UNITED NATIONS UNIVERSITY
UNU-WIDER

WIDER Working Paper 2023/25

Estimating tax gaps in Zambia

A bottom-up approach based on audit assessments

Kwabena Adu-Ababio,[1] Aliisa Koivisto,[2] Eliya Lungu,[3]
Evaristo Mwale,[3] Jonathan Msoni,[3] and Kangwa Musole[3]

February 2023

United Nations University World Institute for Development Economics Research
wider.unu.edu

Abstract: Assessing tax gaps—the difference between the potential and actual taxes raised—plays a vital role in achieving positive domestic revenue objectives through improved and reformed taxation. This is particularly pertinent for growth outcomes in developing countries. This study uses a bottom-up approach based on micro-level audit information to estimate the extent of tax misreporting in Zambia. Our methods predict the extent of tax evasion using a regression and a machine learning algorithm based on a sample of audited firms, after which we estimate tax gaps using a standard approach. We estimate total tax gaps as 56 per cent and 47 per cent for the two approaches, respectively. These gaps are mainly driven by corporate taxes. Applying our gap to key industries shows that the extractives sector in Zambia records the highest gaps in terms of CIT and one of the lowest gaps in terms of VAT.

Key words: tax gap, VAT gaps, bottom-up approach, audits, tax compliance, tax administration, Zambia

JEL classification: H25, H26, H32

Acknowledgements: The results presented here are based on administrative tax data from the Zambian Revenue Authority (ZRA). We thank the staff of the ZRA and ZIPAR (Laban Simbeye, Mbewe Kalikeka, and Sylvia Mwamba) who assisted in the research coordination and retrieving data on specific tax filings. We are grateful for the support of UNU-WIDER through the DRM workstream as well as support to Kwabena Adu-Ababio by the Finnish Cultural Foundation (grant 00210176). We express special appreciation to Jukka Pirttilä and Miri Stryjan for their useful comments and supervision. The results and their interpretation presented in this paper are solely the authors' responsibility.

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] University of Helsinki and UNU-WIDER, Helsinki, Finland, corresponding author: adu-ababio@wider.unu.edu; [2] VATT Institute for Economic Research, Helsinki; [3] Zambian Revenue Authority, Lusaka, Zambia

This study has been prepared within the UNU-WIDER project Building up efficient and fair tax systems – lessons based on administrative tax data, which is part of the Domestic Revenue Mobilization programme. The programme is financed through specific contributions by the Norwegian Agency for Development Cooperation (Norad).

Information and requests: publications@wider.unu.edu

ISSN 1798-7237 ISBN 978-92-9267-333-8

https://doi.org/10.35188/UNU-WIDER/2023/333-8

Typescript prepared by Joseph Laredo.

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

Public revenue mobilization through taxes is an important tool for economies, as it not only provides the means to secure the provision of public goods but also ensures adequate provision of household social protection. However, revenue mobilization faces many challenges that hamper its efficiency and effectiveness. These challenges include tax evasion and non-compliance. The lost revenue caused by these can be measured as a tax gap, referring to the difference between potential revenue and actual revenue mobilized. The problem of lost revenue compounds (especially in low-state-capacity countries, where informality and poor bookkeeping are rife) and contributes significantly to low tax-to-GDP ratios (Slemrod et al. 2019; Waseem 2018).

Due to the parameter's importance in gauging revenue loss, it is essential in the efficient management of domestic revenue. As a result, various methods have been developed and are constantly evolving to estimate the tax gap accurately. Raising revenue imposes a burden on the taxpayer, who would ideally prefer this to be minimal. Therefore, the tax raised is often less than the total amount due. Researchers and policy-makers in tax research and policy have the singular goal of understanding the magnitude and form of such revenue losses and preventing their frequent occurrences.

In this study, using newly available audit data, we employ one of the known methods, the bottom-up approach, to estimate the total tax gap in Zambia. With this method, we aim to tackle some of the endogeneity problems a macro-based top-down analysis might suffer from due to the nature of existing data in Zambia. Currently, there exist no other data to calculate potential tax revenue than the GDP statistics derived from administrative tax data. As the top-down approach stipulates that alternative data should be used to calculate potential tax revenue, it is almost impossible to use this method. We find a solution to this in a bottom-up analysis, which is a substitute micro approach equally used to investigate tax gaps precipitated by evasion and informality. Our approach is to use the information on audit outcomes to estimate the prevalence of tax evasion among taxpayers.

Our chosen approach has the added advantage of predicting the industries and firms, and even the localities, where evasion is the greatest. The scale of tax evasion can be deduced from the audit outcomes if the audits are made randomly. However, in our case where audits are not random, we incorporate auditing probabilities that enable the extrapolation of an estimate of the tax evasion among the full population of firms. The study controls for the audit probabilities by two means: a simple regression predictor and a complex machine learning (ML) predictor based on neural networks.

Responsibility for administering policies that ensure a stable form of fiscal financing lies with the local tax authority. One key measure to enforce taxpayer compliance is the use of effective models to select taxpayers for tax audits. This study uses rich administrative data provided by the Zambian Revenue Authority (ZRA) on all audits, as well as corporate income tax (CIT) and value added tax (VAT) returns information,[1] as the appropriate benchmark for gap estimates for the years 2014 to 2020. We observe information about which firms were selected into tax audits or examinations as compared with firms not selected for the routine check. By concentrating on firms that filed tax returns within this period, we use the audit outcome of the assessed firms to predict the evasion

[1] The audit information covers tax assessments for PAYE, which are included in the CIT returns.

amount for the unaudited firms. These predictions form the basis of our tax gap calculations for the full population of firms.

We begin with a descriptive analysis of how audited firms differ from unaudited firms in terms of location, sector, and return years. We subsequently explain the reasons behind the methods we employ to predict total tax assessments as well as the difficulties encountered therein. We report the tax gap based on VAT and CIT filings using imputed total assessment for all firms while relying on information on the extent of evasion of firms selected into audits by the ZRA. On the basis of our estimates we measure the extent of the tax gap across return years to ascertain the rise or fall in potential taxes recouped as against the actual taxes realized in the tax return register. Finally, we discuss the potential tax revenue estimates in comparison with other factors, such as gross domestic output, in Zambia as well as other tax revenue estimates based on either similar or diverging methodologies and previous study results. The aim of this is to validate the results of our study with a series of robustness checks.

Our results, based on actual and imputed evasion rates, show that tax gaps range between 56 per cent and 47 per cent depending on the prediction method (regression or ML approach) employed. Subsequently, we show that there is a substantive difference between reported and potential tax return filing over the studied years. Our estimation methods show that Zambia could recoup a total of K16.8 billion (~US$940 million)[2] more revenue as compared with the actual revenues received from VAT and CIT filings. We also observe that, as most firms report VAT in refund position, the gaps we observe are mainly driven by CIT filings. Most importantly, the extractives sector in Zambia records the highest gaps in terms of CIT and one of the lowest gaps in terms of VAT. The retail and wholesale industry records the highest VAT gaps. We infer the magnitude of the gap within each industry/sector and suggest alternative auditing approaches that may improve compliance, with the aim of reducing gaps.

Our study is among the first to use a bottom-up approach with audit data to estimate tax gaps in a developing country. Exceptionally, a recent working paper by Best et al. (2021) uses audit information from random but non-comprehensive audits to estimate the VAT gap for Pakistan. The authors consider their evasion estimate of 7.5 per cent as a lower bound due to the non-comprehensive nature of the studied audits, while for small firms their estimate is 80 per cent. Although some revenue authorities in developed countries have used these methods to estimate tax gaps for policy use (e.g. HM Revenue and Customs 2022 for the UK), the bottom-up approach has been less common in academic research due to limited access to data and even more rare in a developing country context. However, other methods have been used to estimate tax gaps in developing countries. Notably, Danquah and Osei-Assibey (2018) use survey information to estimate the tax gap for the non-farm informal sector in Ghana. Their estimated evasion rate is approximately 70 per cent of the potential revenue. There are some other studies that estimate tax gaps in developing countries, yet they usually employ a top-down approach (e.g. Hutton 2017; Jansen et al. 2020; and Ramírez-Álvarez and Carrillo Maldonado 2020 for Ecuador). It can be inferred that such top-down studies are rare in sub-Saharan Africa (SSA), with a few exceptions such as Alexeev and Chibuye (2016) in Zambia, Mascagni et al. (2019) in Rwanda, and Lakuma and Sserunjogi (2018) (a VAT gap analysis) in Uganda.

Alexeev and Chibuye's (2016) ground-breaking study of Zambia provides an overview of the extent of tax compliance in the country. The authors estimate a 30 per cent and 50 per cent VAT gap depending on the study year. In this study we make no distinction between GDP value added

[2] This value records the difference between potential and actual total tax returns during the study period using predictions from a regression model. We record a difference of US$550 million using the ML predictions.

when calculating the potential tax base as well as the actual tax base; thus it is valuable in providing confirmation of the level of evasion with alternative methods. It is nevertheless evident that providing more bottom-up tax gap estimates in a developing country context is an important addition to the literature; indeed we provide a unique example for tax authorities for which the methods based on top-down approaches cannot be replicated (Zídková 2014).

Moreover, we contribute to the literature on firm-level tax compliance. Generally, if tax gaps are large then compliance is poor and vice versa. This points to the importance of key measures implemented to improve compliance. This study is complemented with audit information, which is such a key measure. By employing audit information to measure the magnitude of tax gaps using full population data, we complement other scholarly works using audit information, but with two novelties. On the one hand, we rely on firms' audit status, where we find temporary changes in tax assessments after an audit. This means that our comprehensive tax compliance analysis is based on multiple firm characteristics and uses prediction models to show how actual tax returns deviate from potential tax returns. On the other hand, we apply a machine learning (ML) algorithm, which improves the tax assessment predictions of the sample of unaudited firms. In this sense, our study is related to that of Battaglini et al. (2022), who employ ML techniques to improve audit efficiency using administrative data for self-employed firms in Italy.

Last, our study relates to the broader literature examining the effect of enforcement mechanisms on tax compliance. Almunia and Lopez-Rodriguez (2018), for example, show that monitoring intensity affects tax compliance in terms of whether a firm is audited or not (audit status) in addition to firm characteristics. Our study also relies on an examination of audit status, with the novelty of measuring tax compliance via deviations obtained from tax declarations and subsequent audit outcomes.

The paper is structured as follows. In Section 2, we discuss the institutional setting and the processes that the ZRA follows before firms are selected for audit as well as the types of audits the authority undertakes. Section 3 explains the methods we use in estimating the tax gaps, as well as the principles behind our estimation techniques, how the study implements these estimation methods, and the data employed while implementing the estimation. Section 4 discusses the summary statistics of tax return variables by audit status before examining the results of the tax gap estimations and validating the tax gap estimate. Conclusions and policy recommendations follow in Section 5.

2 Institutional context

This study focuses on tax gaps in CIT and VAT. In Zambia firms are usually liable to VAT at rates of 4 per cent or 16 per cent depending on the category of taxable supplies. Firms with sales above K800,000 in any 12 consecutive months or K200,000 (~US$12,000) in any consecutive three months must register for VAT unless they deal in exempt supplies.

Firms are normally liable for CIT in one of two ways: if their sales exceed K800,000 (~US$50,000) or if they are part of a specified business activity, as prescribed by law, regardless of their sales threshold. The current standard CIT rate as applied to profits stands at 30 per cent.[3] However, not all firms are liable for CIT at this rate, as the rate depends on certain thresholds and industries. Certain industries, such as telecommunications, are liable to the standard 30 per cent rate only on

[3] Until the end of 2021 the rate was 35 per cent.

annual profits of up to K250,000; on profits above this amount, the top marginal tax rate of 40 per cent kicks in. For industries engaging in farming, agro-processing, and non-traditional exports, the applicable CIT rate is 10 per cent. For firms dealing in other kinds of exports but engaging in the manufacture of organic and chemical fertilizers, the applicable rate is 15 per cent. Below the threshold or the industry exclusion criteria, firms are liable for turnover tax. The goal of these CIT rates is to be equitable enough to encourage business growth and prevent underreporting of profits. We summarize the characteristics of the taxes in the Appendix.

