UNU WIDER

2
24

Tax research for development

Toolkit for the estimation of tax gaps using a bottom-up approach

Mostafa Bahbah,[1] Sebastián Castillo,[2] Kwabena Adu-Ababio,[3] and Amina Ebrahim[3]

December 2024

WIDER TECHNICAL NOTE

wider.unu.edu

United Nations University
World Institute for Development Economics Research

Abstract: This technical note relates to the tax gap toolkit, which includes code (Stata do files) and a README file (how to run the code). The note describes the literature and methodology behind the development of the toolkit. The tax gap toolkit is related to the estimation of value-added tax (VAT), corporate income tax (CIT), and personal income tax (PIT) gaps using the bottom-up approach, where operational audits and potential tax are calculated using machine learning.

Key words: toolkit, tax gap, bottom-up approach, operational audits, machine learning

JEL classification: H25, H26, H32

Acknowledgements: The authors gratefully acknowledge inputs from Jukka Pirttilä, Maria Jouste, Gerald Agaba, and Hilja-Maria Takala. The authors acknowledge funding from the International Tax Compact (ITC). The ITC facilitates the Secretariat of the Addis Tax Initiative (ATI), which supported the development of the Tax Gap toolkit. The ITC is funded by the German Federal Ministry of Economic Cooperation and Development, co-funded by the European Union, and implemented by the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. This work is part of UNU-WIDER's Domestic Revenue Mobilization programme, which is funded by Norad.

Supplementary material: The related Stata code and README files are freely available to download from the toolkit's webpage.

Related publications:

- Estimating tax gaps in Zambia: WIDER Working Paper 2023/25
- Estimating the value-added tax gap in Tanzania: WIDER Working Paper 2024/66

Note: This technical note is available in Spanish, French, and Portuguese.

[1] Tampere University, Finland; [2] University of Helsinki, Finland, and Finnish Centre of Excellence in Tax Systems Research (FIT); [3] UNU-WIDER, Helsinki, Finland; corresponding author: amina@wider.unu.edu

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). The Tax Gap toolkit received financial support from the International Tax Compact (ITC), which facilitates the Secretariat of the Addis Tax Initiative (ATI). The ITC is funded by the German Federal Ministry of Economic Cooperation and Development, co-funded by the European Union, and implemented by the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH.

https://doi.org/10.35188/UNU-WIDER/WTN/2024-2

United Nations University World Institute for Development Economics Research – UNU-WIDER

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introduction

Mobilizing domestic revenue to finance public expenditure is an essential focus for many governments. Consequently, revenue authorities play a crucial role in fashioning efficient and effective tax systems that can ensure the optimal amount of revenue with minimum leakages. However, the optimal or maximum achievable revenue falls short of this desired outcome because liable individuals and firms go to extreme ends to avoid or evade such obligations. With this in mind, research aims to measure the gap between actual and potential tax revenue, quantifying the extent of revenue loss that occurs compared to a situation where all individuals and firms fully adhere to tax policy rules.

This technical note illustrates how tax gaps, defined as the difference between actual and potential tax revenue, are calculated based on a bottom-up methodology employing various forms of return and statistical data. Moreover, this note accompanies the tax gap toolkit as a precursor to the general concepts of tax gaps with emphasis on the bottom-up approach to estimating such gaps. Developed by UNU-WIDER, the toolkit simplifies the concepts of the estimation approach and guides users through systematic ways of measurement using available parameters within their reach.

With a broad definition for tax gaps, there is an accompanying variety of estimation approaches to the concept. However, they all address and investigate why actual tax revenue deviates from potential revenue. Some general and commonly used methods concentrate on aggregate macroeconomic indicators as a benchmark for the deviation, while the less employed methods rely on microeconomic data for gap estimation. The choice depends on the access and availability of administrative data depending on the jurisdictions. While this technical note introduces the main approaches to estimating tax gaps, the focus is on the use of micro-level data from operational audits. It is more often the case that operational (or risk-based) audits are conducted in developing countries while random audits are rare.

In the following note, we will first set out definitions for tax gaps in Section 2. Subsequently, the general methods used in estimating tax gaps are discussed in Section 3, where the bottom-up approach forms the bedrock approach, described in its various forms and methods. In Section 4, we describe the components of the toolkit, which comprises two main stages: data cleaning and a machine learning approach to tax gap estimations.

2 Tax gap definition

Tax authorities often face a notable difference between the expected tax revenue and what is collected. This difference, known as the revenue loss, primarily arises when taxes due within a certain period remain unpaid. This tax due from taxpayers represents the amount of tax that could theoretically be collected. This leads to the concept of the tax gap, defined as the difference between the actual revenue collected and the theoretical tax collections under full compliance within the tax code.

From a policy point of view, the tax gap can be characterized more broadly by two main components: the compliance gap and the policy gap. Compliance gap refers to the difference between the actual revenue gathered in a specific year and the maximum possible revenue that could have been obtained based on the economic activities occurring within that period. Policy gaps are a result of legislative decisions meant to modify standard tax regulations by introducing specific exemptions, deductions, or reduced rates for certain cases (Hutton 2017). Changes in the policy framework may cause the policy gap to grow or shrink. For instance, if the zero-tax threshold is raised, allowing a larger portion of income to become tax-free, or if a reduced tax rate is introduced for a specific group of taxpayers, such as small businesses or low-income individuals, the policy gap would increase as less revenue is collected compared to the potential maximum under the standard tax rules. On the other hand, the policy gap could also expand

without any changes in the policy framework due to changes in the tax base composition making a larger portion of net income subject to the standard tax rate (Barra et al. 2023).

The compliance gap consists of two elements: the assessment gap and the collection gap. The assessment gap mainly arises from economic activities that tax authorities are either unaware of or unable to reach, including activities by entities that are either not registered, fail to file, under-report, or misreport their taxes, as evidenced in jurisdictions with high informality. The collection gap refers to the discrepancy between the calculated tax liabilities, accounting for any refunds and withholdings, compared to the taxes that have actually been paid. It encompasses the outstanding tax amounts that tax authorities are aware of but have not successfully recovered, typically because these are tied up in disputes or are considered either too expensive to chase down or impossible to collect through legal means.

The literature also distinguishes four components of the compliance gap that complement the aforementioned assessment and collection gaps (Gemmell and Hasseldine 2014; Durán-Cabré et al. 2019).

1. *Under-reporting component:* the difference between potential and declared tax, reflecting the fraction of tax evaded through non-reporting or under-reporting to the tax authority. This may include reporting less income than earned or claiming more deductions, credits, or other tax benefits than allowed by law, or a combination of both. Also known as the reporing gap or the assessment gap.

2. *Non-filling component:* the potential tax revenue from registered taxpayers required to file a tax return but do not file. Also known as the filing gap or lodgement gap.

