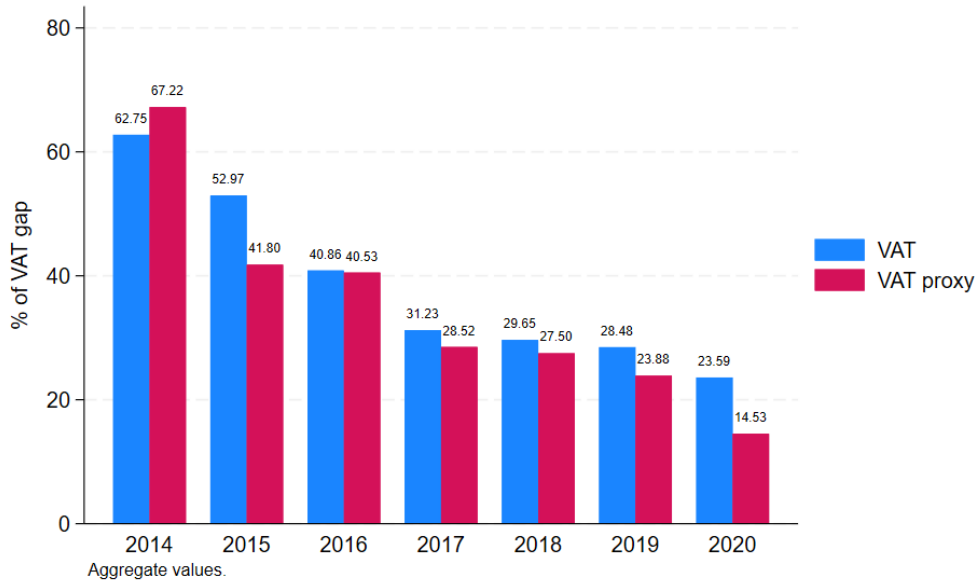
5.1.1 VAT gap for audited firms

The aggregate results between 2014 and 2020 are analysed using only the sample of audited firms. The VAT gap is plotted following Equation (1). The study considers two calculations for the VAT amount: VAT declared in the VAT return data (blue bars in Figure 3) and VAT proxy calculated as sales VAT minus purchases VAT from the VAT return data (red bars in Figure 3).

Figure 3 shows the primary results: using both estimations of the VAT, the VAT gap for audited firms is declining over time. Both variables show the same patterns and similar levels. This is explained by the increased number of days spent during the audit process. However, as the amount of VAT recovered during the audit process is decreasing and costs are increasing, the audit processes are becoming less efficient in recovering VAT, partly explained by the increase in audit costs (calculated as the number of days taken to conduct the audit).

[6] Training and testing samples are part of the calibration of the model. The model is estimated in the training sample and later tested in the testing sample. This procedure iterates until the smallest RMSE is obtained. Both samples are part of the group of audited observations, providing a value for the non-paid VAT. After this process the model is calibrated and it is possible to predict the non-paid VAT in the unaudited observations.

Figure 3: VAT gap for audited firms



Note: yearly aggregate data. The blue bars represent VAT declared on the VAT return, while the red bars represent the calculated VAT proxy (sales less purchases).

Source: authors' calculations based on TRA VAT and audit data.

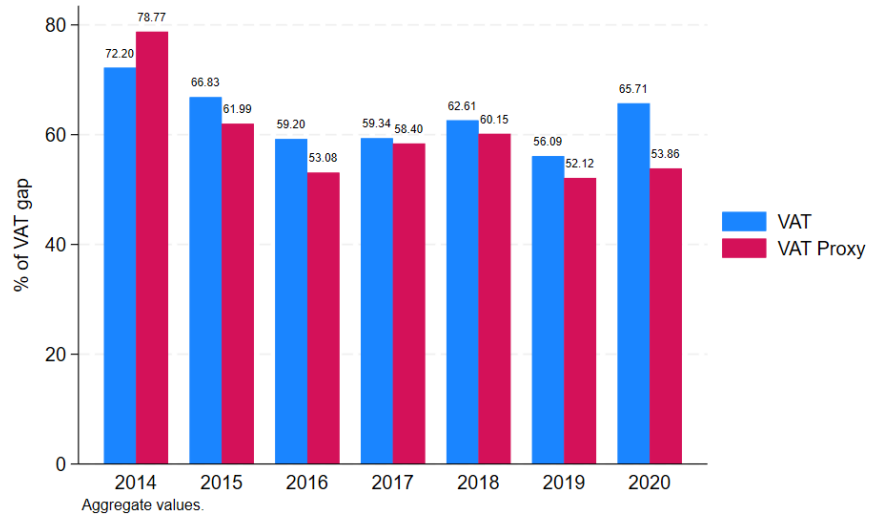
These results only reflect the audited cases. No revenue authority audits all firms. The study follows the estimation approach described in Section 4.2 to extend the analysis to non-audited firms and unaudited periods. The method calculates a potential VAT recovered amount for unaudited firms using all the information from the tax data. This calculation is used to calculate the VAT gap. This methodology is vetted and recommended by the IMF to understand the source of the compliance gap (Barra et al. 2023) better. The following section describes the VAT gap for all VAT-registered firms.

5.1.2 VAT gap for all firms

This section presents the results for all firms in all periods studied for the audited tax regions. As there is no information for all firms during all the periods studied, the potential amount of VAT recovered for unaudited firms results from the ML method. This allows for the estimation of the VAT gap for all VAT-registered firms.

Figure 4 shows the VAT gap for all firms from 2014 to 2020. The average VAT gap is 62% for all firms. This differs considerably from the average VAT gap of 38% for audited firms, described in the previous section. The estimation method partially explains the difference. The ML approach accounts for efficiency changes, ignoring them and predicting evasion that is not sensitive to efficiency changes. As the VAT gap rises, the same patterns appear for efficiency. The amount of evasion per audited day and the VAT gap per audited day are found to be decreasing. This reinforces that increasing the cost of auditing has critical consequences for the erosion of VAT revenue.