Relying on the level of assessments gathered from audits, we briefly discuss the audit processes and how the tax authority selects firms into audits. Before the ZRA conducts an audit or selects firms into audits, there are specific processes and procedures that take place. We first outline these processes and second discuss the type of audits that the revenue authority carries out. The ZRA employs a self-assessment mechanism in which taxpayers are expected to make declarations and payments as prescribed by applicable tax laws. To ensure that taxpayers declare the correct tax at the right time, the ZRA has an audit programme to verify taxpayers' declarations. The authority employs risk-based approaches to select and target cases for audit to ensure efficient use of resources. Case selection is the process of identifying taxpayers to be audited during a specified period. The selection criteria must meet the ZRA's objectives and are structured to ensure that resources are directed towards areas of greatest risk and revenue benefit for the return year in question. The selection is carried out by the Audit and Business Support Unit of Design, Monitoring, and International Relations (DMIR).

The process involves assigning defined risk scores to taxpayers and selecting for audit those that fail the predefined parameters set according to a year's audit strategy, then allocating selected cases to audit centres based on their audit capacity. The setting of risk parameters involves an assessment of unusual items observed during analysis of the taxpayer's records and needing verification. These include inconsistent filing, perceived unusual methods of doing business, consistently low net tax payable, alternation between payment and refund positions, variance between sales and tax thresholds, variance between purchases and sales, and variance between declarations and available third-party data.

The compilation of items needing verification is done at province level by the audit centre within the designated tax office. The centres are not independent contracted entities but form part of the ZRA's organizational framework and are set up locally to monitor returns within their jurisdiction. Once the audit centres have aggregated the flagged cases, they can recommend additions and removals from the pool of compiled cases, giving reasons and subject to approval by the DMIR. The Station Manager of an audit centre (also known as a station) can also recommend a case to be added; this is then approved by DMIR Audit and Business Support via the ZRA's online platform TaxOnline.

The ZRA undertakes two types of audit: a comprehensive audit and an issue audit. A comprehensive audit is detailed and concentrates on broad categories of risk. The audit looks at the affairs of the taxpayer, covering a period of at least one charge year, in totality and may focus on one tax type (single tax type audit) or be integrated to cut across tax types (integrated audit). For an audit to be classified as integrated, a minimum of two tax types must be looked at.

Any audit covering less than one charge year is an issue audit, which concentrates on a specific issue, area, or item in a specific period, which may be one return period, i.e. one month for VAT or one year for CIT, depending on the case. The issues covered by issue audits are summarized in Table 1. If an issue audit covers a period of 12 months for VAT, then it becomes a comprehensive audit. Issue audits are initiated through credibility parameters and can be added to comprehensive audits as checks on specific returns that have failed some predetermined parameters. For example,

refund audits may be generated by questionable refund claims, while a deregistration audit might be generated by a deregistration request. We summarize the different types of audits and the reasons why such an audit may be triggered in Table 1.

Table 1: Audit types implemented by the ZRA

Audit type	Tax type	Example of a trigger
Comprehensive	Income tax	Sales underdeclaration and overclaimed capital allowances
Comprehensive	Income tax, PAYE, and WHT	Variances between income tax expenses, PAYE, and WHT
Comprehensive	VAT	Trend analysis of declarations
Issue	Property transfer tax	Sale of property (shares, mining rights, intellectual property, land)
Issue	Income tax	Overstated purchases
Issue	WHT	Unreported WHT
Issue	PAYE	Variances in gross emoluments from third-party data
Issue (credibility)	VAT	Refund verification; invoice underdeclaration

Note: WHT = withholding tax

Source: authors' compilation.

3 Methodology

In this section, we discuss the conceptual framework of estimating tax gaps using a bottom-up approach. We first discuss the concept of using audits as a basis for calculating tax gaps and its associated problems and how they can be addressed. This is followed by a discussion of how we employ these notions in our work.

3.1 Conceptual framework

In trying to estimate the gap in tax revenue on the basis of audit information, we need to find a methodology that will estimate the extent of evasion among unaudited firms with minimal bias or selection error. This is necessary as audits carried out by the ZRA are not random but targeted (risk-based). The estimated evasion rates are then applied to calculate the overall tax gap. The overall tax gap, as recently calculated in ratio terms by Best et al. (2021) is:

$$Tax\ Gap = \frac{\sum_{pop} estimated\ tax\ evasion}{\sum_{pop} potential\ tax\ income} = \frac{\sum_{pop} estimated\ tax\ evasion}{\sum_{pop} returned\ income + \sum_{pop} estimated\ tax\ evasion} \quad (1)$$

where the estimated tax evasion is the total volume of assessments after an audit. Given the fact that the Zambian tax authorities select a limited number of firms to audit on the basis of their risk criteria, it is likely that the extent of evasion differs among the audited firms in comparison with the full population of firms. It will therefore be erroneous to generalize that the average tax gaps would equal the level of evasion detected in audits. Evasion detected in audits is likely biased upwards if the selection into audits has been successful in targeting particularly risky firms. We present two alternative methodologies, a regression approach and a machine learning algorithm, to estimate the evaded tax for the full population of firms.

The study first uses the information on audited firms and a regression approach to create a prediction model of the extent to which one firm may be evading. Based on these probabilities from the audited firms, the study makes an out-of-sample prediction of the evasion of the

unaudited firms. Using these predicted evasion rates, we can calculate aggregates for an economy-wide gap based on the total sample, including audited and unaudited firms.

Although the above method sets out some factors (explanatory variables) potentially predicting the audit outcome of the firms selected into audits, it may not be exhaustive in terms of capturing inherent interactions and network paths that may affect prediction results. To address the sample selection issue in an alternative way, we apply machine learning (ML) as a second approach. The ML algorithm provides a strategy to predict the assessment of unaudited firms by learning from the characteristics of audited firms. By screening returns based on a priori audit data, we make the assumption that similar tax declarations will be based on similar firm characteristics. We therefore use an ML algorithm to estimate a model for the audit outcomes based on the tax record information of the audited firms. We subsequently use coefficients of the algorithm to predict the audit outcomes of the unaudited firms. The benefit of using ML is that the variables, their interactions, and the functional form captured are chosen in a data-driven manner, which minimizes the risk of making limited or restrictive modelling assumptions. The predictions are generated using the audited population as the training sample, and the predictive performance is assessed using the unaudited population as a testing sample; both samples are drawn from the total population of firms that submitted their tax return in our years of interest. As a final step, to obtain the total tax gap we apply the actual and predicted total tax assessments of all firms, as shown in Equation 1.

3.2 Empirical strategy

Regression approach

To employ the regression approach, we set up a model to predict the audit outcome using parameters from the tax return. We present our model as:

$$y_{it} = \beta_n x_i' + \varepsilon_{it} \tag{2}$$

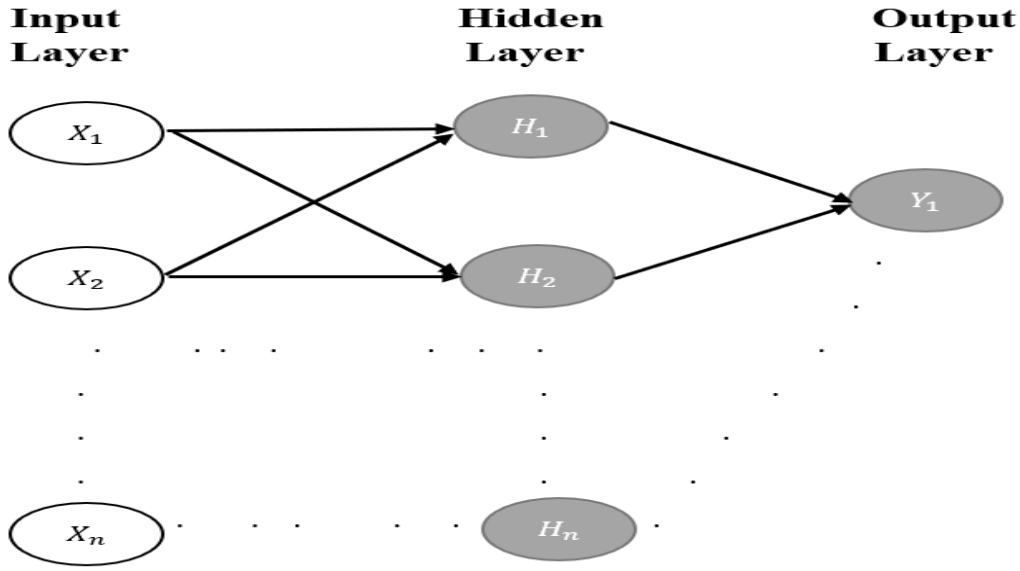
where y_{it} is the level of tax assessment (i.e. tracked evasion) of a firm i after an audit in a particular return year t and x_i' is a vector of explanatory variables based on the tax return information of the firm that potentially affect the audit outcome. The level of assessment y of a firm i in a return period t is the evaded amount observed for audited firms after an audit and can be influenced by the type of audit (comprehensive or issue). However, audit data is very sparse, and splitting tax assessments into types of audits reduces the sample to predict on. As data limitation makes it difficult to check audit type implications, the study pools all tax assessments irrespective of the audit type. The coefficients for each explanatory variable β_n inform on the extent to which the parameter is correlated with detected evasion, while ε_{it} is the error term. The regression framework can also be made relatively non-parametric (i.e. to a specification where the right-hand side variables are not entered linearly). We use Equation 1 in the form of an OLS regression by controlling for the industry, location, and size group effects, among other factors. This provides predictions based on the level of known assessments for the audited sample.

The machine learning approach

In this study, the ML approach employed is based on an Artificial Neural Network (ANN) learning mechanism, which, although recent, is gaining much popularity among scholars (Mullainathan and Spiess 2017). The strength of the method lies in the parallel aggregation of information based on a set of neural networks as well as the ability to make final predictions based on a linear combination of variables ('neurons'), which creates layered hierarchies of linear or binary regressions (Al-Mobayed et al. 2020; Boutaba et al. 2018; Ogwueleka et al. 2015). We improve

predictions based on the complexity of interacting variables. The ML approach controls these interactions using the number of neurons per layer and their subsequent connectivity. Figure 1 presents a stylized model to illustrate how this ML approach is used for our predicted estimates.

Figure 1: Neural network with N inputs and neurons



Source: authors' illustration.

In Figure 1 there are X_n inputs, which in this study are the explanatory variables chosen as factors that potentially affect the possibility of selection into audits, and H_n hidden layers located between the input (first) layer and output (last) layer. These are the levels of interactions between the main input variables, which are set arbitrarily depending on the level of complexity assumed for the audit selection model. The output layer represents the dependent variable of interest, which in our case is the tax assessments after the audit. This is a simple feedforward network, as the inputs for the Y_1 are the outputs from H_1 to H_n.

We partition our data into two and implement the above model on the training data, which includes only audited firms, where the model learns from the selected imputed *X*s. The study limits the learning algorithm to 1,000 repetitions to select the best predictors with minimal error. These predictions are applied to the testing data, which excludes the audited sample to obtain predicted total assessments for unaudited firms. The total assessments are then used as a base for aggregate tax gap calculations.