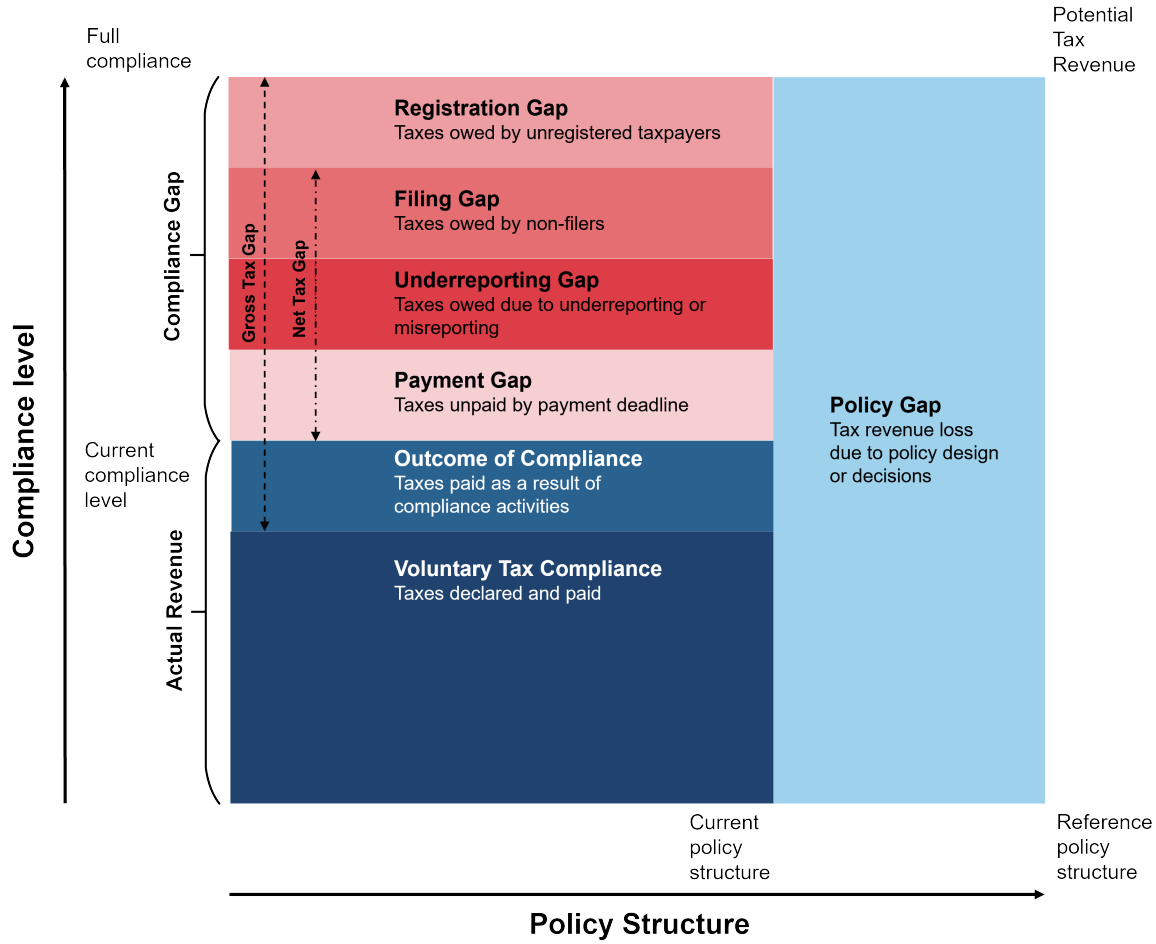
3. *Non-payment component:* the difference between potential and actual tax revenue, reflecting the fraction of tax evaded through non-payment to the tax authority. Also known as the underpayment, non-payment, collection or revenue gap.

4. *Non-registration component:* refers to the difference between the number of entities or individuals that should be registered for tax purposes (such as businesses, self-employed individuals, or property owners) and those that are actually registered. Also referred to as the registration gap.

Finally, from a collection point of view, some revenue authorities define the tax gap into two categories: the gross and the net tax gap.[1] For instance, the Internal Revenue Service (IRS) defines the gross tax gap as the discrepancy between the total true tax liability mandated by law for a specific tax year and the amount of tax that taxpayers voluntarily and timely pay for that year. On the other hand, the net tax gap refers to the remaining amount due of the total tax liability after accounting for all payments made through enforcement actions and voluntary late payments for a specific tax year (Plumley 2005). Figure 1 highlights the key components of the overall tax gap and the overlap between the different definitions of its components.

[1] It is important to note that definitions of the gross and net tax gap could have minor differences among different revenue authorities reflecting the unique tax enforcement environments and administrative priorities of each country.

Figure 1: Tax gap concepts



Note: simplified illustration of tax gap components.
Source: authors' illustration.

3 Tax gap methodologies

There are two general approaches to estimating the tax gap—the top-down and bottom-up approaches. The top-down approach uses aggregate-level data such as macroeconomic indicators or national account data to comprehensively assess all tax losses by measuring the gap as the difference between estimated potential revenue and actual revenues. However, it cannot determine the origins of the tax gap or explain why certain areas or activities remain untaxed. In contrast, the bottom-up approach uses micro-level data from tax administrations, including outcomes from random or operational audits aimed at particular criteria or other general administrative data from tax authorities. These data can be used to assess the extent of non-compliance of particular segments of the tax system, specific groups of taxpayers, or types of non-compliance (Hutton 2017).

3.1 Advantages and disadvantages of the bottom-up approach

Advantages

The bottom-up approach in estimating tax gaps offers several advantages over other methodologies, particularly in its ability to provide detailed insights (granular estimations) based on fiscal audits. Here are the key advantages:

- *Improved precision through detailed data*: The bottom-up method leverages granular data from financial audits, allowing for more precise estimations of the tax gap. This technique stands in contrast to top-down strategies, which depend on broad economic indicators and might overlook subtleties in the actions of individual taxpayers or particular industries.

- *Detailed insights for precise actions*: The bottom-up tactics deliver a detailed understanding of tax compliance at the individual or firm level. This level of detail enables tax authorities to craft precise interventions for particular industries, taxpayer categories, or instances of non-compliance, enhancing the efficiency and impact of enforcement measures (Hutton 2017).

- *Addressing selection bias in tax gap estimation*: Selection bias presents a major obstacle in accurately estimating the tax gap due to the non-representative nature of taxpayers chosen for fiscal audits. Using the bottom-up approach, particularly when integrated with machine learning techniques, can effectively mitigate this bias. This method does not depend on presumptions regarding the data's distribution, thereby providing robustness against any biases that might distort the tax gap estimation (Alaimo Di Loro et al. 2023).

- *Sector-by-sector analysis*: The bottom-up approach allows for a detailed sector-by-sector analysis of tax compliance gaps. Such detailed insights enable tax authorities to direct their compliance strategies more precisely, concentrating on sectors with the most significant gaps. This targeted approach could lead to improved efficiency in tax collection without the need for broad-based increases in audit or enforcement activities (Barra et al. 2023; Hutton 2017).

- *Adaptability to different tax types*: The flexibility of the bottom-up approach allows it to be adapted for estimating gaps in various types of taxes, including the value-added tax (VAT), corporate income tax (CIT), and personal income tax (PIT). This adaptability is crucial as different taxes face different types of compliance challenges and tax evasion tactics.

- *Enhanced tax compliance*: A bottom-up approach can provide insights into taxpayer behaviour, enabling the verification or refinement of existing models for identifying and managing risk. Pinpointing specific mistakes is also easier, allowing for effective redirection to alternative management approaches, such as enhancing taxpayer education, improving services, or conducting further audits and reassessments (Barra et al. 2023).

- *Allow for upper and lower bounds in the estimates*: The bottom-up approach enables the application of multiple techniques to the same taxpayer unit, in addition to enabling the statistical sensitivity analysis of the findings (Barra et al. 2023).

Disadvantages

Despite the strengths of the bottom-up approach, the literature indicates that it has the following limitations (Warren 2018; FISCALIS Tax Gap Project Group 2018).