Figure 4: VAT gap for all firms



Note: yearly aggregate data. The blue bars represent VAT declared on the VAT return, while the red bars represent the calculated VAT proxy (sales less purchases).

Source: authors' calculations based on TRA VAT and audit data.

As audits are determined within each tax region, the study does not extrapolate to the tax regions where audit data is unavailable. However, the data description shows that the selected regions reasonably estimate all of Tanzania. Precise estimation for each tax region would require dedicated audit data from all tax regions.

5.1.3 VAT gap by sector, firm size, and declaration amount

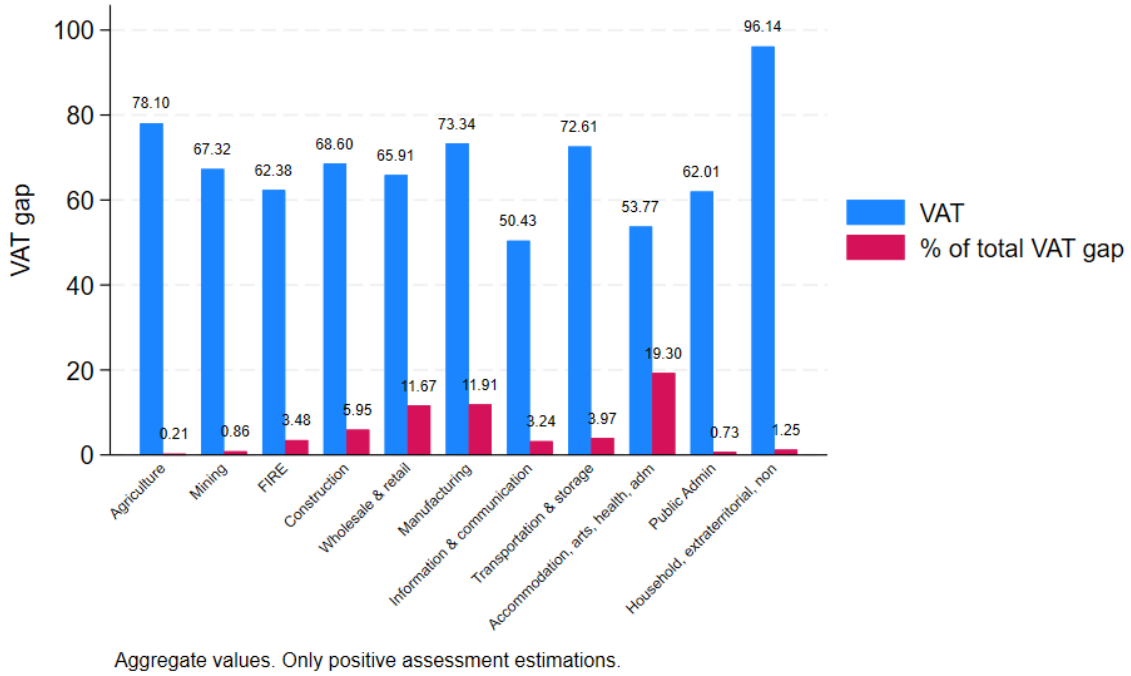
The strength of the methodology enables the estimation of the VAT gap by sector, firm size, or other relevant categories. This section describes the VAT gap results by sector and firm size.

Figure 5 shows the estimation for the entire period by industry. The average VAT gap by economic activity is 68% (blue bars), showing that evasion behaviour is a critical issue within industries. The sector with the largest VAT gap is households, followed by agriculture. The households sector encompasses self-employment activities associated with self-reporting. The agriculture sector includes many exemptions that reduce tax liability. The rate gap is also shown as an indicator of the contribution to the overall VAT gap (red bars).

While agriculture appears to have a very large VAT gap, the contribution of that agricultural VAT gap to the total VAT gap (red bar) is low. When considering which sector contributes the most to the overall VAT gap, the biggest contributor is the accommodation, arts, health, admin, and professional activities sector.

This indicates that the agricultural sector could present a substantial informality rate and, to combat this, it would be useful to conduct a closer study of this sector. To increase revenue, it would be better to focus compliance efforts on the wholesale and retail and accommodation, arts, health, admin, and professional sectors.

Figure 5: VAT gap by Industry

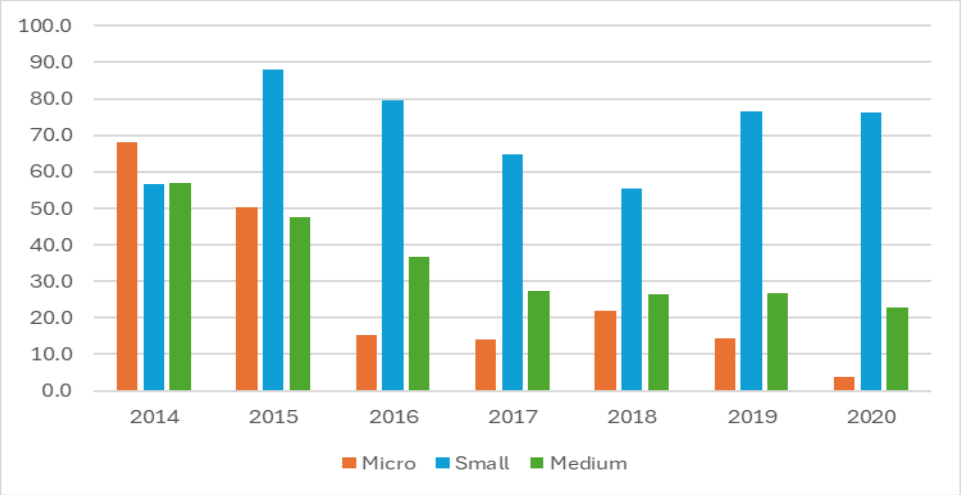


Note: amounts are expressed in TZS millions. Variables are the total for the return year period 2014–2020. FIRE means the finance, insurance and real estate. Only positive VAT declarations are considered.