3.3 Data and predictions

There are three different datasets in our analysis: monthly VAT returns, annual CIT returns, and audit assessments on all Zambian firms for the years 2014–21. VAT and CIT return data include information on monthly and annual declarations, respectively, as well as firm characteristics. The source of all datasets is the ZRA, which performs risk-based audits on selected firms. We are obliged to employ the VAT and CIT datasets together, as audit information is silent on whether an audit was triggered by a VAT or CIT discrepancy. As one solution to not knowing the actual share of the VAT in detected overall firm-level evasion, we may assume that the share of VAT in the evaded amount corresponds to the share of VAT in the firm's total tax payment. However, the study can match each taxpayer to its respective VAT or CIT declaration and the subsequent outcome of the audit. With the combined data, there is the opportunity to split the calculated tax gaps attributable to VAT and CIT based on audit status.

The data comprise 133,516 annualized VAT returns, CIT returns, and audit assessments and identify 45,085 unique firms over the seven-year period. There are 8,472 firms that have undergone audits at least once over the study period, the highest number of audits being in 2018 (on 1,857 firms) and the lowest in 2020 (on 37 firms).[4] The remaining 36,613 firms have filed CIT and VAT returns but have never undergone audits. With these data we can capture both CIT and VAT gaps using knowledge of audit status and tax declarations or assessments by applying the proposed regression approach and ML algorithm. Moreover, using the proposed models permits the calculation of the estimated evasion comprising the tax assessment for audited and unaudited firms, which we then use as a base to estimate tax gaps.

If all firms report truthfully irrespective of audit status, reported tax declarations should not show major deviations or gaps across firms based on the size of tax assessments after audits. However, if firms are in the habit of misreporting in their CIT or VAT filings according to the size of their tax assessments, then we expect to find significant gaps due to misreporting. The resulting gap due to the audit status will inform us, at least to some extent, as to the declaration that the tax authority is likely to receive.

Our tax gap estimate is a lower bound of the tax gap for two reasons related to the data. First, as this study approach is based on audits, we rely on the predefined 'gap' within the data as the difference between the audit outcome and the tax declared of audited firms. This gap may not always be positive but rather zero or negative depending on the results from the audits. Hence, if the detected evasion is smaller than the actual evasion, our tax gap results are based on lower-bound estimates of evasion. Second, our tax data are based on registered firms, so the tax gap prediction leaves out the fully informal firms in the economy.

3.4 Data description

We first discuss the distribution of audits and the distribution of specific variables across the audited and unaudited samples and ask why certain years, provinces, or sectors have a higher or lower potential for selection into audits. We also discuss the variables or risk factors used with the audited sample to predict the outcome tax assessment for firms selected into audit. Table 2 presents a summary of the audit status of the sample of firms.

Table 2: Audit status across return year

Return year	Unaudited	Audited	Total
2014	18,543	921	19,464
2015	20,832	695	21,527
2016	22,596	597	23,193
2017	19,947	1,020	20,967
2018	17,837	1,857	19,694
2019	13,567	679	14,246
2020	14,388	37	14,425
Total	127,710	5,806	133,516

Source: authors' calculations based on ZRA administrative data.

The data include 133,516 firm-year observations in total, 127,710 of them unaudited and 5,806 in the audited group over the period 2014–20. The sharp increase in the number of audits from 2017 to 2018 is attributable to the Tax Amnesty programme, an amnesty on interest and penalties

[4] This is in part due to the delayed launch of the audit module from the filing software TaxOnline II, as well as to reasons attributable to the COVID-19 pandemic.

introduced by the ZRA in 2017 and extended to 2018. During this period, taxpayers were expected to submit outstanding tax returns and pay all principal tax liabilities for tax periods prior to 1 March 2017, after which all interest and penalties accrued for the said period would be waived in full. This resulted in an increase in the number of audits for both CIT and VAT. In addition, the ZRA prioritized tax audits in 2018. A notable decline in the number of audits was observed in 2020 due to a transition in tax administration systems; the ZRA was moving to a new tax administration system (TaxOnline 2) from TaxOnline 1. As a result, most audit functions could not be performed by the new system in this period.

Table 3 presents a summary of the audit status of firms by location for both the audited and unaudited groups. Due to data quality issues on some tax returns and audit data, allocation by location was not always possible. The number of firms was thus reduced to 51,860.

Table 3: Audit status by location

Location	Unaudited	%	Audited	%	Total
Central Province	1,527	3.15	97	2.91	1,624
Copperbelt Province	12,826	26.46	1,095	32.89	13,921
Eastern Province	722	1.49	98	2.94	820
Luapula Province	124	0.26	20	0.60	144
Lusaka Province	30,181	62.19	1,736	52.15	31,917
Muchinga Province	66	0.14	8	0.24	74
North Western Province	626	1.29	111	3.33	737
Northern Province	127	0.26	18	0.54	145
Southern Province	2,025	4.17	132	3.97	2,157
Western Province	307	0.63	14	0.42	321
Total	48,531	100	3,329	100	51,860

Note: the total is the number of taxpayers who report their location in the administrative data.

Source: authors' calculations based on ZRA administrative data.

Lusaka and Copperbelt Provinces account for the largest proportion of both groups. This is because most of the businesses registered for VAT and CIT are concentrated in these regions (ZRA statistics bulletin). In addition, over 70 per cent of business registrations are in Lusaka and the Copperbelt, with Lusaka accounting for over half (2021 PACRA Annual Report).

Table 4 displays the sector-wise breakdown of the audited and unaudited groups that filed returns during the period. Wholesale and retail trade/repairs and manufacturing account for approximately 50 per cent of all audits conducted during the period. This is largely due to the nature of their businesses, which require VAT registration, as well as the large daily volume of VAT transactions they conduct.

Construction, and mining and quarrying have the next highest number of audits due to the complexity and specialization of their business activities, while the agriculture, forestry, and fishing sector is audited primarily to ensure that there is no abuse or misclassification of inputs that are supposed to be standard-rated, as the sector allows certain exempt inputs. The above-mentioned activities, together with transportation and storage; professional, scientific, and technical activities; other service activities; administrative and support services; and accommodation and food service activities account for 95 per cent of all audits during the period.

Table 4: Audit status by sector

Sector	Unaudited	%	Audited	%	Total
Accommodation and food service activities	3,210	2.52	78	2.34	3,288
Activities of extraterritorial organisations	147	0.12	-	-	147
Activities of households	19	0.01	-		19
Administrative and support services	5,994	4.71	139	4.17	6,133
Agriculture, forestry, and fishing	7,082	5.56	216	6.59	7,298
Arts, entertainment, and recreation	643	0.50	20	0.60	663
Construction	9,450	7.42	256	7.69	9,706
Education	1,882	1.48	1	0.03	1,883
Electricity, gas, steam, and aircon	460	0.36	23	0.69	483
Financial and insurance activities	1,692	1.33	25	0.75	1,717
Human health and social work	1,181	0.93	9	0.27	1,190
Information and communication	1,514	1.19	37	1.11	1,551
Manufacturing	7,268	5.71	468	14.05	7,736
Mining and quarrying	3,602	2.83	224	6.73	3,826
Other service activities	13,501	10.60	182	5.47	13,683
Professional, scientific, and tech. activities	7,021	5.51	196	5.89	7,217
Public administration and defence; social security	87	0.07	3	0.09	90
Real estate activities	3,256	2.56	47	1.14	3,303
Transportation and storage	4,934	3.87	211	6.34	5,145
Water supply, sewerage, and waste	320	0.25	16	0.48	336
Wholesale and retail trade, repairs	54,075	42.47	1,179	35.41	55,254
Total	127,338	100	3,330	100	130,668

Source: authors' calculations based on ZRA administrative data.

Table 5 shows the summary statistics for the audited and unaudited firms. The number of audited observations was 8,472, with 8,399 positive assessments. The total number of unaudited observations was 127,668 firm–year pairs. Table 5: Statistics for selected variables by audit status (K)

	Audited		Unaudited	
	Mean	SD	Mean	SD
Total sales	108,051,338	814,061,700	22,205,476	271,394,744
Total purchases	56,397,417	407,955,050	12,538,693	133,055,770
Total VAT	-6,910,931	81,358,223	241,741	17,983,587
Total output VAT	7,192,081	47,236,431	2,255,640	22,479,639
Salaries/Wages	7,072,042	62,917,877	1,141,616	16,031,725
Gross profits	46,675,437	517,791,088	3,428,643	123,826,130
Total assessment	273,853	6,842,049	-	-
Corporate income tax	2,329,210	31,317,879	186,435	10,232,159
Observations	8,399		127,668	

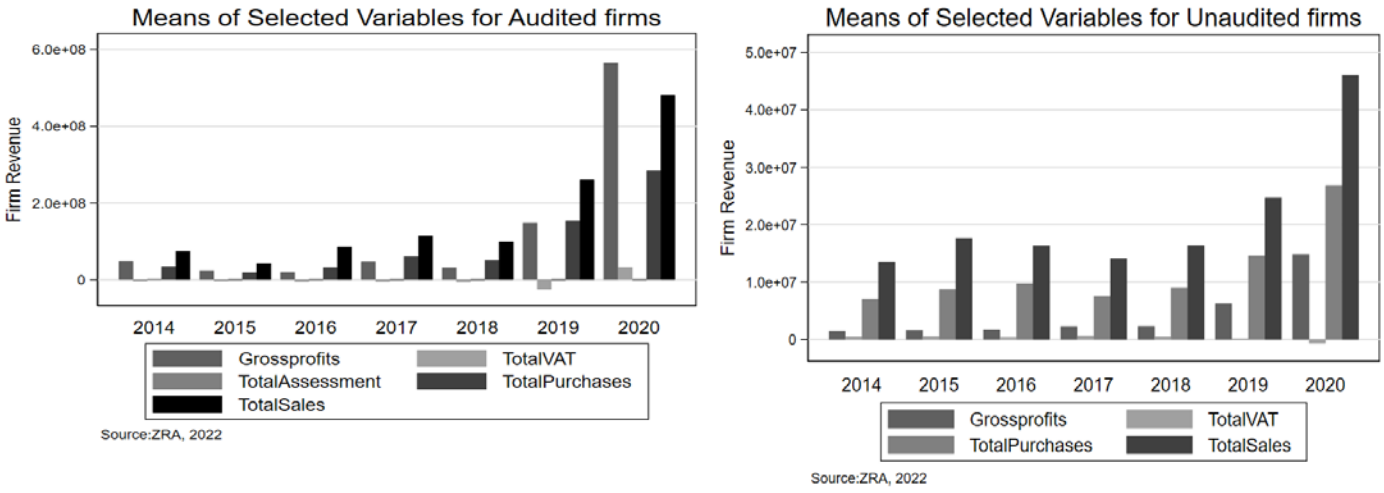
Source: authors' calculations based on ZRA administrative data.

For the audited population, the mean total sales in the period 2014–20 were K108.1 million. Half of the population recorded total sales above K4.1 million (median), the other half falling below this amount. The mean total sales of the unaudited firms in the period was K22.2 million with a median of K1.4 million and a standard deviation of K271.4 million. The amount of CIT was significantly higher for audited firms than unaudited firms, as the former tended to be larger. Tax authorities are generally interested in scrutinizing firms with higher tax revenue potential and this is no different in Zambia, where audited firms on average pay about K2 million in income tax compared with less than K200,000 for unaudited firms. While the audited firms are generally larger than unaudited firms, there is high variation in both groups, which means that the groups overlap in size, enabling us to use the audited group to predict evasion amounts for the unaudited.