- *Endogeneity:* This method relies heavily on existing knowledge and data within the revenue administration, making it less effective at capturing unknown factors or unobserved issues.

- *Challenges in accounting for unknowns:* Since the method is based on known data and operational results, it struggles to account for factors that are not easily observed, like under-reported income. It also does not cover the informal economy since only registered taxpayers can be selected for audit. As a result, estimates for these unknowns often involve rough adjustments, which can reduce accuracy.

- *Narrow focus:* This approach works from the specific to the general, zooming in on individual taxpayers. While this provides detailed insights, it may overlook macroeconomic trends or patterns.

- *Aggregation risk:* Bottom-up approaches only estimate components of the tax gap, requiring aggregation to estimate the total gap. However, this process carries a risk of double-counting and overestimating the total tax gap, especially when overlaps exist between different gap components.

3.2 Audit type

Tax authorities usually rely on audit information to predict tax evasion and estimate tax gaps. These audits can be categorized into two main types: random and operational audits. Both types serve different purposes and have unique methodologies that provide insights into taxpayer compliance.

Random audits

Random audit initiatives involve selecting taxpayer samples through a random process, aiming to accurately reflect the broader population. When conducting these audits, all selected taxpayers undergo a thorough examination to identify any discrepancies between what they reported on their taxes and what they are legally required to report. The findings from these audits provide a reliable measure of the overall level of compliance within the sample group. To extrapolate the sample results to the total population, we must ensure that the selection process is completely random and involves no selection criteria (Barra et al. 2023).

Random audits have downsides according to Feinstein (1999), including high costs for both tax offices and taxpayers, especially those who comply with tax laws. There is also a delay between the period the data covers and when the results are available. The financial returns are usually lower than those of targeted audits since they examine both compliant and non-compliant taxpayers, unlike targeted audits, which focus on those more likely to evade taxes. Additionally, they cannot detect unregistered taxpayers, leading to underestimations of some tax gaps.

Finally, revenue authorities might be reluctant to conduct random audits due to reasons related to the public image of authority for taxpayers. Random audits could be perceived as over-reach or unfair scrutiny by compliant taxpayers, leading to negative public sentiment and diminishing trust in the tax authority.

Operational audits

Operational audits are based on risk assessment and target specific taxpayers chosen according to criteria defined by tax authorities' risk analysis. These audits can focus on one or multiple types of taxes, and single or multiple periods. Consequently, this type of audit might not be representative of the entire population due to the selection criteria, as not all taxpayers have the chance of being selected as in a random audit. Tax administrations implement bottom-up gap estimation using non-random audit data with the help of techniques aimed at inferring the traits of the wider population from the unrepresentative sample.

3.3 Bottom-up estimation procedures

Several procedures can be used to perform a bottom-up approach. They all use audit information to predict behaviour in unaudited firms or periods. In this section, we review the most common estimations and highlight their main characteristics (pros and cons).

Regression techniques

Regression techniques are considered common in bottom-up literature and can adjust for selection bias. They can also help determine traits that can predict whether a taxpayer will be compliant and estimate the extent of non-compliance. These regression techniques include the Heckman approach and the propensity score matching approach.

Heckman approach. The Heckman approach addresses selection bias, which occurs during the operational audit process, leading to endogeneity in the subset of audited taxpayers. This method, founded on Heckman's (1979) work, involves a two-stage estimation process. The first stage identifies the likelihood of an observation being included in the sample, essentially calculating the probability of a taxpayer being selected for an audit, using a probit regression equation. The second stage focuses on estimating the variable of interest, which in this case is the amount recovered from the audit. This is done by considering explanatory variables and a specific regressor that adjusts for selection bias. This particular regressor, known as the inverse Mills ratio, is derived from the parameters estimated in the selection equation. The outcome equation is then calculated using OLS regression, incorporating a factor from the first stage's equation.

The FISCALIS Tax Gap Project Group (2018) points out that two important considerations must be taken into account when estimating the tax gap using the Heckman method. Firstly, the selection equation needs to be powerful in explaining the outcomes, since the method relies heavily on the equation's ability to predict non-compliance. Secondly, the equation must include at least one variable that influences the selection for audit but does not impact non-compliance itself. This helps avoid problems with inaccurate estimates due to multicollinearity. Essentially, for accurate tax gap estimation, it is necessary to have data on factors that lead to being audited, which are not directly related to the level of non-compliance, and, in practice, this exclusion restriction is hard to satisfy.

Propensity score matching approach. The propensity score matching method is used to correct for selection bias based on weights on the data. This method starts by calculating a 'propensity score' for each entity, using statistical models to predict their likelihood of being noncompliant or audited. A binary selection model computes the propensities using probit or logit. Once these scores are estimated, the method pairs entities that have been audited with those that have not but share similar propensity scores. The approach used to match observations could be nearest neighbor, caliper, kernel, or local linear. After matching, the final step is to assign a value to the unaudited returns. This value, referred to as N, is an imputed or estimated value of what the unaudited return would have reported if it had been audited. The imputation is based on the actual values observed in the matched audited returns. This step is necessary to estimate what the tax compliance would have been for the unaudited group if they had been subject to an audit.

Clustering approach

This approach categorizes both audited and non-audited taxpayers into clusters based on significant variables used to select the company for auditing, such as company size, geographical region, and industry sector. It allows for the calculation of the overall tax gap by summing the estimated gaps for each cluster. These estimates are derived by applying a scaling factor to the audit results of the audited taxpayers, thereby projecting these findings onto the broader population within each cluster. Although straightforward to apply and easy to implement, this method only partially corrects for selection bias, resulting in outcomes that are not entirely reliable.

Extreme values approach

The extreme value approach leverages the selection bias in operational auditing toward taxpayers with higher levels of non-compliance. It deals with the behaviour of extreme (maximum or minimum) values in a dataset, rather than the average values, suggesting that, regardless of the overall distribution of the data, the extreme values often follow a generalized Pareto distribution. This posits that insights into the overall rate of tax non-compliance among large corporations can be derived from a limited number of extreme cases (namely, the most significant tax evaders). This approach is applicable when the data exhibit characteristics of the Pareto distribution—a form of power-law distribution indicating that a small fraction of cases contribute disproportionately to the total value observed in the data, as when tax under-reporting is heavily skewed (with a few large corporations accounting for most of the gap) (Bloomquist et al. 2014).

Machine learning approaches

The application of machine learning (ML) approaches to economics studies, although quite new, is witnessing a gradual increase, particularly in taxation-related research, such as tax evasion, fraud, and compliance prediction, as well as improving tax audit and tax gap estimation. While the research in this area generally relies on traditional methods to make predictions, these methods suffer limitations related to the dependency on linear regression methods and the strict distribution assumptions they have. In reality, data often exhibit more complex patterns, which makes these methods not flexible enough for prediction. Therefore, some studies have started to adopt machine learning methods to improve the prediction outcomes.