Source: authors' calculations based on TRA VAT and audit data.

To unpack the underlying causes of the VAT gap, the study focuses on the whole sample in two divisions based on the yearly VAT declaration and sales (see Brockmeyer et al. (2024) for an explanation of the usefulness of this). For sales, the study considers the annual total output as a proxy for sales and divides the sample into three using the sales distribution. Hence, for each year, the sample is split into three groups that capture a third of the sales in the corresponding year. As there are no large firms in the sample, we classify the groups as micro, small, and medium-sized firms. Figure 6 shows the VAT gap across years by firm size.

Figure 6: VAT gap by firm size

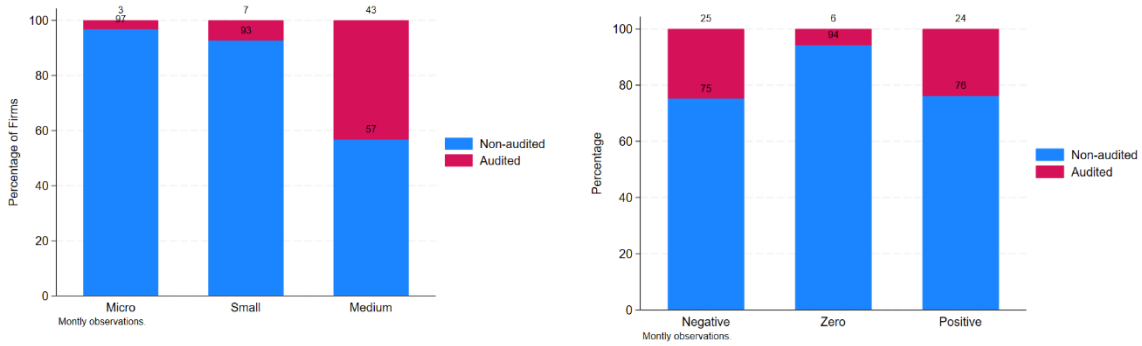


Note: only positive VAT declarations are considered.

Source: authors' calculations based on TRA VAT and audit data.

The VAT gap is large for micro and medium-sized firms in 2014 but decreases over time and is high for small firms and remains high over the study period. More than 40% of medium-sized firms are audited while the revenue collection from micro firms may be too small against the cost of audit. Only 7% of small firms are audited and display a high VAT gap, signalling the need for further investigation to consider whether additional audits are recommended for this group.

Figure 7: Percentage of audited and unaudited firms by firm size and VAT amount declared



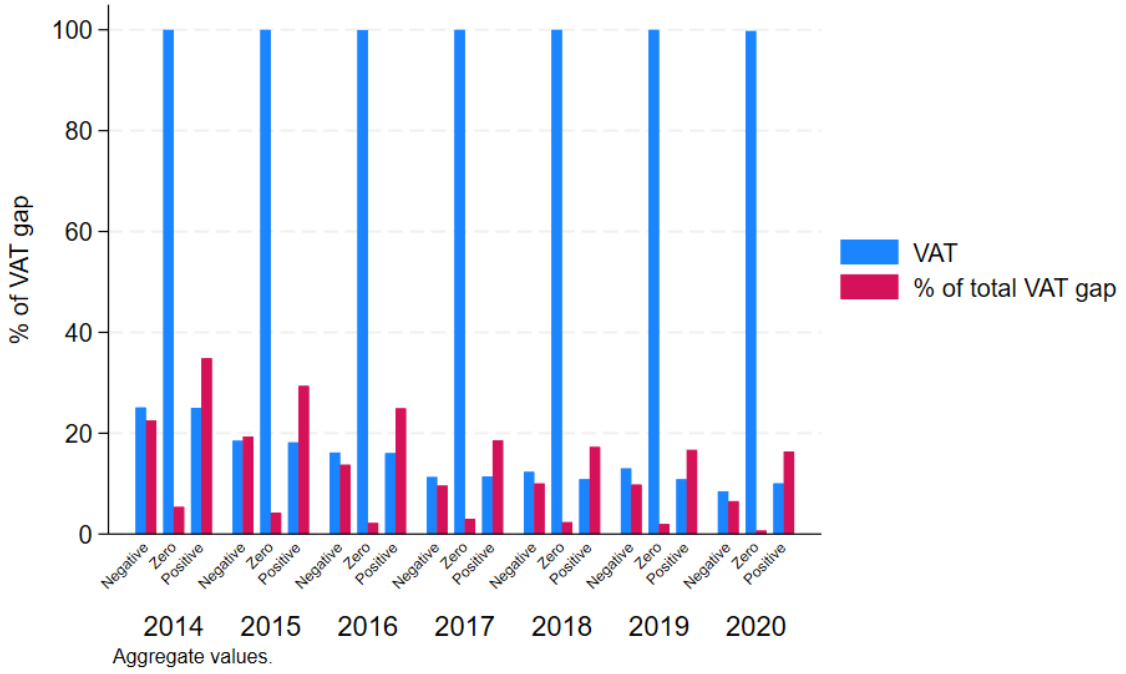
Source: authors' calculations based on TRA VAT and audit data.

Lastly, the study considers the reporting VAT amount to split the group into three categories. Recall that in Section 2.2. the study described the invoice–credit VAT system in Tanzania. There is a compliance risk in the over-declaration of input tax or under-declaration of output tax, as VAT-registered traders declare their output tax, input tax, and tax payable independently. The large informal sector in Tanzania and suboptimal usage of tax invoices and receipts increase the risk of undermining VAT collection (Sokolovska and Sokolovskyi 2015). VAT non-compliance mainly appears as traders overclaiming input tax, failing to issue receipts, deliberately falsifying invoices and receipts, and potentially colluding with buyers (Fjeldstad et al. 2020; Wilks et al. 2019).