For the audited businesses, 50 per cent had gross profits above K1 million while the other 50 per cent were below the same amount in the period 2014–20. The mean gross profits were K46.7 million with a standard deviation of K517.8 million, compared with K3.4 million, with wide profit variations among businesses as reflected in a standard deviation of K123.8 million, for the unaudited businesses. In terms of total purchases, the mean of the unaudited businesses stood at K12.5 million, the median was K0.7 million, and the standard deviation was also high at K133.1 million. Salaries/wages among the unaudited population recorded a mean of K1.1 million with a median of K0.02 million and a standard deviation of K16.0 million.

Figure 2 plots the evolution of various key variables across years and by audit status. In the period 2014–20, fluctuations were observed in the mean for total purchases and the mean gross profits. For the audited businesses, whose population stood at 5,743, the highest mean for total purchases was recorded in 2020 at K285.0 million with a standard deviation of K634.1 million, and the lowest was recorded in 2015 at K18.8 million with a standard deviation of K788.0 million. Consequently, the highest mean for gross profits was also recorded in 2020 at K564.4 million with a standard deviation of K975.8 million, and the lowest mean for gross profit was recorded in 2015 at K238.0 million with a standard deviation of K333.1 million.

Figure 2: Statistics for selected variables by audit status by year



Source: authors' calculations based on ZRA administrative data.

Among the unaudited businesses (127,668), fluctuations were also observed in the mean for total purchases. The highest mean for total purchases was recorded in 2020 at K26.8 million with a standard deviation of K276.6 million, and the lowest mean for total purchases was recorded in 2014 at K7.0 million with a standard deviation of K48.4 million. However, the mean for gross profits rose steadily throughout this period. The lowest mean gross profit was recorded in 2014 at K14.0 million with a standard deviation of K18.8 million, and the highest mean gross profit was recorded in 2020 at K148 million with a standard deviation of K276.6 million. This indicates a 946.5 per cent increase in the mean for gross profits (see Figure 2).

In Table 6, we show the average VAT paid for both the audited and unaudited firms in the different economic sectors. The average VAT paid by the audited firms by sector indicates a negative contribution to the overall VAT payments. This implies that, on average, after being audited, firms were found to be in a refund position. The major driver for this position is the firms in the mining and quarrying sector. This is due to the nature of business and the huge trade volumes in this sector. The mineral outputs from the mining firms are exported. The export of these goods from Zambia is zero-rated provided firms show requisite evidence of exportation. This automatically

puts the mining firms in a refund position, hence explaining the high negative amounts in VAT payments.

Table 6: Summary of total VAT paid by sector

Sector	Audited		Unaudited		Total	
	Mean	SD	Mean	SD	Mean	SD
Accommodation and food service activities	684,036	2,702,828	345,772	2,258,885	361,310	2,281,341
Activities of extraterritorial organisations	-	-	2,057,267	3,037,541	2,057,267	3,037,541
Activities of households	-	-	48,406	50,400	48,406	50,400
Administrative and support service	-3,291,356	37,800,000	2,563,408	42,600,000	2,234,595	42,400,000
Agriculture, forestry, and fishing	172,138	2,925,185	116,557	1,894,050	121,470	2,005,875
Arts, entertainment, and recreation	234,770	942,697	465,664	854,101	452,836	859,469
Construction	-522,061	10,300,000	207,231	2,378,438	162,047	3,450,534
Education	5,220	-	172,724	884,072	170,461	878,212
Electricity, gas, steam, and aircon	-4,090,086	21,500,000	6,034,258	53,700,000	5,055,855	51,600,000
Financial and insurance activities	3,855,176	8,059,102	2,651,245	8,245,357	2,715,421	8,231,538
Human health and social work activities	-813,734	2,366,358	-416,277	5,089,652	-441,292	4,958,586
Information and communication	4,489,074	26,000,000	2,671,481	13,900,000	2,771,706	14,800,000
Manufacturing	-2,270,977	14,900,000	508,011	7,885,202	247,377	8,820,837
Mining and quarrying	-94,400,000	298,000,000	-8,046,715	91,600,000	-22,400,000	151,000,000
Other service activities	-26,137	6,129,873	422,917	4,207,029	399,817	4,326,785
Professional, scientific and tech activities	762,147	6,828,479	507,112	2,774,482	522,738	3,173,781
Public admininstration and defence; soc. sec.	-28,100,000	41,500,000	452,292	6,191,336	-1,799,350	13,700,000
Real estate activities	-1,317,793	3,691,328	212,394	1,896,991	129,347	2,061,062
Transportation and storage	-643,922	3,155,581	126,970	1,700,470	59,617	1,884,809
Water supply, sewerage, and waste	856,127	2,577,089	69,920	661,191	143,055	1,015,531
Wholesale and retail trade, and repairs	-290,847	5,806,004	146,154	5,944,548	123,353	5,938,069
Total	-6,910,931	81,400,000	241,672	18,000,000	-217,341	27,000,000

Source: authors' calculations based on ZRA administrative data.

For the unaudited firms, we find that their average contribution to VAT paid is positive. The major drivers are the energy; financial and insurance activities; and information and communication sectors. The total VAT reported by audited firms was found to be negative.

4 Results

In this section, we discuss the main results of the study. We present results for tax gaps based on our regression and ML models.

4.1 Estimation of tax gap with regression approach

The initial method employed is a regression based on the full firm population data on operational tax audits over the return period 2014–20. The regression predicts the evasion amount for unaudited firms using the audited firms' data. We note from the descriptive summaries that the ZRA conducted an audit for 18.8 per cent of the total population of unique firms at some point in the seven-year period. Although the data provide the assessment for the audited sample, the study predicts the assessments for the unaudited cohort using a regression approach.

The probability of selection into an audit is obtained on the basis of the risk parameters specified by the ZRA. The study limits itself to parameters that have a suitable representation within CIT and VAT filings when combined with audited filings. In our prediction model, we employ parameters such as gross profits, tax on taxable profit, cost of sales, tax on taxable profit, and gross salaries. It is important to note that in our prediction equation the VAT-related variables are excluded due to data inadequacy. Ideally, our prediction equation can be improved in many ways: First, it should include control variables that take into account local exempt sales and export of exempt goods and services, such that tax assessments are not predicted for firms reporting zero-rated exports and exempt sales. Second, VAT on import services is levied as reverse VAT and could be used in the prediction equation by including such controls. However, we encounter challenges when we try to improve our prediction equation. When we include such controls, we lose 63 per cent of the prediction power for unaudited firms, as most firms have these variables missing. With a reduced predicted sample there is the possibility of a selection leading to biased total gaps. Prediction is only possible if the variables used are not missing for the unaudited firms. Therefore, the study uses feasible non-missing variables as considered by the tax authority when selecting firms into audits. Moreover, in the prediction equation we include controls specifying the sector or industry firms select into. We believe that such controls potentially take account of firms reporting zero-rated exports or exempt sales as well as import VAT.

About 2,330 audits can be tracked from existing CIT and VAT filings. In this manner the study can predict the total assessment outcome of similar firms with unaudited status. We show results of the regression prediction model in Table 7, which gives the coefficients for both alternative models. Column (1) restricts the risk parameter variables to levels without any transformation, while column (2) uses an alternative to log transformations on the risk parameters. This avoids the bias that might affect our results by adjusting and including zero or below zero variable cells. An inverse hyperbolic sine transformation is an alternative to log transformations where some of the variables used as prediction parameters—such as gross profits taxable profit, and tax on taxable profit—may take on zero or negative values. In column (3), we further restrict the tax audit outcomes by assuming that the previous year's audit outcomes influence selection into a scrutiny assessment. In column (4), we respecify our model by lagging the regressors for gross and taxable profits.

Table 7: Results of regression prediction model for audit outcomes

Variables	(1) assesstot	(2) assesstot	(3) L1.assesstot	(4) assesstot
Gross profit	-0.0002	0.729***	0.0008***	0.0027*
	(0.0002)	(0.163)	(0.0001)	(0.0015)
L1.Gross profit				-0.0022**
				(0.0008)
Taxable profit	0.0003***	-0.304*	0.0001	0.0005**
	(0.0001)	(0.162)	(0.0004)	(0.0002)
L1.Taxable profit				0.005**
				(0.0002)
Cost of sales (expenditures)	0.0001	0.259*	0.0002***	-0.0001
	(0.0000)	(0.125)	(0.000)	(0.0001)
Tax on taxable profit	-0.0007	0.043**	-0.0015	-0.0100**
	(0.0007)	(0.128)	(0.0022)	(0.0026)
Total employment expenditures	0.0045*	-0.310**	-0.0061***	0.045
	(0.0019)	(0.123)	(0.0011)	(0.0027)
Sector FE	Yes	Yes	Yes	Yes
Area FE	Yes	Yes	Yes	Yes
Return year FE	Yes	Yes	Yes	Yes
Observations	2,334	289	630	1,680
Overall R-squared	0.030	0.381	0.057	0.555
Number of tpins	1,736	248	437	1,276

Note: the dependent variable is the additional tax owed or refunded. Columns (1)–(4) show prediction estimates of tax audits in Zambia. Column (1) shows estimates using the risk parameter variable in levels. In column (2) we show prediction estimates using inverse hyperbolic sine transformation changes in risk parameter variables. In column (3) we show a lagged specification of the dependent variable. In column (4) we lag 2 regressors (gross and taxable profit) while keeping the dependent variable at levels. Robust standard errors clustered at the sector level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Source: authors' calculations based on ZRA administrative data.

From column (1) we can predict assessments on a greater number of unaudited firms (2,334) as compared with the smaller numbers in columns (2), (3), and (4) for the same cohort of unaudited firms (289, 628, and 1,680, respectively). As we seek a tax gap that is more representative of the total population of firms whether audited or not, we use the model prediction results in column (1) for the study to obtain the audit outcome for the unaudited firms. Although the number of audited firms is greater in the administrative data, we use the smaller sample in the analysis, as we encounter problems when combining audit data with VAT and CIT filings. Some audited firms are missing tax returns because the tax authority rejects their filings due to discrepancies observed after an audit; hence, the respective assessments generated for such cases cannot be linked to any return filings for VAT or CIT. This limits the number the audit cases whose assessments can be directly linked to returns data.

After predicting the true tax liability for the unaudited sample according to the specification in Equation 2, we use the firm-level evasion estimates to calculate a nationwide tax gap. The tax gap calculations based on Equation 1 show that the total tax gap is 55.6 per cent[5] after trimming[6] for outlier tax returns. We classify this gap as a general tax gap, although it is based on VAT and CIT

[5] We also noisily estimate the total VAT gap as 84 per cent and the total CIT gap as 63 per cent. These estimates are described as noisy because they are very vague and exploratory, since we do not know whether the audits were VAT or CIT triggered.

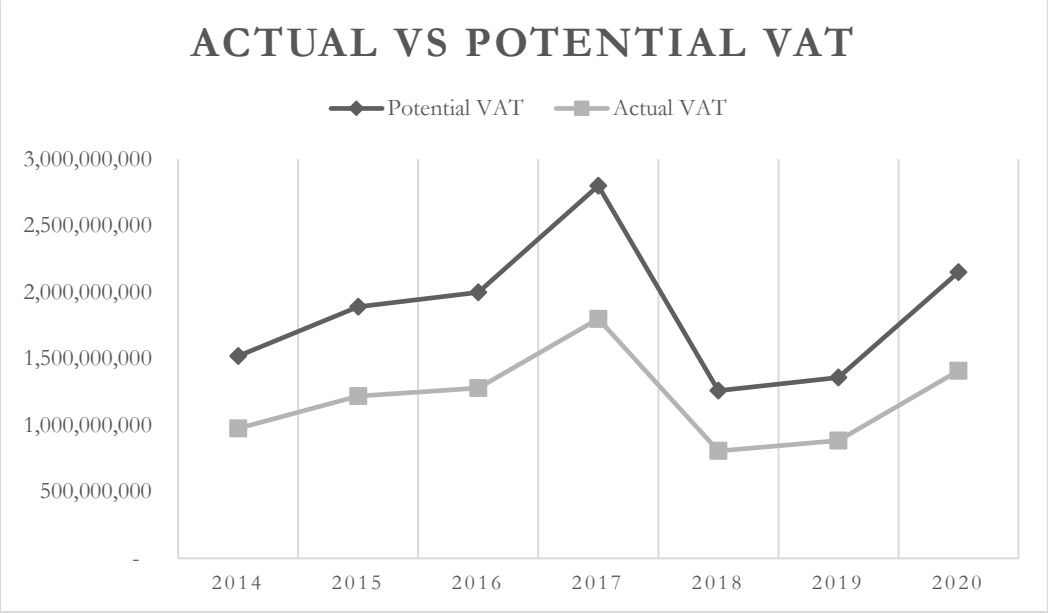
[6] We winsorize the variable by 1 per cent each end.

audits, because we continue to use predicted assessments based on audits that cannot be classified as VAT or CIT targeted. Having more precision in the prediction regression for audits as well as obtaining all filings for all audited firms might improve the results. We apply the estimated tax gap on actual VAT and CIT filings to obtain potential filings for all firms.