As an illustration of using machine learning, Pérez López et al. (2019) employed multilayer perceptron (MLP) neural network models to predict tax fraud by using comprehensive Spanish PIT returns data. This ML method was able to predict tax fraud probability and the likelihood of involvement in fraud-related practices for each taxpayer. Zumaya et al. (2021) utilized two ML algorithms, including artificial neural networks (ANNs) and random forest (RF), in addition to complex network analysis to predict VAT evasion in Mexico by analyzing the transactional data and the interaction networks of taxpayers. The paper found that the combination of these three methods enabled the identification of new potential suspects by learning patterns from known evaders. Ioana-Florina and Mare (2021) tried to predict taxpayer's propensity to evade taxes based on their trust in the fiscal system using a multilayer perceptron (MLP) neural network model. This approach demonstrated superior predictive performance, surpassing that of the binary logistic regression model.[2]

On the other hand, machine learning methods are also used to enhance tax audit efforts. For instance, Howard et al. (2020) assessed the potential of machine learning techniques to enhance the selection process for correspondence audit cases by the Internal Revenue Service (IRS). The study discovered that for some audit categories, ML methods outperform traditional approaches in ranking and selecting tax returns for correspondence audits. These methods not only yield higher revenue but also lower the no-change ratio, meaning fewer audits result in no adjustments compared to other methods. Similarly, Battaglini et al. (2024) used Italian administrative tax data to explore the potential of machine learning techniques such as random forest in enhancing the discovery of tax evasion detection and recovery by improving the process of selecting taxpayers for audit. The paper indicates that in some scenarios, ML could improve the prediction of evasion detection by up to 83 per cent and recover up to 65 per cent of detected evasion.

[2] See also Alsadhan (2023); Baghdasaryan et al. (2022); Holtzblatt and Engler (2022); Murorunkwere et al. (2022, 2023); Raikov (2021); Savić et al. (2022) for other examples of using machine learning methods in predicting tax fraud and evasion behaviour.

The research on estimating the tax gaps was not far from these new developments. Given the limitations of previously mentioned tax gap estimation approaches that rely on traditional regression methods to make predictions, some researchers and tax revenue authorities started to incorporate the use of semi-parametric techniques into traditional methods and began to use machine learning to improve the prediction outcomes. While machine learning is superior in prediction tasks compared to traditional approaches, it is also effective in addressing selection bias for tax gap estimations that are based on operational audits.

To address the issue of selection bias in the context of tax gap estimations, it is crucial to distinguish between the two primary types of selection bias: causal and sample selection bias. Causal selection bias affects the estimation of unbiased parameters in causal analysis, such as when treatment and control groups are not randomly assigned, leading to biased estimations of treatment effects. However, our focus is on sample selection bias, which occurs when the training sample used to build a predictive model differs from the prediction sample. In the case of tax gap estimations based on operational audits, this bias arises because the training sample consists of audited taxpayers selected based on some known criteria from tax authorities and not representative of the whole taxpayer population, whereas the prediction sample includes unaudited taxpayers. This discrepancy can lead to biased predictions if not properly addressed.

A crucial aspect of handling sample selection bias is distinguishing between biases arising from observable versus unobservable factors. Observable selection bias occurs when the selection process, such as the decision to audit, is based on known and measurable variables. In such cases, if the probability of being audited can be accurately estimated using these observable covariates, the bias can be corrected by including these covariates in the machine learning model. This methodology corresponds with the strategies outlined by Brewer and Carlson (2024), who advocate for controlling selection bias by adjusting for observable factors. By calculating and integrating the probability of selection into the model, it is possible to mitigate the selection bias, presuming that the audit decisions are driven primarily by observable data.[3]

In scenarios where the selection process is governed by unobservable factors that are not captured in the dataset, the complexity of the bias increases. Traditional methods may not be sufficient to counteract this form of bias. In such cases, more advanced techniques are required to address the selection bias based on unobservables, such as incorporating a control function into the ML model based on Heckman's method (Brewer and Carlson 2024). In recent literature, there are notable examples of integrating machine learning approaches into traditional methods as well as studies that estimate tax gaps using mainly machine learning techniques.

Alaimo Di Loro et al. (2023) proposed a machine learning-based method consisting of two steps of the gradient boosting algorithm. This method addresses the selection bias stemming from the reliance on non-random audit data and provides accurate predictions. Firstly, the method estimates the propensity scores of the likelihood of a taxpayer being audited using a classification model based on gradient boosting with classification and regression trees (CART) as base learners. To do that, the data is divided into training and test sets, and during the training process, the important covariates are selected. This step will lead to having the predicted probabilities of each firm being audited based on their covariates. Secondly, the method employs a regression model using gradient boosting with CART as base learners to predict the potential tax base, including the undeclared VAT, hence the evaded amounts for each firm. In this step, the propensity scores previously obtained are used to create weights for each taxpayer,

[3] Presumably, tax authorities have information about how to decide who to audit. This information is usually reserved, yet it can be used in the machine learning model to accurately predict outcomes. Our advice is not to share the relevance of the covariates in the prediction since this information is related to the audit process. However, those results can also be used to improve the audit decision-making process.

correcting for any over- or under-representation in the audited sample. The comparison of this ML approach with the traditional Heckman model reveals the superiority of ML in capturing the variability in the potential tax base and providing more accurate predictions for the tax gap estimation.

Adu-Ababio et al. (2024) employed supervised machine learning algorithms with tax returns and audit data to estimate tax gaps in Zambia. The main machine learning algorithm utilized in this study was the artificial neural network (ANN) in two stages. The first stage relies only on the audited data to create iterations of multiple versions of training and testing datasets randomly, with 90 per cent of the data used for training the model. In every iteration, the algorithm learns from the training set by analysing various tax-related parameters. Then the algorithm uses what has been learned to predict tax evasion rates using testing data. Then, the algorithm compares the actual and predicted tax evasion rates. If these predictions do not match closely with the actual rates, some improvements could be made to the model. This process is repeated until reaching a satisfactory performance. For the second stage, the model is deployed using the full sample where the audited data is used in the training set and the unaudited data forms the testing set. Once the model learns from the selected explanatory variables, the model predicts the tax evasion from the testing data and then uses the predicted and actual tax evasion to estimate tax gaps. The authors also employed other machine learning algorithms, such as random forest, to verify the stability and reliability of the main method, and the results were relatively close.

Following the same line, the study by Ebrahim et al. (2024) used administration tax and audit data to estimate the VAT gap in Tanzania using machine learning, specifically the random forest algorithm. This approach aimed to predict tax evasion amounts for non-audited and audited firms in periods when no audit were carried out. The authors compared the performance of the ML approach with the traditional OLS regression. They found a significant reduction in root mean square error (RMSE) and higher R-square values when using the random forest algorithm, indicating higher accurate prediction performance. The results reveal a VAT gap averaging around 62 per cent, with considerable differences among various economic sectors. The agriculture sector, in particular, showed the largest VAT gap, highlighting significant tax evasion in this area.

Other advances in ML techniques involve the use of former regression approaches. Chudý et al. (2020) applied a semiparametric sample selection of the Heckman model to estimate Slovakia's corporate income tax (CIT) gap. This extension of the Heckman model outperforms the traditional Heckman model as it allows a more relaxed normality assumption and better modeling of the complex data structures and handling of non-linear relationships and heteroscedasticity inherent in the data. In the first stage of this model, the selection equation was estimated using a nonparametric method like kernel smoothing, providing flexible approximations of distributions. In the second stage of the model, the outcome equation incorporated these estimates from the first stage to provide a more robust correction for selection bias and capture the more complex relationships that a linear regression model might overlook. In addressing selection bias and providing better predictions, the paper found that this approach performed better compared to some other alternative approaches, like propensity score matching and weighted OLS linear regression.