First, Figure A1 (in the appendix) shows the distribution of firms by VAT amount declared. The figure shows that firms bunch around zero VAT declaration (without credit claiming). This means that more firms report zero VAT than positive or negative VAT. The sample is divided between negative, zero, and positive VAT declarations. The monthly mean of TZS66,689 is used as a threshold, such that the zero-group threshold is between the mean and is negative.

The study examines firms in the negative VAT category (referring to firms that claim more input tax than output tax), zero VAT (firms that claim equal, or close to equal, parts of input and output tax), and positive VAT (output tax greater than input tax). This enables examination of the differential behaviour by firms should the audit treatment rely on the declared VAT amount. Figure 7 shows that the audit rates for negative and positive VAT groups are around the same; the audit rate for firms that declare zero VAT is very low. If firms know that the probability of being audited is low when they declare zero VAT, then the incentive to declare zero, or close to zero, increases and, with it, the possibility of VAT evasion. This result emerges in Figure 8. VAT-registered firms that report zero, or close to zero, VAT are found to have the highest VAT gaps. However, the zero VAT groups of firms contribute very little in terms of the overall VAT gap. Instead, firms that declare a positive VAT show the largest participation in the overall VAT gap. Regarding the VAT declaration, the largest VAT gap is in the group that declares around zero VAT. This number has been stable over the years. For the other groups the VAT gap has decreased over the years.

Figure 8: VAT gap by groups that declared VAT



Source: authors' calculations based on TRA VAT and audit data.

This section of the study demonstrated the VAT gap for audited firms, all firms, and groups of firms to examine the source of VAT non-compliance. The next section considers the efficiency of audits in light of the VAT gap analysis.

5.2 Discussion

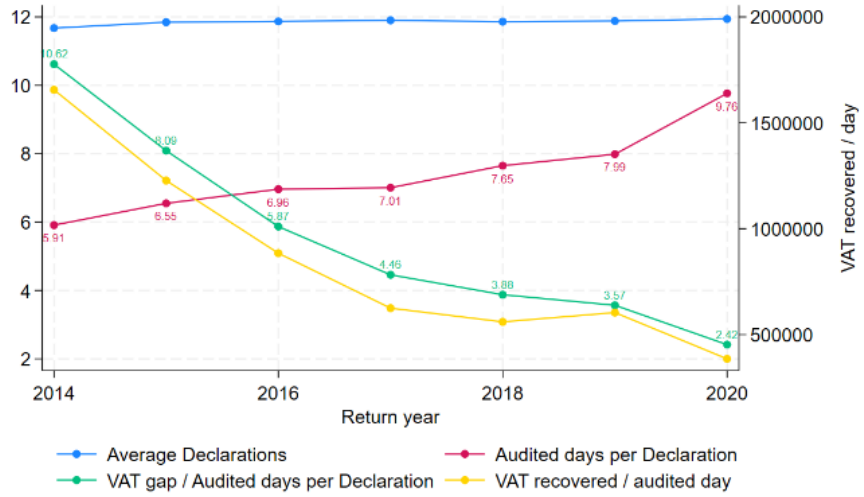
The aim of the study, methodology, and use of microdata is to understand the source of the VAT gap and suggest how the gap can be reduced. This subsection explains the source of the gap by considering the reporting and evasion behaviour of firms. To understand where resources can be reconsidered or reallocated, the study calculates the efficiency measures described in Section 4.1.

The audit data is rich in that it includes the start and end period of the audit examination. For each audited taxpayer, the length of the audit (number of days) can be checked against the amount of VAT recovered during the audit. The cost of auditing ratio is the number of days taken to complete the audit examination over the number of declarations made in the audit examination period and refers to Equation (2) in Section 4.1. This measure gives an average cost, capturing how many days it took to audit a firm per declaration made in the same period. For example, a rate of 9 indicates that auditing the firm took nine days per VAT declaration made in the same period.

The cost-effectiveness ratio is the VAT gap divided by the cost of auditing, as referred to in Equation (3) in Section 4.1. The ratio captures how much of the VAT gap is discovered per audit cost. For example, a ratio of 28/10 = 2.8 means that 2.8 percentage points of the VAT gap are discovered per ten days of auditing cost.

The efficiency analysis is composed of different ratios, where the critical elements are the auditing days per firm's declarations (red line), the evaded amount discovered per auditing day (yellow line), and the VAT gap discovered per auditing process—the efficiency measure (green line). Figure 9 shows the first results of the efficiency measures.

Figure 9: Efficiency measures over time



Source: authors' calculations based on TRA VAT and audit data.

First, note that the average yearly declaration (blue line) is stable, but the number of auditing days per declaration increase (red line). This means that the cost of auditing has been increasing over the years. The VAT recovered per day audited (yellow line) has decreased across the years, meaning that less VAT is being discovered per day. The VAT gap over audited days per declaration (green line) has also decreased over time, indicating that it has become more difficult to find undeclared VAT over the years. A decreasing trend shows that the auditing process is discovering less VAT gap per audited day. This is explained, in part, by the increased auditing costs depicted by the red line.

Figure 3 and Figure 4 show that the VAT gap has decreased over time but remains substantial. At the same time, the efficiency measures show the rising costs of auditing. It is possible that firms are finding more sophisticated ways to evade paying taxes, making it harder for undeclared VAT to be discovered.