In Figures 3–5, we observe the trends in actual and potential VAT returns, CIT returns, and an aggregate of these returns when we apply the tax gap estimate. We obtain actual annual VAT and CIT figures from the annual or monthly returns reported by firms to the tax authority. Potential VAT and CIT figures are those adjusted using our tax gap estimate as calculated using the methods discussed. We observe the highest potential annual VAT return of K2.8 billion (~US$150 million) in 2017, while the highest potential CIT return is K12 billion (~US$633 million) as compared with approximate actual annual returns of K1 billion (~US$55 million) and K6.2 billion (~US$350 million) in 2014 and 2020, respectively. Gaps in the VAT returns tend to be large and constant in magnitude, while those in the CIT returns tend to be smaller in the earlier years (2014–16) but larger in the more recent years (2017–20). The aggregate graph in Figure 5 reflects CIT as the main driver of tax revenue rather than VAT. Most VAT returns for the firms are in net position. The aggregation nets out the refunds—hence the low VAT values as compared with the CIT returns, which are mostly reported in debit position.

Based on our gap estimate of 56 per cent, the country could have made K17 billion (~US$900 million) in additional domestic revenue from VAT and CIT returns. Most of this gap is generated by firms paying CIT rather than net payable VAT. From tax records we observe that big corporations mostly report credit positions, especially for VAT, as shown in Table 6, where the audited group in sectors like manufacturing, mining, construction, and real estate record negative VAT positions.

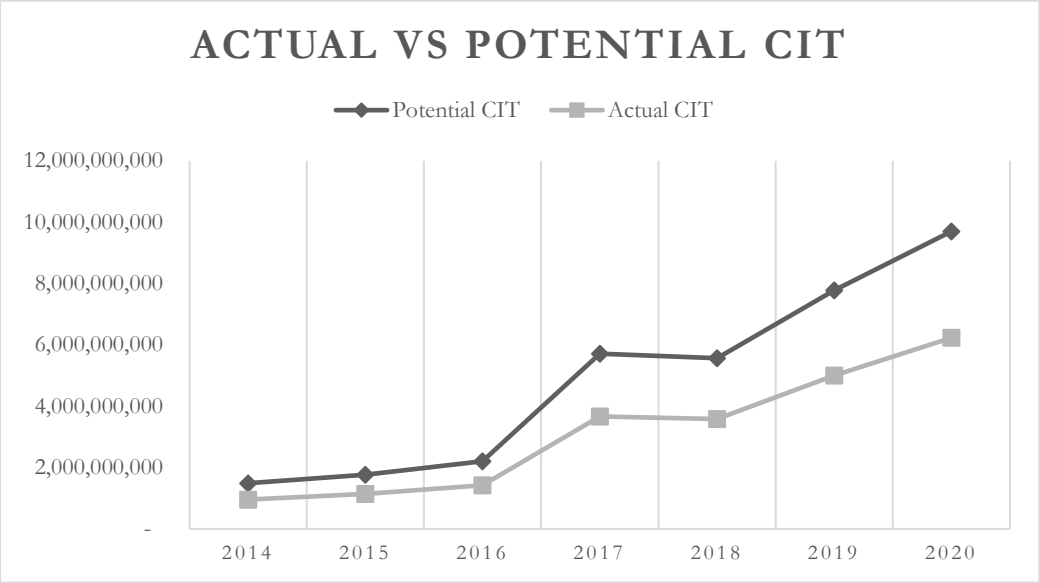
Figure 3: Total actual vs. potential VAT in ZMK



Note: the potential VAT is the aggregated potential VAT of all filing firms in the year, where the potential VAT for each firm is the reported tax + the predicted evasion. Return values are based on study data and may deviate from official values.

Source: authors' calculations based on ZRA administrative data.

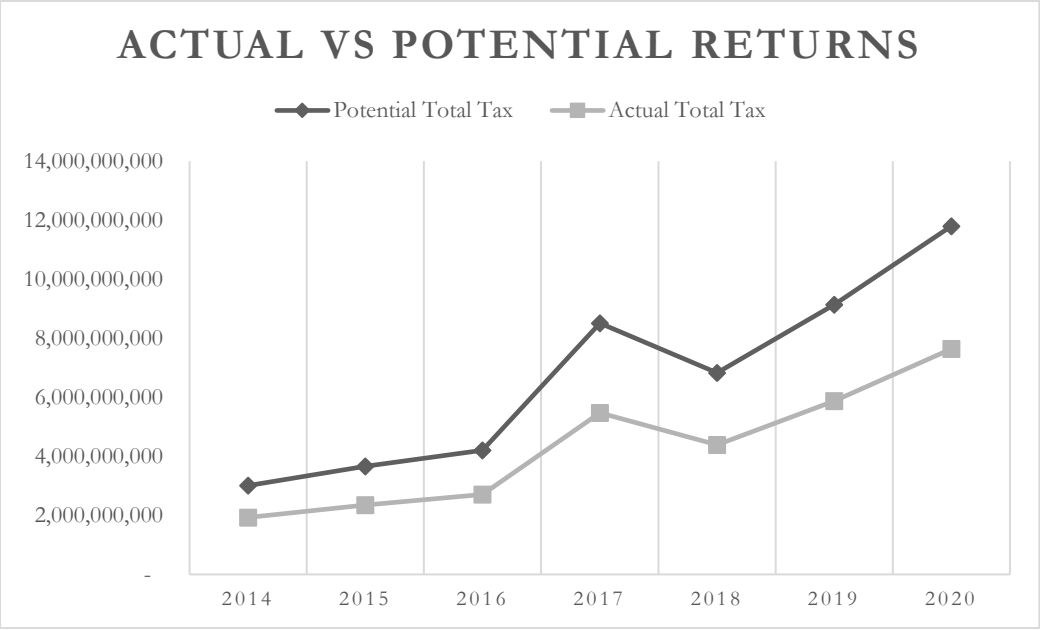
Figure 4: Total actual vs. potential CIT in ZMK



Note: the potential CIT is the aggregated potential CIT of all filing firms in the year, where the potential CIT for each firm is the reported tax + the predicted evasion. Return values are based on study data and may deviate from official values.

Source: authors' calculations based on ZRA administrative data.

Figure 5: Total actual vs. potential combined VAT and CIT filings in ZMK



Note: the potential tax is the aggregated potential revenue of all filing firms in the year, where the potential tax revenue for each firm is the reported tax + the predicted evasion. Return values are based on study data and may deviate from official values.

Source: authors' calculations based on ZRA administrative data.

4.2 Estimation of tax gap with machine learning approach

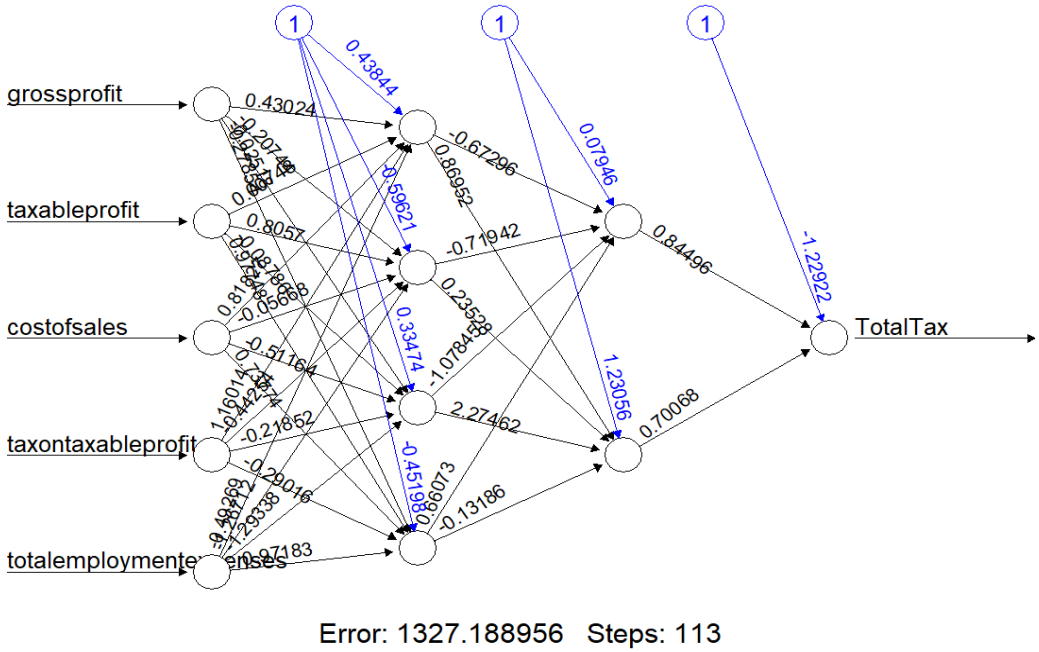
The second approach builds an Artificial Neural Network (ANN) of tax parameters that may be connected to the total tax assessment of an audit (evasion). There are three reasons for the choice of the model: first the model's ability to learn and model complex non-linear relations by mimicking real life scenarios; second, the model's generalizing capabilities by studying initial input

variables and inferencing unseen relationships on new data through accurate predictions; and third, the unrestricted nature of input variables the model can take, as no specific prior distribution is required before the model can function (Hoepner et al. 2021).

In this estimation, we build an input-dependent hidden-layered neural network model by interacting all variables that may be connected to the outcome of an audit (total tax assessment). For comparability we select the same variables as those used in the regression approach, i.e. gross profit, taxable profit, tax on taxable profit, cost of sales, employee expenditure, and return year controls. Data constraints prevent us from including additional variables, as the ANN approach is unresponsive to the missing cells common to most non-filing firms. The advantage of this approach is that it extends the model specification with numerous interaction terms based on the input variables introduced into the ANN through the hidden layers. The latter is a vector of integers specifying the number of hidden neurons (vertices) in each layer where variable interactions take place.

In this study, we specify two hidden layers with four and two vertices of neurons, forming neural networks (interaction terms) that we believe will be powerful enough to provide robust predictions of the extent of evasion. We train our ANN model to select the best prediction after 1,000 repetitions with the least error based on the audited sample. We show our best model prediction of total tax assessments with the least error in Figure 6, where each network branch shows a coefficient for each learning interaction. From the figure we observe that the maximum number of steps that leads to a stop in the neural network's training process is 113 after running 1,000 iterations. We are unable to list all the interaction terms, as the ML approach relies on numerous combinations to provide the estimates.

Figure 6: ANN visual model



Note: in addition to the variables in the visualization, the model includes industry and year variables.

Source: authors' calculations based on ZRA administrative data.