Tax authorities have also started to use ML techniques to improve their estimations of tax gaps or auditing processes. The Italian Revenue Agency (n.d.) used machine learning along with other traditional methods to estimate the VAT gap in the machine-learning-assisted approach. The initial step of this approach aims to address selection bias that stems from using non-random audits by using logistic regression to divide the population into groups, with each group having a similar probability of being audited. Then, the population is stratified into quintiles based on these probabilities, which enables the audited taxpayers to be representative of the entire population in each group. In the second step, machine learning, specifically bagging regression trees, is employed to make predictions within each stratum. The last step is aimed at improving the prediction accuracy by using the predictive mean matching (PMM) model which uses the initial predictions to match each non-audited taxpayer (referred to as the recipient)

with an audited taxpayer (referred to as the donor) based on the similarity of their predicted values. This process ensures that the imputed values reflect the true distribution of the target variable, allowing for accurate inferences on various distributional characteristics beyond just the averages.

The Canada Revenue Agency (2019) employs an unsupervised machine learning technique to identify clusters within a population, similar to the first step mentioned previously for Italy, where elements in each cluster are more similar to each other than to those in other clusters. This machine learning algorithm automatically categorizes firms into clusters based on specific characteristics, assuming that unaudited firms in each cluster share the same non-compliance ratio to reported gross revenue as audited ones. This approach was used to deliver an upper-bound estimate and was combined with the extreme value approach to provide a lower-bound estimate of the tax gap.

Summary

Estimating tax gaps using a bottom-up approach could be achieved using different methods of estimation. The method used depends on the context and available data. Generally, using a bottom-up approach could be based on random or risk-based audit data. Many researchers see that relying on random audit data is the ideal way to make bottom-up tax gap estimation. However, in many cases, tax authorities tend to prefer doing a risk-based audit which introduces some challenges to the estimation given that the taxpayers selected for audit could differ significantly from other taxpayers. Risk-based audits do not represent the overall non-compliant population. In this case, traditional OLS estimation might not be the optimal choice for researchers because of the selection bias of the audit process. Therefore, researchers are using other methods to reach unbiased estimations. In the following, we summarize the key insights about the methods mentioned in this section.

While the two-stage Heckman approach is considered one of the most commonly used methods to account for selection bias, sometimes its exclusion restriction is difficult to satisfy. This might lead to inflated standard errors due to multicollinearity and a tendency to under-report the tax gap, as tax avoidance and undetected non-compliance are often overlooked. Propensity score matching helps eliminate selection bias by creating matched groups of compliant and non-compliant taxpayers based on observable characteristics, which allows for more accurate attribution of differences in tax compliance outcomes to non-compliance rather than unobserved factors. Some revenue authorities use the clustering approach to detect anomalous behaviours and uncover tax under-reporting within specific clusters, and then estimate the tax gap for each cluster by extrapolating the audit findings from the audited taxpayers to the entire population within that specific cluster. On the other hand, the extreme value approach is more straightforward and cost-efficient in terms of time and resource usage than other approaches. However, it requires more assumptions, especially concerning the setting of the tail in the Pareto distribution, which it relies on for modeling.

In contrast, estimation techniques based on machine learning offer significant advantages over the aforementioned methods, particularly when handling complex, non-linear relationships and unobserved factors influencing selection bias. Machine learning methods may be preferred for their flexibility and superior predictive performance.

4 The toolkit

In this section, we explain the toolkit's components. The goal of the toolkit is to estimate the tax gap on VAT (value-added tax), CIT (corporate income tax), or PIT (personal income tax). The toolkit has two main elements: data cleaning and estimation. The data cleaning process is to ensure the harmonization and consistency of required data files for the bottom-up estimation. Moreover, it helps to align the general requirements in the machine learning (ML) estimation. This is important since data comes from

different sources and periods, and standardizing it simplifies the estimation process. The estimation is based on the random forest methodology, a machine learning technique. A technical explanation of this is provided in Appendix A.

4.1 Data cleaning

The data cleaning process can be divided into three main stages: the first two deal with administrative return tax (VAT, CIT, and PIT) and audit data, and the last one demonstrates combining these data files for subsequent analysis. This step aims to process different data sources, harmonize them, and build a unique structure that combines information about taxpayers, tax declarations, and audit outcomes or assessments.

Usually, information about tax returns (VAT, CIT, or PIT) is contained in different files than from audits as the latter is conducted after firms or individuals file their returns. However, tax returns can contain at least two sets of filing for the same taxpayer. This may be due to the taxpayer updating the return at some point within or outside the filing period. This is a common issue of duplication that arises in tax administrative databases. In such instances, the same piece of information is replicated for the same taxpayer. In other words, for a taxpayer in a particular return year, there are two or more replications of the same information. One of the main aims of the data cleaning section is to ensure that each taxpayer is uniquely identified by their identifiers and the filing return year. In the first stage of the toolkit, we provide possible scenarios creating such duplication errors and demonstrate how the user can individually deal with them. It is important to address any duplicates in all required tax return and audit data files, irrespective of whether they come in single or multiple files. In the case of multiple files, the approach is to first deal with duplicates and then to append the respective data sets into a single file.

In this stage of the toolkit, we also address problems observed in audit data concerning audit periods and how they relate to specific return periods. In some cases, the audit data is identified by the assessment year doubling as the return year for the filing. At times it is rather the audit year that doubles as the return year. Whatever the case, it is necessary to identify the specific year in the audit data that corresponds to the return year and append them to obtain a single file if the data is in multiple files. This ensures that each audit assessment is correctly linked to a specific return period.

At the end of these first two steps, we aggregate the tax return data to the annual level. The aggregation usually happens for VAT and PIT but not for CIT, as it is always reported annually. As we estimate the tax gaps annually, we also ensure that audit assessments relate to annual outcomes even if audits were conducted for multiple return years. This procedure ensures that we have one tax return or audit outcome (if the taxpayer is audited) per taxpayer per year.

Finally, we combine the required data files bearing in mind that variables from tax returns and audits are in two separate files. It is important to understand the merging process as it shows how well we have cleaned and dealt with duplicates in all data files. The aim is to merge information for the same unit (taxpayer) in the same period (year-month). Moreover, we want the information provided by the audit data, such as the audit outcome for a particular return year, to be merged with the tax record in the corresponding return filing period. For example, we merge the tax record of the 2018 return year with the audit outcome about tax misreporting in 2018 if the firm was audited. There will be no audit outcome information if the firm is unaudited. This is a usual problem the user will face as the audits are conducted retroactively on a limited number of taxpayers based on past declarations. We explain how to obtain the audit outcomes for these unaudited firms in the next stage of the toolkit.

4.2 Machine learning estimation

The bottom-up approach is followed to estimate the tax gap. This approach requires as input the taxpayers' audit outcome, which we proxy as tax misreporting. This variable is obtained and found in the audit data after the audit process. As the variable is only available for audited firms, there is a necessity to estimate or predict it for unaudited taxpayers and periods. Predictions on unaudited taxpayers and periods are needed because the information about misreporting from audits is contingent on specific times and units. Hence, an audited taxpayer in return year 3 is unaudited in return year 2, meaning that we need to include a prediction for unaudited periods to ensure we have all the necessary information. An estimation procedure is necessary to obtain predictions about tax misreporting accurately.

In the toolkit, we follow the random forest method to predict tax misreporting in unaudited firms and periods. This methodology allows for granular estimation, better capturing potential outliers or atypical values than linear prediction. The random forest must be tuned by choosing two critical parameters: the number of iterations (or trees) and the number of uses to predict in each split. To do this, it is necessary to use data that contains the variables to predict tax misreporting. Therefore, first, the dataset is split into audit and unaudited data. The first will be used to tune the model, and the latter will be used to predict.