The study considers the cost of auditing and the efficiency measure, by industry. The average VAT gap by economic activity is 68%, with all the industries showing a rate larger than 50%. This shows that evasion is a common issue across industries. The sector with the largest VAT gap is households, followed by agriculture. The mining industry shows a large efficiency rate (last column in Table 4), indicating that targeting those firms could increase the amount of VAT recovered at a lower cost. However, we also provide the rate gap to identify the industries that make the largest contribution to the erosion of VAT collection. The accommodation, arts, health, admin, and professional, the manufacturing, and the wholesale and retail sectors have the largest rate gaps. This suggests that agriculture could present a large informality rate (similar to household as an industry) but that auditing the agriculture sector would only make up a small proportion of the VAT gap. On the other hand it is better to focus on the wholesale and retail, and accommodation, arts, health, admin, and professional sectors to increase revenue. An interesting case is the manufacturing sector, which shows a large VAT gap and rate gap, indicating that increased auditing of those firms would reduce informality and increase revenue simultaneously.

Table 4: Summary of VAT gap and efficiency by economic activity

Economic activity	VAT gap	Rate gap	Audited day/ declaration	Evasion/ audited day	VAT gap/ efficiency
Agriculture	78.2	0.2	5.7	7,569,094	13.7
Mining	67.7	0.9	6.2	6,680,066	11.0
FIRE	62.2	3.5	7.2	6,443,763	8.6
Construction	69.0	6.1	7.3	8,434,876	9.4
Wholesale & retail	66.0	11.8	7.4	6,824,557	8.9
Manufacturing	71.8	11.1	8.2	8,764,143	8.8
Information & communication	50.5	3.3	7.3	11,158,407	6.9
Transportation & storage	72.7	4.0	8.3	3,634,134	8.7
Accommodation, arts, health, admin, professional	53.9	19.5	7.0	10,749,831	7.7
Public admin	62.2	0.7	7.3	7,047,842	8.6
Household, extraterritorial, non-business	96.0	1.2	3.4	56,822,220	28.2

Note: amounts are expressed in TZS millions. Variables are the total for the return year period 2014–20. FIRE means the finance, insurance, and real estate sector.

Source: authors' calculations based on TRA VAT and audit data.

The literature shows that not all firms behave in the same way with regard to tax avoidance and evasion. This study therefore looks at the behaviour of firms in relation to firm size. The same efficiency measures described above are calculated for micro, small, and medium-sized firms across the study time period. Table 5 shows the results for audited firms, where Panel A portrays the impact for micro firms, Panel B for small firms, and Panel C for medium-sized firms.

Table 5: Summary of VAT gap and efficiency by size groups

	Years						
	2014	2015	2016	2017	2018	2019	2020
Panel A: Micro							
VAT gap	68.0	50.4	15.3	14.0	21.8	14.2	3.7
Rate gap	6.8	4.2	1.0	1.2	1.5	0.2	0.1
Audited day/declaration	6.2	5.0	4.5	3.4	4.4	4.1	5.7
Evasion/audited day	3,525,024	2,393,376	732,383	1,046,474	866,835	233,369	84,357
VAT gap/efficiency	11.0	10.0	3.4	4.1	5.0	3.5	0.7
Panel B: Small							
VAT gap	56.5	88.0	79.4	64.9	55.6	76.5	76.3
Rate gap	1.4	2.0	2.5	1.9	1.3	1.9	0.5
Audited day/declaration	5.9	5.4	5.1	6.0	4.5	6.7	7.5
Evasion/audited day	425,401	567,756	707,850	358,760	400,370	496,767	159,873
VAT gap/efficiency	9.6	16.3	15.7	10.8	12.3	11.5	10.2
Panel C: Medium							
VAT gap	56.9	47.7	36.8	27.2	26.5	26.8	22.8
Rate gap	54.6	46.8	37.3	28.1	26.9	26.4	23.0
Audited day/declaration	5.9	6.8	7.3	7.4	8.3	8.3	10.0
Evasion/audited day	1,666,298	1,235,699	904,599	645,300	560,119	620,027	402,002
VAT gap/efficiency	9.6	7.0	5.0	3.7	3.2	3.2	2.3

Note: yearly aggregate data.

Source: authors' calculations based on TRA VAT and audit data.

Micro firms show a higher VAT gap, with a stable VAT gap rate over the years. The small and medium-sized firms have a small VAT gap and display different patterns across the years. The VAT gap in small firms increased in 2015, decreased in 2016–18, and increased again in 2019 and 2020. The VAT gap appears to decrease over time for medium-sized firms. The cost-effectiveness ratio is larger in micro and small firms than in medium-sized firms, but the opposite happens with the rate gap. In order to increase revenue, reallocating resources to auditing medium-sized firms could yield the best results.

Table 6 shows the results for audited firms, where Panel A depicts the results for firms that declare a negative VAT, Panel B for those declaring around zero VAT, and Panel C for those with a positive VAT declaration.

Table 6: Summary of VAT gap and efficiency by declaration groups

	Years						
	2014	2015	2016	2017	2018	2019	2020
Panel A: Negative							
VAT gap	25.1	18.5	16.2	11.3	12.3	13.0	8.5
Rate gap	22.5	19.4	13.7	9.6	10.0	9.8	6.5
Audited day/declaration	6.0	6.5	7.5	7.1	8.5	8.4	10.4
Evasion/audited day	1,780,635	1,516,386	969,149	697,894	631,736	701,142	355,046
VAT gap/efficiency	4.2	2.8	2.2	1.6	1.5	1.6	0.8
Panel B: Zero							
VAT gap	100.0	100.0	100.0	100.0	100.0	100.0	99.8
Rate gap	5.4	4.2	2.2	3.0	2.4	2.0	0.7
Audited day/declaration	6.2	6.7	5.7	5.0	4.8	7.1	9.3
Evasion/audited day	1,912,822	1,136,920	707,224	878,965	797,732	581,721	154,387
VAT gap/efficiency	16.2	14.9	17.6	20.0	20.7	14.2	10.7
Panel C: Positive							
VAT gap	25.0	18.2	16.1	11.4	10.9	10.9	10.0
Rate gap	34.9	29.4	24.9	18.6	17.3	16.7	16.4
Audited day/declaration	5.8	6.5	6.9	7.3	7.7	7.9	9.5
Evasion/audited day	1,553,519	1,101,668	861,558	567,291	506,173	560,602	426,963
VAT gap/efficiency	4.3	2.8	2.3	1.6	1.4	1.4	1.1

Note: yearly aggregate data.