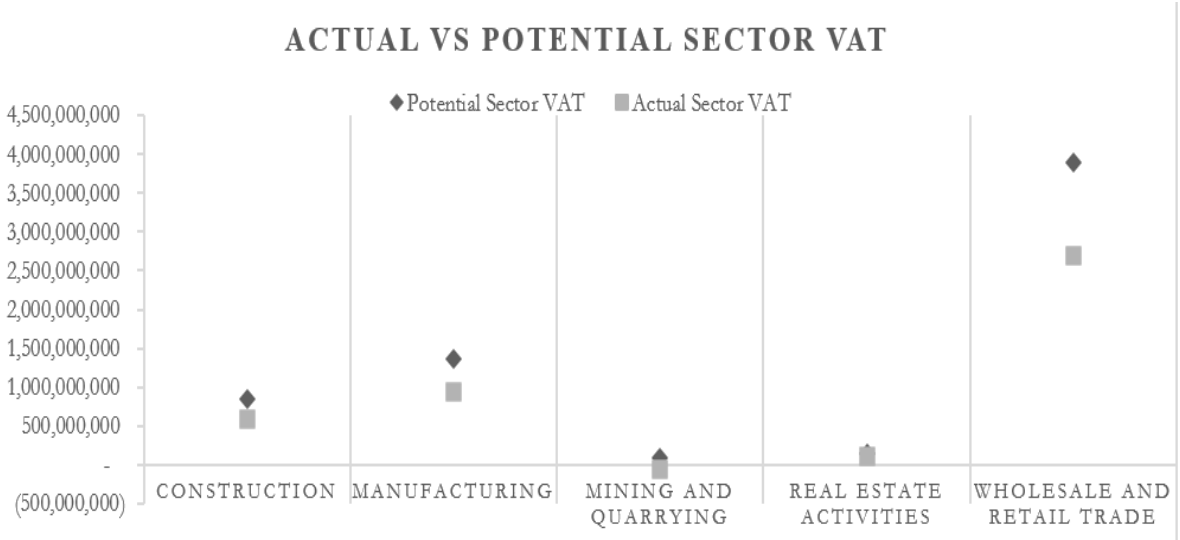
We use the coefficients obtained to predict the true tax liability of the unaudited sample. This is done as the ANN process learns from the characteristics of the audited sample, through which we obtain total tax assessment estimates for the full population of unaudited firms. Unaudited firms

are those for which we have no reported true tax assessments, but we believe that our ML model based on all specified inputs (firm characteristics) can learn and train the data through neural pathways or nets to produce robust tax assessment estimates. The method of prediction here is superior to the regression approach not only in introducing additional interacting covariates, but in two other ways. First, the inclusion of two hidden layers with four and two node points improves the goodness of fit by introducing additional interaction terms. Second, the steps ANN takes (113 in this study) to train the data in model to mimic the characteristics of audited firms improves its predictive power. Last, the number of iterations run in the model before settling on the optimum leads to better results. Having estimates of the true tax liability for the full population, we apply Equation 1 again for the total gap estimate based on the ML approach.

After the prediction exercise from the ML approach, we obtain a total tax gap estimate based on both predicted and actual assessments of 46.5 per cent.[7] Although this is marginally lower than the prediction from the regression approach, we believe that there is an increase in the precision of predicting total audit assessments. Results are improved because our ML approach trains the model to improve the predictions. We apply the estimated tax gap on actual VAT and CIT filings to obtain potential filings for all firms and show the results in the graphs in the Appendix.

As we believe in the robustness of the ML approach, we show the sectoral breakdown of potential and actual VAT, CIT, and aggregate filings over the the return years. We use the absolute value of predictions even though the actual total return may be negative. For instance, in the mining and quarrying sector actual total VAT was in a refund position. However, we predict a positive potential total VAT return based on our gap estimate. In Figures 7–9 we present actual and potential filings for five key sectors of the Zambian economy: construction, manufacturing, mining and quarrying, real estate activities, and wholesale and retail trade.

Figure 7: Total actual vs. potential VAT by key sector in ZMK

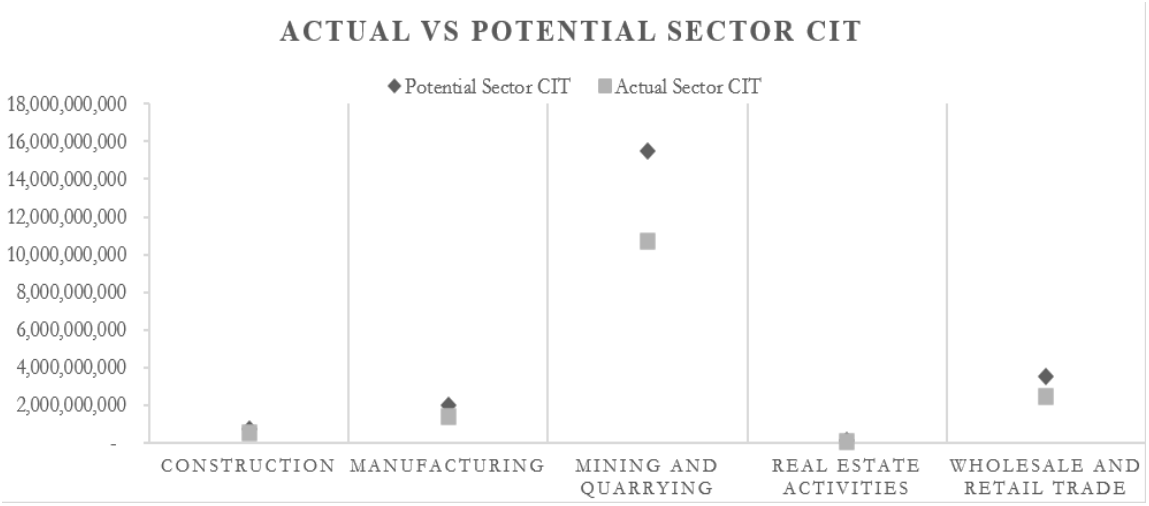


Note: the potential VAT is the aggregated potential VAT of all filing firms in the sector, where the potential VAT for each firm is the reported tax + the predicted evasion. Return values are based on study data and may deviate from official values.

Source: authors' calculations based on ZRA administrative data.

[7] We again noisily estimate the total VAT gap as 78 per cent and the total CIT gap as 54 per cent. These estimates are described as noisy because estimates are very vague and exploratory, since we do not know whether audits are VAT or CIT triggered.

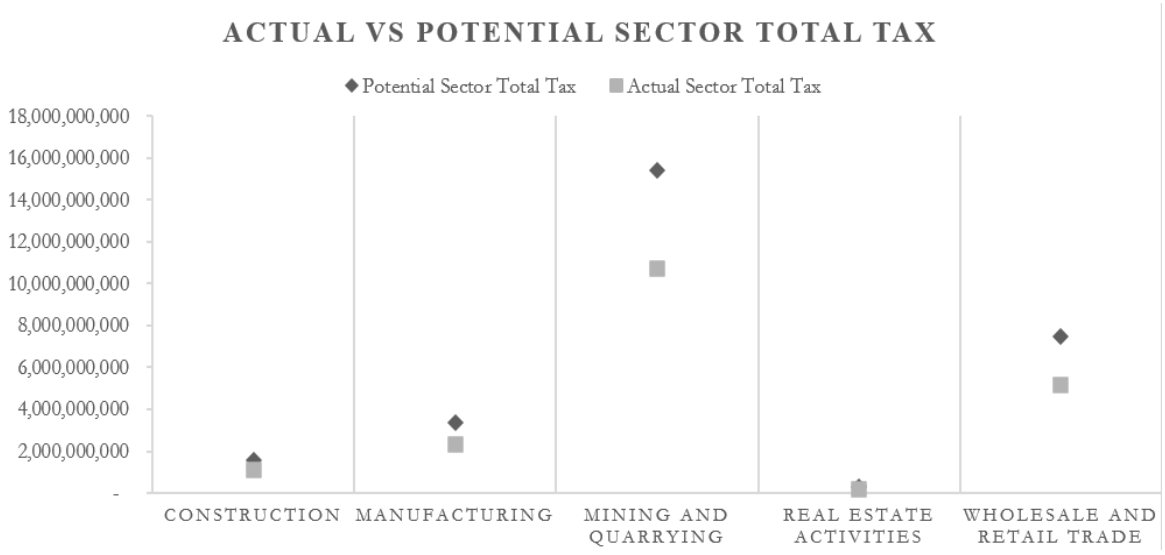
Figure 8: Total actual vs. potential CIT by key sector in ZMK



Note: the potential CIT is the aggregated potential CIT of all filing firms in the sector, where the potential CIT for each firm is the reported tax + the predicted evasion. Return values are based on study data and may deviate from official values.

Source: authors' calculations based on ZRA administrative data.

Figure 9: Total actual vs. potential sector returns by key sector in ZMK



Note: the potential revenue is the aggregated potential tax of all filing firms in the sector, where the potential tax revenue for each firm is the reported tax + the predicted evasion. Return values are based on study data and may deviate from official values.

Source: authors' calculations based on ZRA administrative data.

We observe interesting results based on sectoral disaggregation of potential and actual VAT and CIT returns. In general, the gaps observed in VAT returns are much lower than those observed in CIT returns. Although the wholesale and retail sector records large VAT gaps, this is mainly due to the high number of firms recorded in the data for that sector, which affects the actual value of reported tax. Moreover, VAT gaps for the mining and quarrying and the real estate sectors are minimal compared with the other key sectors. This is without doubt due to the nature of VAT returns in the extractives sector, which mainly exports its final output, so that not much VAT is paid on final output in the mining and quarrying industry. However, we observe that in CIT gap sectoral comparisons, the extractives sector records quite high gaps compared with the other key sectors. The real estate sector continues to record the smallest gap (K88 million) over the period

as compared with the K4.7 billion in the extractives sector. Between these two extremes and in increasing order are construction (K492 million), manufacturing (K1 billion), and wholesale and retail (K2.3 billion).

4.3 Validation of tax gap estimates

We perform sensitivity checks on the gap estimates to confirm the stability of our results. First, we compare our estimates with other known studies for deviations or similarities. Although not many studies employ bottom-up approaches, it is interesting to know how our estimates compare with other tax gap estimates irrespective of the models or approaches used. In addition to this, we check on the reliability of the ML model based on the ANN mechanism. Employing another ML approach, in the form of Random Forest, we check the closeness of our standard results and make conclusions therefrom.

We compare our gap estimates with those of Alexeev and Chibuye (2016), who assess similar tax gaps in Zambia using a top-down approach and report VAT gaps over the period 2009–11 as between 50 per cent and 30 per cent of total VAT liability. Our estimates of 56 per cent and 47 per cent (on average 52 per cent) for total CIT and VAT gaps fall in the same ballpark. However, given our employed methods and the data discrepancy noted with the top-down approach in Zambia, we argue for higher precision with our bottom-up approach.

Our comparison with GDP reveals smaller rates than the 2.5 per cent average obtained by Alexeev and Chibuye (2016). However, as our estimate is calculated only for tax-filing firms in the VAT and CIT net and leaves aside the fully informal economy, it is likely that the monetary value of the revenue loss is lower bound. Table 8 shows the total tax gap obtained from VAT and CIT returns using the ML approach,[8] GDP in local currency units, and the percentage of the gap to GDP. The percentage ranges between 0.5 and 1 in the period 2014–20, as compared with Alexeev and Chibuye's (2016) reported VAT-to-GDP gaps of 3 per cent in 2009 and 2 per cent in both 2010 and 2011.