Splitting the audit data into training and testing samples is necessary in the tuning process. This is to improve the estimation accuracy since the methodology uses the training data to learn about the variables and later contrasts the prediction with the real value in the testing data. By doing this process, the two critical parameters are obtained. Besides that, the parameters ensure that the prediction error, in other words, the difference between the prediction and the real value, is the minimum possible. The toolkit runs a prediction with a regression model to compare to the ML predictions. This is to show the accuracy of the prediction and helps to validate the prediction model.

Finally, the tax gap is obtained. Firstly, the model is run only in the audited data since those observations have misreporting information. In this step, the model estimates the index (or weight) each auxiliary variable (or covariant) must have. Later, the model predicts the unaudited data using the optimal index, and the misreporting predictions are obtained. The tax gap is obtained by adding up the misreporting tax (predicted or discovered by audit) with the tax declaration, obtaining the potential revenue. The tax gap is the rate between misreporting and potential tax, indicating the percentage of the potential tax not collected due to misreporting. This variable is obtained by the type of group (such as industry), showing the granularity of the methodology.

5 Concluding remarks

In this tax gaps estimation bottom-up toolkit, we aimed to develop a practical framework for estimating tax gaps in value-added tax (VAT), corporate income tax (CIT), and personal income tax (PIT) using a bottom-up methodology. The toolkit is designed for tax authorities and policy-makers to estimate the difference between the actual tax revenue collected and the potential revenue that could have been gathered under full compliance with tax regulations. It provides a standardized setting applicable to developing countries given their context and available resources. Our approach is based on applying a machine learning algorithm using administrative micro-tax returns and audit data to predict tax misreporting and non-compliance and then estimating the tax gaps at both the aggregate level and by specific sectors or regions.

In this note, we went through the definitions of tax gaps to provide an understanding of its components. The method target is to estimate the under-reporting, misreporting, and non-compliant tax gaps among registered taxpayers. Then, we go through the traditionally employed procedures pointing out

the advantages of using machine learning estimation using a bottom-up approach over other alternative estimations.

The toolkit can be divided into two major stages: data management and machine learning analysis. Within the data management stage, the tax and audit datasets are prepared for an analysis procedure through cleaning, handling of duplications, and merging to ensure harmonization of tax and audit data, allowing a seamless move to the stages of machine learning. Machine learning predicts tax misreporting for taxpayers or periods not covered by audits through the application of random forest algorithms. Such models have the capability of providing a proper estimation since they are trained on audit data and can accurately estimate tax evasion for unaudited cases. This enables estimations of the tax gap in a comprehensive manner. In comparison to traditional regression models, machine learning models out perform OLS estimations and improves predictive power.

Finally, some suggestions for future work involve expanding and improving the current toolkit in the following several ways. It would be possible to use other machine learning algorithms, such as neural networks, and compare the accuracy of prediction across different methods. Generalization of the toolkit into other programming languages besides STATA, thus expanding its reach, is yet another area that would be considered. There is a need for further work regarding the way the toolkit could be implemented across different national contexts. Finally, the toolkit may itself provide a point of departure in future research on taxpayer behaviour with respect to assisting authorities in the design and implementation of better enforcement strategies and compliance measures.

References

Adu-Ababio, K., Koivisto, A., and Mwale, E. (2024). *Estimating tax gaps in Zambia.* (Preprint)

Alaimo Di Loro, P., Scacciatelli, D., and Tagliaferri, G. (2023). '2-step Gradient Boosting approach to selectivity bias correction in tax audit: an application to the VAT gap in Italy'. *Statistical Methods & Applications*, 32(1): 237–270. https://doi.org/10.1007/s10260-022-00643-4

Alsadhan, N. (2023). 'A Multi-Module Machine Learning Approach to Detect Tax Fraud'. *Computer Systems Science and Engineering*, 46(1): 241–253. https://doi.org/10.32604/csse.2023.033375

Athey, S., and Imbens, G. W. (2019). 'Machine learning methods that economists should know about'. *Annual Review of Economics*, 11(1): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Baghdasaryan, V., Davtyan, H., Sarikyan, A., and Navasardyan, Z. (2022). 'Improving tax audit efficiency using machine learning: The role of taxpayer's network data in fraud detection'. *Applied Artificial Intelligence*, 36(1): 2012002. https://doi.org/10.1080/08839514.2021.2012002

Barra, P. A., Hutton, M. E., and Prokof'yeva, P. (2023). *Corporate Income Tax Gap Estimation by using Bottom-Up Techniques in Selected Countries: Revenue Administration Gap Analysis Program*. Washington, DC: International Monetary Fund. https://doi.org/10.5089/9798400246265.005

Battaglini, M., Guiso, L., Lacava, C., Miller, D. L., and Patacchini, E. (2024). 'Refining public policies with machine learning: The case of tax auditing'. *Journal of Econometrics*, September(-): 105847. https://doi.org/10.1016/j.jeconom.2024.105847

Békés, G., and Kézdi, G. (2021). 'Regression Trees'. In *Data analysis for business, economics, and policy* (pp. 417–437). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108591102.015

Bloomquist, K. M., Hamilton, S., and Pope, J. (2014). 'Estimating Corporation Income Tax Under-Reporting Using Extreme Values from Operational Audit Data'. *Fiscal Studies*, 35(4): 401–419. https://doi.org/10.1111/j.1475-5890.2014.12036.x

Brewer, D., and Carlson, A. (2024). 'Addressing sample selection bias for machine learning methods'. *Journal of Applied Econometrics*, 39(3): 383–400. https://doi.org/10.1002/jae.3029

Canada Revenue Agency (2019). *Tax gap and compliance results for the federal corporate income tax system.*

Chudý, M., Gábik, R., Bukovina, J., and Šrámková, L. (2020). *Searching for gaps: Bottom-up approach for Slovakia.* Institute for Financial Policy (IFP).

Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). 'Random forests'. In C. Zhang and Y. Ma (eds), *Ensemble machine learning: Methods and applications* (pp. 157–175). New York: Springer. ttps://doi.org/10.1007/978-1-4419-9326-7_5

Durán-Cabré, J. M., Esteller Moré, A., Mas-Montserrat, M., and Salvadori, L. (2019). 'The tax gap as a public management instrument: application to wealth taxes'. *Applied Economic Analysis*, 27(81): 207–225. https://doi.org/10.1108/AEA-09-2019-0028

Ebrahim, A., Castillo, S., Leyaro, V., Swema, E., and Haule, O. (2024). *Estimating the Value-Added Tax Gap for SMMEs in Tanzania: An Empirical Analysis.* (Manuscript)

Feinstein, J. S. (1999). 'Approaches for estimating noncompliance: examples from federal taxation in the United States'. *The Economic Journal*, 109(456): 360–369. https://doi.org/10.1111/1468-0297.00439

FISCALIS Tax Gap Project Group (2018). 'The Concept of Tax Gaps: Corporate Income Tax Gap Estimation Methodologies'. Working paper 73 – 2018. Luxembourg: Publications Office of the European Union. (European Commission's Directorate-General Taxation and Customs Union) https://doi.org/10.2778/83206

Gemmell, N., and Hasseldine, J. (2014). 'Taxpayers' behavioural responses and measures of tax compliance 'gaps': A critique and a new measure'. *Fiscal Studies*, 35(3): 275–296. https://doi.org/10.1111/j.1475-5890.2014.12031.x

Hartshorn, S. (2016). *Machine learning with random forests and decision trees: A Visual guide for beginners*. Kindle edition.