Source: authors' calculations based on TRA VAT and audit data.

The largest VAT gap is shown for the group of firms that bunch around zero, with a stable rate over the years. The groups of firms that declare positive or negative VAT show smaller and decreasing VAT gap rates. The cost-effectiveness of auditing is larger for firms that declare around zero VAT, indicating that reallocating resources to auditing more intensively in this sector could heavily reduce the VAT gap. However, the rate gap is larger for the positive VAT declaration group, showing that to increase revenue, it would be better to audit this group more intensively. Therefore, increased auditing of firms that declare around zero VAT increases formality but focusing on the group of firms that declare positive VAT increases revenue. Interestingly, the difference in the rate gap between those two groups is not so large, indicating that focusing only on firms that declare around zero VAT will not recover large revenue losses

The VAT gap results for audited firms and all firms differ. Both results show a similar pattern, with the VAT gap decreasing over time. The difference lies in the groups' levels of the VAT gap.

The difference in levels comes from the cost of auditing. In the efficiency analysis the cost of auditing is shown to increase across the period, and medium-sized firms and firms that declare favourable VAT are the group that drives this fact. The ML technique addresses this pattern as it uses the percentile of the VAT declaration and yearly sales distribution. This means that predictions are made taking account of the changes in auditing cost, thereby bringing to light the underlying value. When auditing costs rise it becomes more costly to discover non-paid VAT, with the result that the amount found after auditing is less than before auditing. Given this, it is possible to estimate a potential unpaid amount in a context where the cost remains constant. This is what drives the difference in the total VAT gap.

6 Conclusions

Looking only at the results from the data, it appears that the VAT gap has been decreasing over time. The data also shows that some firms are being repeatedly audited, learning from this experience and seeking new ways to hide their taxes. This is also seen in the data as the costs to recover the same amount of undeclared VAT have been increasing each year. When the data is used to predict undeclared VAT in unaudited firms and unaudited periods, the VAT gap trend remain the same (decreasing), but the rate is higher when taking account of the increasing costs of discovering undeclared taxes.

Further inspection of the firms shows some interesting facts. First, targeting firms is effective as a small number of audited firms are found to evade large amounts. However, this could also mean that firms declare strategically in order to expose themselves to less auditing. The cost of auditing firms in the negative and positive VAT declaration groups and medium-sized firm groups appears high. These groups may encompass more sophisticated firms or those with more resources to evade. Reallocating resources to increase auditing in a particular sector could be beneficial from different perspectives but also could lead to an exponential increase in cost. For instance, increasing auditing of the group of firms that declare around zero VAT as well as micro and small firms would seem to be effective in increasing formality or decreasing the VAT gap. However, this would not be so effective in terms of revenue. To increase revenue, it would be better to reallocate resources to auditing firms with positive VAT declarations and medium-sized firms. However, auditing such groups would entail large costs, potentially producing an exponential increase. A balance between both strategies and a specification strategy, by targeting a particular group, could improve the VAT gap and increase revenue.

References

Adu-Ababio K., A. Koivisto, E. Lungu, E. Mwale, J. Msoni, and K. Musole (2023). 'Estimating Tax Gaps in Zambia: a Bottom-up Approach Based on Audit Assessments'. WIDER Working Paper 2023/25. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2023/333-8.

Annacondia, F., and W. van der Corput (2012). 'VAT Registration Thresholds in Europe'. *International VAT Monitor*, 20(6). https://doi.org/10.59403/27z6s15

Athey, S., and G.W. Imbens (2019). 'Machine Learning Methods that Economists Should Know About'. *Annual Review of Economics*, 11: 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Barra, P.A., E. Hutton, and P. Prokof'yeva (2023). 'Corporate Income Tax Gap Estimation by Using Bottom-Up Techniques in Selected Countries: Revenue Administration Gap Analysis Program'. *Technical Notes and Manuals*, 2023(006): A001. https://doi.org/10.5089/9798400246265.005.A001

Best, M., J. Shah, and M. Waseem (2021). '*Detection without Deterrence: Long-Run Effects of Tax Audit on Firm Behavior*'. Mimeo: University of Manchester.

Breiman, L. (2001). 'Random Forests'. *Machine Learning*, 45: 5–32. https://doi.org/10.1023/A:1010933404324

Brockmeyer, A., G. Mascagni, V. Nair, M. Waseem, and M. Almunia (2024). 'Does the Value-Added Tax Add Value? Lessons Using Administrative Data from a Diverse Set of Countries'. *Journal of Economic Perspectives*, 38(1): 107–132. https://doi.org/10.1257/jep.38.1.107

Cnossen, J. (2019). *Modernizing VATs in Africa*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198844075.001.0001

Cnossen, S. (1991). 'Key Questions in Considering a Value-Added Tax for Central and Eastern European Countries'. Working Paper No. 1991/069. Washington, DC: International Monetary Fund. https://doi.org/10.5089/97814519606

Crowe Horwath International (2016). *Africa VAT/GST Guide 2016 (A Concise Overview of All 54 VAT/GST Systems in Africa)*. Port Louis: Crowe Horwath International.

Durán-Cabré, J.M., A.E. Moré, M. Mas-Montserrat, and L. Salvadori (2019). 'The Tax Gap As a Public Management Instrument: Application to Wealth Taxes'. *Applied Economic Analysis*, 27(81): 207–25. https://doi.org/10.1108/AEA-09-2019-0028

Ebrill, L., M. Keen, and V. Perry (2001). *The Modern VAT*. Washington, DC: International Monetary Fund. https://doi.org/10.5089/9781589060265.071

Fjeldstad, O.-H. (1995). 'Value Added Taxation in Tanzania?'. CMI Working Paper WP 1995:5. Bergen: Chr. Michelsen Institute.