Table 8: Tax gap–GDP nexus using ANN predictions

Series name	2014	2015	2016	2017	2018	2019	2020
GDP (current LCU)	167,053	183,381	216,098	246,252	275,174	300,448	332,223
Tax gap (VAT + CIT)	864	1,050	1,210	2,440	1,960	2,630	3,400
Gap % of GDP	0.52	0.57	0.56	0.99	0.71	0.88	1.02

Note: GDP and total tax gap in ZMK'millions.

Source: authors' calculations based on ZRA administrative tax return data and WDI.

Similarly high gaps have been recorded with the bottom-up method in other developing countries. We also compare our gap estimates with those of Best et al. (2021), who study heterogenous tax gaps by firm size in Pakistan. The authors observe that tax gaps are particularly high for smaller firms (bottom 20 per cent of the distribution), where the evasion rate exceeds 80 per cent. Administrative data indicate that there is a greater share of smaller firms in Zambia, where about 43 per cent of firms record annual turnover between K800,000 (US$50,000) and K900,000 (US$60,000). From these facts our average tax gap estimate of 52 per cent is a not particularly high estimate of the actual tax compliance conditions in the country.

[8] See similar results for the regression approach in the Appendix.

Using a Random Forest application to predict audit outcomes for a similar testing sample of unaudited firms, we estimate tax gaps for comparability with our standard ML approach. In essence the approach produces predictors from random vector combinations independently sampled from a distribution of decision trees collectively referred to as the forest (Breiman 2001). The reason for choosing this approach is its similarity to a neural network, although it formulates variable interactions into layers of average decision trees rather than neural nets. In our application, we run 100 iterations, through which the model produces a slightly higher tax gap of 50.4 per cent,[9] which is quite close to both our standard ML approach and the estimate based on the regression approach. The actual and evasion volumes used in the Random Forest method could be enhanced to approach our standard results. However, for comparability using similar base parameters we keep predictions as they are and conclude that the results based on predictions from our ANN are stable. We base such conclusions on the idea that predictions do not go through an extra layer of polishing before tax gap estimations but derive solely from what the ML model learns using the training data.

5 Conclusion

This paper uses the universe of Zambian value added and corporate income tax filers, together with audit information, to study the difference between potential and actual tax liability. The audit information is used as a base to predict evasion for firms that were unaudited. The study assesses how far the revenue authority might be from its revenue potential. The challenge in this nature of research lies in the precision of evasion rate predictions for the unaudited sample, which usually goes unnoticed. The study takes rudimentary to complex steps to improve the predictions of total tax assessments used to calculate tax gaps. In this way, we are able to judge the sensitivity of our estimates when applicable methods change.

First, we generalize our problem as a basic regression model, where we estimate an equation for evasion using specific risk parameters as defined by the audit department of the tax authority. These coefficients become the basis for predicting the audit outcomes for the larger population of unaudited firms. Using these predictions, the study finds total tax gaps of 56 per cent of total VAT and CIT liability, where the latter is the major driver of the gap estimate, as can be seen from the potential volumes when the estimate is applied to the actual volumes. This estimate is similar in magnitude to evidence from other developing countries (e.g. Best et al. 2021).

As the regression may not be exhaustive in terms of the covariates, the study devises other means to measure the tax gap. The study calls on a machine learning algorithm based on Artificial Neural Network technology, which has the potential to learn from a training sample to make predictions on a testing sample. In our case, we train the algorithm on the sample of audited firms to predict audit outcomes for the sample of unaudited firms. Combining these outcomes for all the firms, the study finds a total tax gap of 47 per cent, which is slightly lower than our estimates obtained from the regression model.

Although the first of its kind, our study generates results that are stable, given our robustness checks on the gap estimates. Most importantly, we apply another ML algorithm based on the

[9] We similarly estimate 80.3 per cent and 57.5 per cent as the VAT and CIT gaps, respectively. These estimates include all data points for actual and predicted audit outcomes as used in our standard ML approach. We obtain fairly close estimates for total, VAT, and CIT gaps as 48.5 per cent, 79.1 per cent, and 55.6 per cent, respectively, when using a winsozorised sample at the 1st and 99th percentile of the total tax assessment distribution.

Random Forest approach, which produces slightly higher gap estimates (when predictions are untrimmed). With extra effort on prediction data, gap estimates seem to approach estimates in standard methods.

We rely on the ML method estimate to further calculate potential VAT and CIT returns in specific industries. The wholesale and retail sector records the highest potential VAT, while the extractives sector records the highest potential CIT. This reflects the fact that wholesale and retail trade attracts more VAT while extractives trade attracts less, as most of its output is exported. Our primary estimates are for total tax gaps even though we also estimate gaps for VAT and CIT separately. We cannot clearly delineate the motive for triggering an audit as VAT or CIT, since audits conducted by the Zambian Revenue Authority within our period of interest probed all the filings of a firm irrespective of the trigger. Thus, we suggest interpreting the separate estimates for CIT and VAT gaps with caution, as the identification based on the audit status might bias our results. We nevertheless present these estimates to show how the gaps might deviate if audits were specific to VAT or CIT.

To sum up, given our tax gap estimates of around 50 per cent, there is potential for the ZRA to increase its tax revenue significantly with the current tax parameters via enforced compliance. As this study has shown, VAT audits might prove cost-effective if targeted at firms in the wholesale and retail trade, while CIT audits should be targeted at the extractives industries or firms that export most of their output. An additional recommendation to the audit department within the revenue authority is to specify and group audits as VAT- or CIT-triggered. As this study concentrates on the compliance gap, there is need for further studies to investigate the tax gaps due to policy lapses in the economy.

References

Alexeev, M., and B. Chibuye (2016). 'Estimating the Value Added Tax (VAT) Gap in Zambia: 2009–2011'. Lusaka: International Growth Center.

Al-Mobayed, A.A., Y.M. Al-Madhoun, M.N. Al-Shuwaikh, and S.S. Abu-Naser (2020). 'Artificial Neural Network for Predicting Car Performance Using JNN'. *International Journal of Engineering and Information Systems*, 4(9): 139–45.

Almunia, M., and D. Lopez-Rodriguez (2018). 'Under the Radar: the Effects of Monitoring Firms on Tax Compliance'. *American Economic Journal: Economic Policy*, 10(1): 1–38. https://doi.org/10.1257/pol.20160229

Battaglini, M., L. Guiso, C. Lacava, D.L. Miller, and E. Patacchini (2022). 'Refining Public Policies with Machine Learning: the Case of Tax Auditing'. NBER Working Paper w30777. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w30777

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

Boutaba, R., M.A. Salahuddin, N. Limam, S. Ayoubi, N. Shahriar, F. Estrada-Solano, and O.M. Caicedo (2018). 'A Comprehensive Survey on Machine Learning for Networking: Evolution, Applications and Research Opportunities'. *Journal of Internet Services and Applications*, 9(1): 1–99. https://doi.org/10.1186/s13174-018-0087-2

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

Danquah, M., and E. Osei-Assibey (2018). 'The Extent and Determinants of Tax Gap in the Informal Sector: Evidence from Ghana'. *Journal of International Development*, 30(6): 992–1005. https://doi.org/10.1002/jid.3361

HMRC (2022). 'Measuring Tax Gaps 2022 Edition: Tax Gap Estimates for 2020 to 2021'. London: HM Revenue and Customs. Available at: https://www.gov.uk/government/statistics/measuring-tax-gaps (accessed 27 January 2023).

Hoepner, A.G., D. McMillan, A. Vivian, and C. Wese Simen (2021). 'Significance, Relevance and Explainability in the Machine Learning Age: an Econometrics and Financial Data Science Perspective'. *The European Journal of Finance*, 27: 1–7. https://doi.org/10.1080/1351847X.2020.1847725

Hutton, E. (2017). 'The Revenue Administration–Gap Analysis Program: Model and Methodology for Value-Added Tax Gap Estimation'. Washington, DC: International Monetary Fund. https://doi.org/10.5089/9781475583618.005

Jansen, A., W. Ngobeni, A. Sithole, and W. Steyn (2020). 'The Corporate Income Tax Gap in South Africa'. WIDER Working Paper 2020/40. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2020/797-2

Lakuma, C.P., and B. Sserunjogi (2018). 'The Value Added Tax (VAT) Gap Analysis For Uganda'. Research Series 145. Kampala: Economic Research Policy Centre.

Mascagni, G., D. Mukama, and F. Santoro (2019). 'An Analysis of Discrepancies in Taxpayers' VAT Declarations in Rwanda'. ICTD Working Paper 92. Brighton: Institute of Development Studies.

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

Ogwueleka, F.N., S. Misra, R. Colomo-Palacios, and L. Fernandez (2015). 'Neural Network and Classification Approach in Identifying Customer Behavior in the Banking Sector: a Case Study of an International Bank'. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 25(1): 28–42.

Ramírez-Álvarez, J., and P. Carrillo Maldonado (2020). 'Indicator of the Efficiency of Value Added Tax and Income Tax Collection in Ecuador'. *CEPAL Review*, 2020(131): 69–86. https://doi.org/10.18356/16840348-2020-131-4

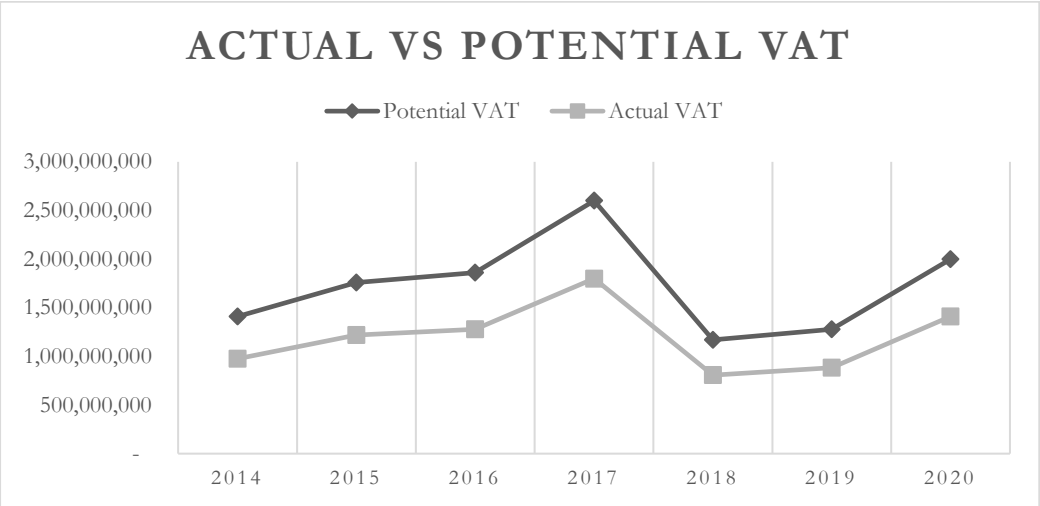
Slemrod, J., O.U. Rehman, and M. Waseem (2019). 'Pecuniary and Non-Pecuniary Motivations for Tax Compliance: Evidence from Pakistan'. NBER Working Paper w25623. Bonn: National Bureau of Economic Research. https://doi.org/10.3386/w25623

Waseem, M. (2018). 'Taxes, Informality and Income Shifting: Evidence from a Recent Pakistani Tax Reform'. *Journal of Public Economics*, 157: 41–77. https://doi.org/10.1016/j.jpubeco.2017.11.003

Zídková, H. (2014). 'Determinants of VAT Gap in EU' [*sic*]. *Prague Economic Papers*, 23(4): 514–30. https://doi.org/10.18267/j.pep.496

Appendix: Additional results

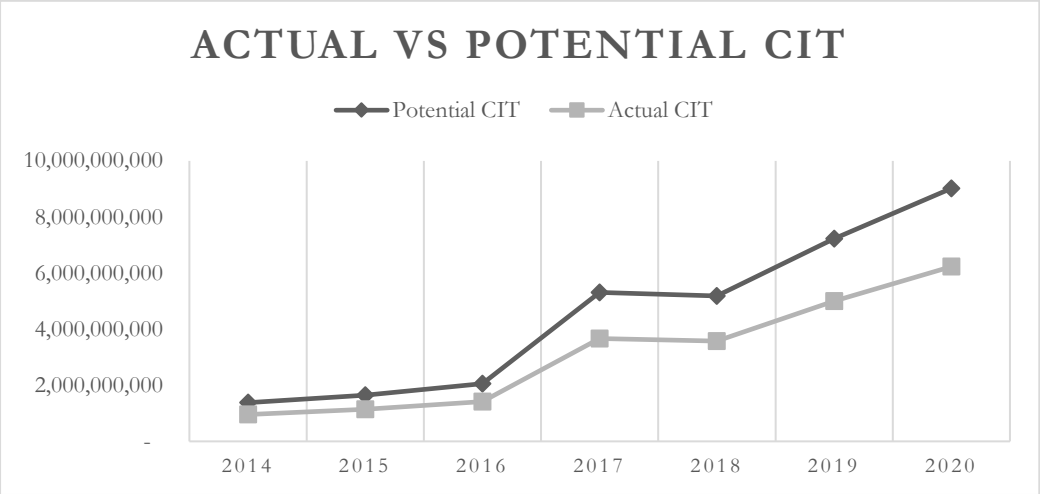
Figure A1: Mean actual vs. potential VAT in ZMK using ML



Note: return values are based on study data and may deviate from official values.