Heckman, J. J. (1979). 'Sample selection bias as a specification error'. *Econometrica*, 47(1): 153–161. https://doi.org/10.2307/1912352

Holtzblatt, J., and Engler, A. (2022). *Machine Learning and Tax Enforcement.* Tax Policy Center, Urban Institute & Brookings Institution.

Howard, B., Lykke, L., Pinski, D., and Plumley, A. (2020). 'Can Machine Learning Improve Correspondence Audit Case Selection? Considerations for Algorithm Selection, Validation, and Experimentation'. In A. Plumley (ed.), *The IRS Research Bulletin: Proceedings of the 2020 IRS / TPC Research Conference* (pp. 147–169). Internal Revenue Service.

Hutton, M. E. (2017). *The Revenue Administration–Gap Analysis Program: Model and Methodology for Value-Added Tax Gap Estimation*. International Monetary Fund. https://doi.org/10.5089/97814755836618.005

Ioana-Florina, C., and Mare, C. (2021). 'The utility of neural model in predicting tax avoidance behavior'. In I. Czarnowski, R. Howlett, and L. Jain (eds), *Intelligent Decision Technologies: Proceedings of the 13th KES-IDT 2021 Conference* (pp. 71–81). https://doi.org/10.1007/978-981-16-2765-1_6

Italian Revenue Agency (n.d.). *Italy: VAT gap estimation via bottom up approach.*

Murorunkwere, B. F., Haughton, D., Nzabanita, J., Kipkogei, F., and Kabano, I. (2023). 'Predicting tax fraud using supervised machine learning approach'. *African Journal of Science, Technology, Innovation and Development*, 15(6): 731–742. https://doi.org/10.1080/20421338.2023.2187930

Murorunkwere, B. F., Tuyishimire, O., Haughton, D., and Nzabanita, J. (2022). 'Fraud detection using neural networks: A case study of income tax'. *Future Internet*, 14(6): 168. https://doi.org/10.3390/fi14060168

Pérez López, C., Delgado Rodríguez, M. J., and de Lucas Santos, S. (2019). 'Tax fraud detection through neural networks: An application using a sample of personal income taxpayers'. *Future Internet*, 11(4): 86. https://doi.org/10.3390/fi11040086

Plumley, A. (2005). 'Preliminary update of the tax year 2001 individual income tax underreporting gap estimates'. In *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association* (Vol. 98, pp. 19–25).

Raikov, A. (2021). 'Decreasing tax evasion by artificial intelligence'. *IFAC-PapersOnLine*, 54(13): 172–177.

Savić, M., Atanasijević, J., Jakovetić, D., and Krejić, N. (2022). 'Tax evasion risk management using a Hybrid Unsupervised Outlier Detection method'. *Expert Systems with Applications*, 193(May): 116409. https://doi.org/10.1016/j.eswa.2021.116409

Schonlau, M., and Zou, R. Y. (2020). 'The random forest algorithm for statistical learning'. *The Stata Journal*, 20(1): 3–29. https://doi.org/10.1177/1536867X20909688

Varian, H. R. (2014). 'Big data: New tricks for econometrics'. *Journal of Economic Perspectives*, 28(2): 3–28.

Warren, N. (2018, April). 'Estimating Tax Gap is Everything to an Informed Response to the Digital Era'. In *13th International Revenue Administration Conference on Tax System Integrity in a Digital Age* (p. 1-41). Available at: https://ssrn.com/abstract=3200838 (last revised: June 23, 2019)

Zacharis, N. Z. (2018). 'Classification and regression trees (CART) for predictive modeling in blended learning'. *IJ Intelligent Systems and Applications*, 3(1): 9. https://doi.org/10.5815/ijisa.2018.03.01

Zumaya, M., Guerrero, R., Islas, E., Pineda, O., Gershenson, C., Iñiguez, G., and Pineda, C. (2021). 'Identifying tax evasion in Mexico with tools from network science and machine learning'. In O. Granados and J. Nicolás-Carlock (eds), *Corruption networks: Concepts and applications* (pp. 89–113). Cham: Springer. https://doi.org/10.1007/978-3-030-81484-7_6

A Appendix: Random forest algorithm

Random forest is considered one of the most widely used and best-performing ensemble machine learning algorithms for prediction tasks (Athey and Imbens 2019).[4] Unlike traditional regression models, which assume linearity and struggle when the number of observations is less than the independent variables, random forest can handle nonlinear relationships in the data and avoids the problem of estimating more parameters than the data points can support. Besides, it captures better the existence of outliers and atypical values, producing more accurate predictions in such cases (Athey and Imbens 2019). It achieves this by not using all predictor variables at the same time, resulting in better predictions than traditional regression (Schonlau and Zou 2020). Apart from being simple to use, random forest is straightforward to understand and quick to implement. Additionally, it performs well when compared to other machine learning algorithms (Varian 2014).

In essence, random forest can enable us to predict the target variable (y) using input variables (x). It is essentially a collection of decision trees created using random subsets of data. But what are decision trees, and how are they used to create a random forest model? To answer this, the document starts by explaining the concepts of decision trees and how they work, and then moves on to explain how we can build a random forest model and use it to accomplish prediction tasks.

A1 Decision trees

Decision trees are a type of supervised learning algorithm used for both regression and classification tasks. They work by splitting the data into subsets based on the values of input variables (x) to predict values (y). This splitting process continues until the data within each subset are as homogeneous as possible with respect to the target variable. It is also known as the classification and regression trees (CART) algorithm, which is a way of finding the best split at each step to maximize prediction accuracy.

CART algorithm

CART types:

- **Classification trees** are a type of decision tree algorithm used for classifying categorical target variables. They work by segmenting the predictor space into distinct regions, with each region corresponding to a specific class label. The goal is to determine which category the target variable belongs to based on the input features.

- **Regression trees** are a type of decision tree algorithm designed to predict continuous target variables. They divide the predictor space into regions and provide a continuous value as the output for each region.

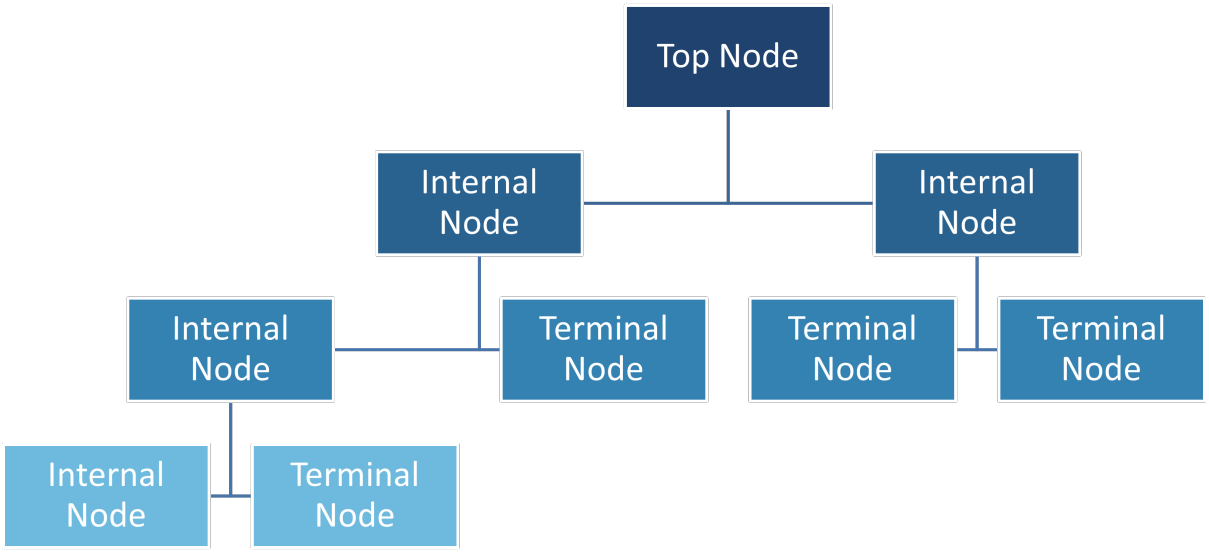
How does the CART algorithm work?