Fjeldstad, O.-H., C. Kagoma, E. Mdee, I. Sjursen, and V. Somville (2020). 'The Customer is King: Evidence on VAT Compliance in Tanzania'. *World Development*, 128. https://doi.org/10.1016/j.worlddev.2019.104841

Gemmell, N., and J. Hasseldine (2014). 'Taxpayers' Behavioural Responses and Measures of Tax Compliance "Gaps": a Critique and a New Measure'. *Fiscal Studies*, 35(3): 275–96. https://doi.org/10.1111/j.1475-5890.2014.12031.x

Gyoshev, S., C. Kotsogiannis, K. Tester, and T. Pavkov (2023). 'Evasion, Avoidance, and the VAT Threshold'. Mimeo

Keen, M. (2012). 'Taxation and Development—Again'. International Monetary Fund Working Paper, Fiscal Affairs Department.

Kopczuk, W. and J. Slemrod (2006). 'Putting Firms into Optimal Tax Theory'. *American Economic Review*, 96(2): 130–34. https://doi.org/10.1257/000282806777212585

Mrema, N.G.F. (2012). 'The Importance of Audit Trail in VAT Input Credit System: A Case of Tanzania'. Presentation to the 15th Annual Conference of the Value Added Tax Administrators in Africa (VADA), Arusha, Tanzania.

Mullainathan, S., and J. Spiess (2017). 'Machine Learning: an Applied Econometric Approach'. *Journal of Economic Perspectives*, 31(2): 87–106. https://doi.org/10.1257/jep.31.2.87

Pomeranz, D. (2015). 'No Taxation without Information: Deterrence and Self-Enforcement in the Value Added Tax'. *American Economic Review*, 105(8): 2539–69. https://doi.org/10.1257/aer.20130393

Schonlau, M., and R.Y. Zou (2020). 'The Random Forest Algorithm for Statistical Learning'. *The Stata Journal*, 20(1), 3–29. https://doi.org/10.1177/1536867X20909

Slemrod, J., M. Blumenthal, and C. Christian (2001). 'Taxpayer Response to an Increased Probability of Audit: Evidence from a Controlled Experiment in Minnesota'. *Journal of Public Economics*, 79(3): 455–83. https://doi.org/10.1016/S0047-2727(99)00107-3

Sokolovska, O. and D. Sokolovskyi (2015). 'VAT efficiency in the countries worldwide'. MPRA Paper 66422. Munich: University Library of Munich.

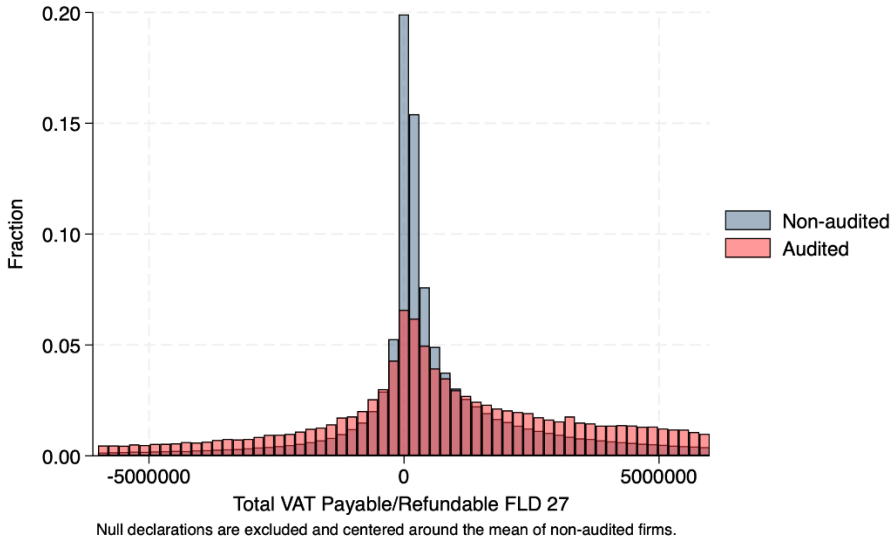
Tanzania Revenue Authority (2019). *Estimation of Tax Gap for Tanzania for the Period 2015/16 - 2017/18 Financial Year*. Unpublished.

Wilks, D.C., J. Cruz, and P. Sousa (2019). 'Please Give Me an Invoice: VAT Evasion and the Portuguese Tax Lottery'. *International Journal of Sociology and Social Policy*, 39(5/6): 412–26. https://doi.org/10.1108/IJSSP-07-2018-0120

World Bank (1991). *Lessons of Tax Reform*. Washington, DC: World Bank.

Appendix

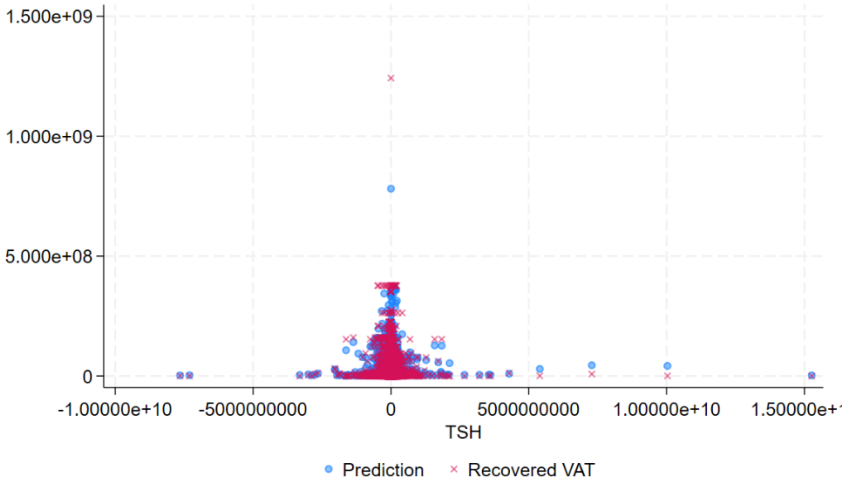
Figure A1: Distribution of VAT payment by audit status



Note: null declarations are excluded and centred around the mean of non-audited firms.