Source: authors' calculations based on ZRA administrative data.

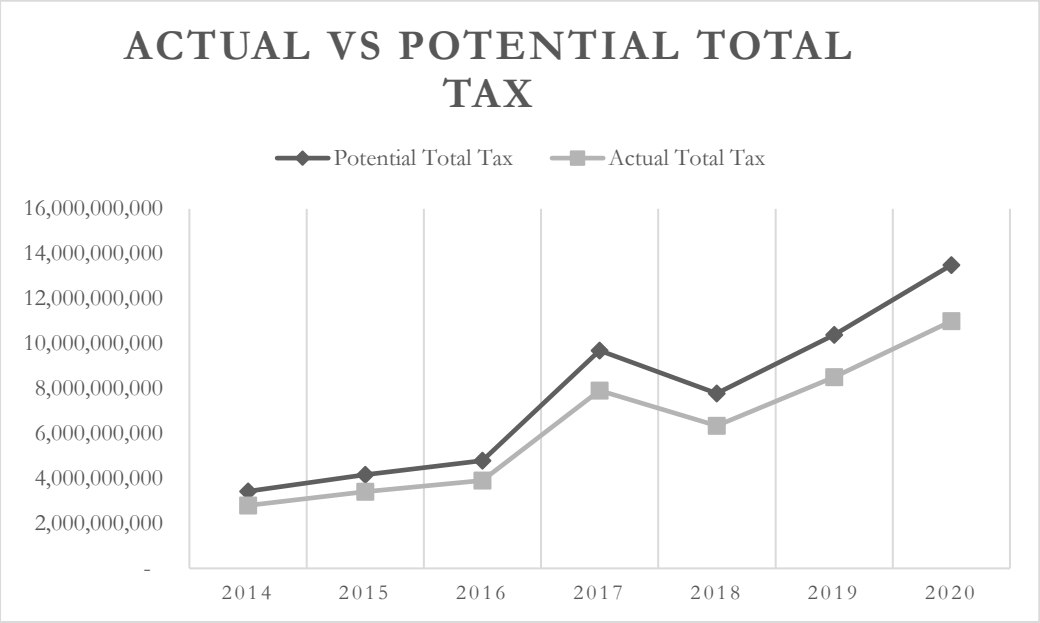
Figure A2: Mean actual vs. potential CIT in ZMK using ML



Note: return values are based on study data and may deviate from official values.

Source: authors' calculations based on ZRA administrative data.

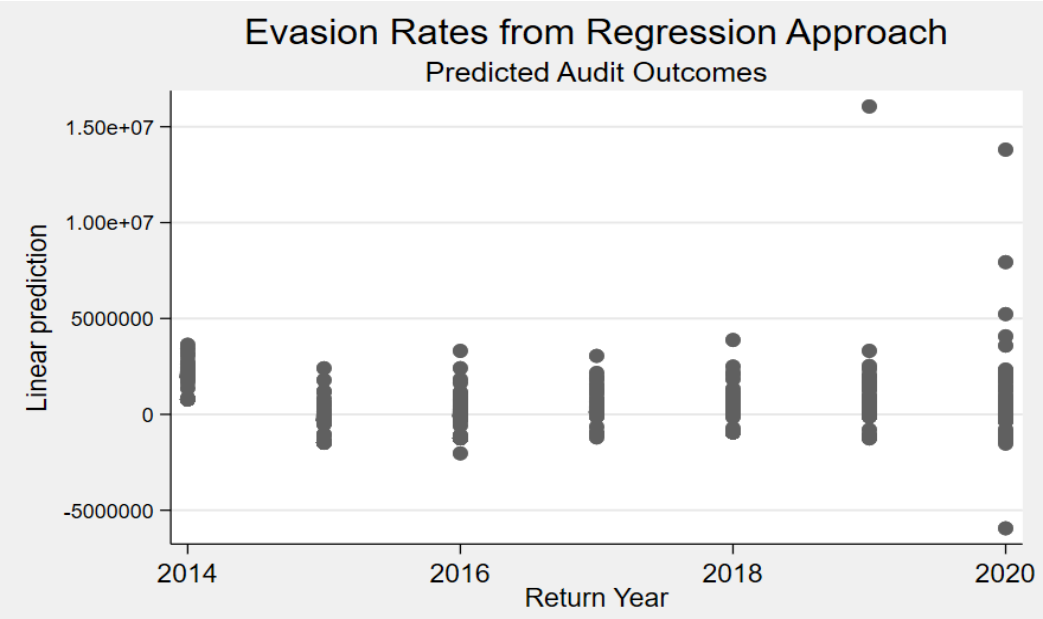
Figure A3: Actual vs. potential combined VAT and CIT filings in ZMK using ML



Note: return values are based on study data and may deviate from official values.

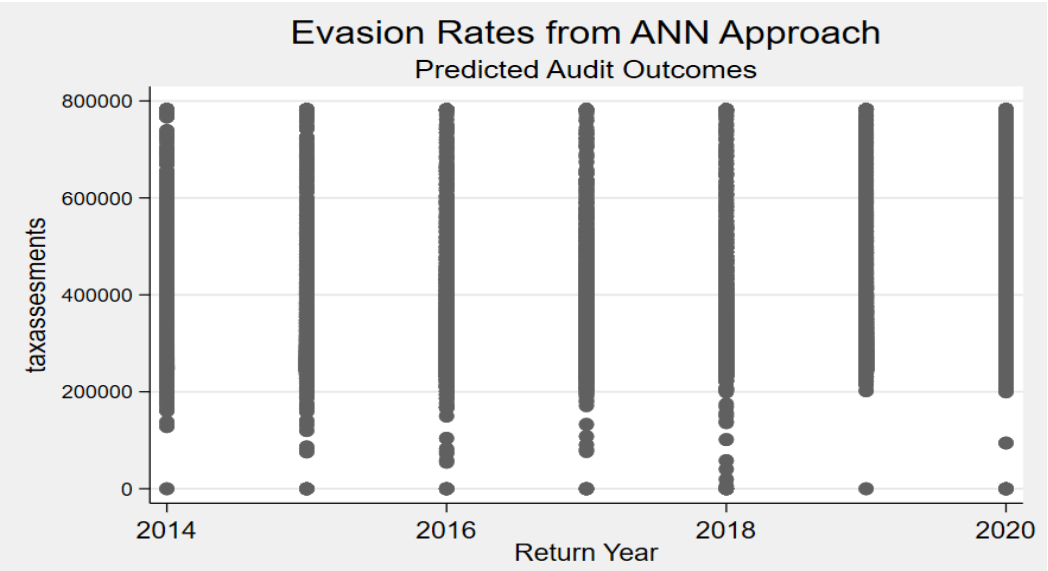
Source: authors' calculations based on ZRA administrative data.

Figure A4: Predicted evasion rates using regression estimates



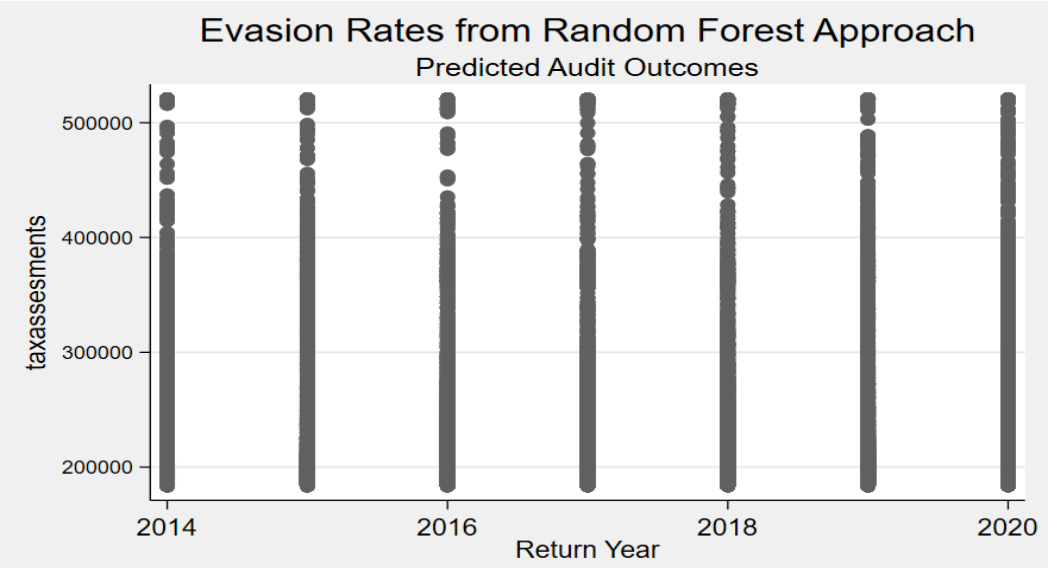
Source: authors' calculations based on ZRA administrative data.

Figure A5: Predicted evasion rates using ANN estimates



Source: authors' calculations based on ZRA administrative data.

Figure A6: Predicted evasion rates using Random Forest estimates



Source: authors' calculations based on ZRA administrative data.

Table A1: Tax gap–GDP nexus based on regression predictions

Series name	2014	2015	2016	2017	2018	2019	2020
GDP (current LCU)	167,053	183,381	216,098	246,252	275,174	300,448	332,223
Total Tax Gap	1,070	1,310	1,500	3,040	2,440	3,270	4,230
Gap % of GDP	0.64	0.71	0.69	1.23	0.89	1.09	1.27

Notes: GDP and total tax gap in ZMK'millions.

Source: authors' calculations based on ZRA administrative tax returns data and WDI.

Table A2: Characteristics of Zambia's taxes in 2022

	CIT	VAT	Turnover tax
Rates	Standard rate: 30% Other rates: 40%, 15%, 10%	Standard rate: 16%	Standard rate: 4%
Thresholds	Annual gross sales/Turnover exceeding K800,000	Annual gross sales/ Turnover exceeding K800,000	Annua gross sales/ Turnover below K800,000
Liability and exemptions	30% for certain industries with annual profits below K250,000 40% for certain industries with annual profits exceeding K250,000 10% for industries in farming and agro-processing 15% for certain industries in manufacturing	16% across all industries above the threshold	4% across all industries below the threshold

Source: authors' calculations based on ZRA administrative data.