Source: authors' calculations based on TRA VAT and audit data.

Figure A2: Scatterplot non-paid VAT vs prediction



Source: authors' calculations based on TRA VAT and audit data.

Table A1: Summary statistics by audited tax regions

	2014		2015		2016		2017		2018		2019		2020	
Variable	Non-audit	Audit	Non-audit	Audit	Non-audit	Audit	Non-audit	Audit	Non-audit	Audit	Non-audit	Audit	Non-audit	Audit
VAT payable	38	179	94	247	177	310	170	310	350	310	198	397	166	398
Total VAT due or carried forward	-1,006	-2,971	-1,097	-3,506	-1,358	-3,953	-1,865	-4,586	-2,253	-5,486	-2,926	-5,705	-3,445	-5,905
Total output	111	459	7,166	13,371	11,203	13,896	12,445	14,665	14,609	15,721	17,846	22,366	1,798,624	19,350
Total input	4,586	11,635	6,497	11,747	8,564	1,677,081	10,186	11,222	12,416	12,341	13,472	17,570	57,663	15,513
Net profits	-4,186	-10,770	768	1,887	2,617	-1,662,760	2,424	4,127	2,417	4,083	4,581	4,961	1,420,843	4,670
VAT recovered	0	0.66	0.16	6.71	6.22	44.26	20.64	108.60	0.71	31.61	0.30	42.26	0.10	49.75
Total tax recovered	0	2.61	0.61	13.55	15.51	132.30	36.15	338.30	1.47	103.70	3.09	139.40	0.59	194.5
Number of firms	5,778	7,849	6,670	8,594	8,268	9,312	10,021	10,009	11,709	10,846	13,462	11,885	15,315	13,160
Number of audited firms	0	1,507	0	1,671	0	1,824	0	1,949	0	2,052	0	2,120	0	2,106
Number of firms with VAT registration number	56,229	81,734	68,920	90,328	84,218	99,227	104,712	107,538	122,505	116,117	140,305	126,981	161,372	139,546
Number of monthly declarations	10.81	11.27	11.27	11.37	11.19	11.41	11.29	11.48	11.33	11.47	11.33	11.46	11.36	11.41
Number of audited days (VAT)	1,588	20,828	3,448	44,844	4,232	67,024	569	60459	506	70,515	391	78,541	84	105,238

Note: the tax year runs from 1 July to 30 June. Net profit is obtained by subtracting total inputs from total outputs. Monetary amounts are expressed in TZS billions.

Source: authors' calculations based on TRA VAT and audit data.

Table A2: Summary statistics in audited tax regions by audited firms

	2014		2015		2016		2017		2018		2019		2020	
Variable	Non-audit	Audit	Non-audit	Audit	Non-audit	Audit	Non-audit	Audit	Non-audit	Audit	Non-audit	Audit	Non-audit	Audit
VAT payable	86	92	110	136	139	171	136	175	109	201	161	236	150	248
Total VAT due or carried forward	-1,551	-1,419	-1,724	-1,782	-1,813	-2,139	-2,091	-2,495	-2,458	-3,028	-2,364	-3,342	-2,314	-3,592
Total output	180	279	4,813	8,558	4,826	9,070	4,518	10,147	5,009	10,712	9,938	12,428	6,316	13,035
Total input	4,259	7,376	4,070	7,677	1,669,852	7,229	3,390	7,833	3,804	8,538	8,468	9,102	4,743	10,770
Net profits	-3,891	-6,879	863	1,024	-1,664,854	2,094	1,426	2,701	1,465	2,619	1,068	3,892	1,792	2,878
VAT recovered	0	0.66	0	6.71	0	44.26	0	108.60	0	31.61	0	42.26	0	49.75
Total tax recovered	0	2.61	0	13.55	0	132.30	0	338.30	0	103.70	0	139.40	0	194.50
Number of firms	6,342	1,507	6,923	1,671	7,488	1,824	8,060	1,949	8,794	2,052	9,765	2,120	11,054	2,106
Number of audited firms	0	1,507	0	1,671	0	1,824	0	1,949	0	2,052	0	2,120	0	2,106
Number of auditing processes (VAT)	0	310	0	586	0	821	0	735	0	787	0	839	0	904
Number of compliance firms	0	31	0	54	0	66	0	68	0	68	0	86	0	148
Number of firms with VAT registration number	64,928	16,806	71,348	18,980	78,373	20,854	85,080	22,458	92,393	23,724	102,471	24,510	114,882	24,664
Number of monthly declarations	11.19	11.60	11.27	11.72	11.32	11.77	11.39	11.81	11.38	11.81	11.37	11.82	11.31	11.88
Number of audited days (VAT)	0	20,828	0	44,844	0	67,024	0	60,459	0	70,515	0	78,541	0	105,238

Note: the tax year runs from 1 July to 30 June. Net profit is obtained by subtracting total inputs from total outputs. Monetary amounts are expressed in TZS billions.

Source: authors' calculations based on TRA VAT and audit data.

Table A3: Accuracy of ML approach

Statistic	Training Data		Testing Data	
	OLS	ML	OLS	ML
R-square	0.04	0.96	0.04	0.78
RMSE			19,699,902	9,732,259
			3.4	1.7
Mean Variable	5,489,586		5,730,170	
			0.98	0.48
Std Variable	20,044,263		20,122,841	

Note: the ordinary least squares (OLS) estimation used the same variables as the ML.

Source: authors' calculations based on TRA VAT and audit data.