The structure of building a decision tree using CART begins with the **top node**, which represents the entire dataset. This top node is the starting point of the tree. From this point, the algorithm identifies the best attribute to split the dataset and labels the node accordingly. This creates branches that lead to **internal nodes**, where each internal node represents a decision based on the value of the chosen attribute. The data is further split at each internal node, continuing to generate more branches and nodes. This process repeats, creating a hierarchical structure. The endpoints of these branches are the **terminal nodes** which provide the final prediction, as shown in Figure A1. In classification tasks, the

[4] An ensemble method combines multiple simple models, known as weak learners, to create a single, stronger predictive model.

prediction at a terminal node is the majority class of the observations in that node, and in regression tasks, it is the average value of the observations.

Figure A1: Decision tree structure



Source: authors' illustration.

In regression tasks, CART utilizes residual reduction as its splitting criterion. This involves partitioning the data at each node to minimize the average squared difference between the predicted and actual values, aiming to achieve the lowest residual error. For classification tasks, CART uses Gini impurity to assess all potential splits, choosing the one that most effectively reduces impurity and thus increases the purity of the resulting subsets. Gini impurity quantifies the likelihood of misclassifying a random instance based on the majority class within a subset. This splitting process is recursive, continuing until certain stopping criteria are met. These criteria include reaching a node where all records share the same target value, the node size being below a user-defined threshold, the tree achieving its predefined maximum depth, having fewer than a minimum number of cases in a node, or when further splitting does not significantly enhance purity (Zacharis 2018).

An essential risk when using decision trees is model overfitting. This can happen if the model grows without constraints, such as when a regression tree continues to split until each terminal node contains only a single observation. While this may result in an almost perfect fit to the training set (see Note for the definition), it negatively impacts the model's ability to generalize to new, unseen data. Overfitted models typically perform well on training data but poorly on validation or test data because they have learned the noise rather than the signal.

To address overfitting issues, CART utilizes a pruning technique once the tree has been fully grown. Pruning involves cutting back the tree to eliminate nodes that add minimal predictive value, thereby simplifying the model and improving its generalization. A widely used technique is cost complexity pruning, where a large tree is initially grown using a very small complexity parameter to ensure all potential splits are evaluated. Then, splits are removed sequentially, and the model's performance is re-evaluated using cross-validation. This process continues until further pruning does not enhance the model's fit (Békés and Kézdi 2021).

Figure A2 presents a pseudocode example for a tree-growing algorithm that explains classification and regression tasks. Let's consider a scenario where we aim to construct a decision tree for predicting a target variable using a dataset X, which contains multiple covariates A, and the target variable y. The task parameter indicates whether we are dealing with classification or regression.

The algorithm starts by initializing a single tree T with a top node. If all stopping criteria have been met the algorithm proceeds to label the node. For classification tasks, the node is labeled with the most common class among the samples in X. For regression tasks, the node is labeled with the mean value of y.

If the stopping criteria have not been met, the algorithm searches for the best attribute $a \in A$ that splits the dataset X most effectively. Classification tasks are done using an impurity function such as Gini impurity. For regression tasks, the algorithm aims to minimize the variance within the nodes. The node is then labeled with the attribute a.

Figure A2: Tree growing algorithm pseudocode for both classification and regression tasks

Algorithm 1 Tree growing algorithm `growingtree`(X, A, y, task)

Require: Training dataset X, attribute set A, output variable y, task (classification or regression)
Ensure: Decision tree

Begin a single tree T with a top node
if all stopping criteria have been met **then**
    if task == classification **then**
        T has one node with the most common class in X as label
    else
        T has one node with the mean of y in X as label
    end if
else
    find $a \in A$, that best splits X using impurity function (for classification) or minimizing variance (for regression)
    Label node with a
    for possible value v of a **do**
        $X_v =$ the subset of X that have $a = v$
        $A_v = A - a$
        `growingtree`$(X_v, A_v, y, \text{task})$
        connect the new node to the top node with label v
    end for
end if
return `pruningtree`(X, A, y, task)

Note

In machine learning, we split the data into two main subsets:

Training set: This subset is used to build models such as regression trees and random forests. It includes input features (independent variables) and the target variable (dependent variable). The model learns patterns and relationships from this data.
Testing set: This subset is used to evaluate the model's performance. The testing set is not seen by the model during the training phase, allowing for an unbiased assessment of how well the model generalizes to new, unseen data.

Next, the algorithm iterates over all possible values v of the chosen attribute a. For each value v, it creates a subset of X where the attribute a takes on the value v. It also updates the attribute set A by removing the attribute a. The algorithm then recursively calls itself to grow the tree further, using the subset of X and the updated attribute set A. This recursive process continues, connecting new nodes to the top node with labels corresponding to the values v.

Once the tree has grown to its full extent based on the initial criteria, the algorithm proceeds to prune the tree. The pruning process involves using a separate pruning function that evaluates whether removing certain nodes and branches improves the tree's performance on a testing dataset. This is done using cross-validation techniques to ensure the tree generalizes well to unseen data.

By iterating this process, the tree-growing algorithm constructs a decision tree that partitions the dataset X into increasingly smaller regions. Each terminal node of the tree corresponds to a specific region in the feature space. In classification tasks, the leaf node represents the majority class within that region, while in regression tasks, it represents the average value of y.

A2 Random forest

Decision trees, although useful, have notable limitations, particularly their tendency to overfit data despite pruning. In real-world scenarios, data can be messy and contain anomalies that do not generalize well. Decision trees might create very specific splits that fit the training data but fail to perform accurately on new, unseen data. Random forests address this issue by utilizing multiple decision trees and averaging their results. Simply generating multiple trees from the same dataset does not solve the problem, as it would produce similar trees. Instead, random forests create trees using random subsets of the data. This process of using varied subsets ensures that the trees are different, which helps to smooth out anomalies and improve overall prediction accuracy by combining the diverse trees into a more robust model.

Bootstrap aggregation and selection criteria

In random forests, randomness is introduced in two primary ways. First, by selecting a random subset of data for each tree, and second, by choosing a random subset of predictor variables for each split in the tree. Each tree in a random forest is constructed using a technique called **bootstrap aggregation, or bagging**. The bagging algorithm works by first taking multiple random samples from the original dataset. Let's say we take B samples, where B is a large number, usually in the hundreds. For each sample, a large decision tree is created without any simplification. These trees are then used to make predictions. The algorithm creates B prediction rules from these trees and combines them. In a setup where we test the model's accuracy, B predictions are made for each data point based on the results from each of the B trees. The final step is to average these B predictions to get the final predicted value.

Random forests also introduce randomness by limiting the features considered at each split. Rather than evaluating all predictor variables (x variables) at each branching point, the algorithm randomly selects only a subset of these variables for consideration. The size of this subset is usually predetermined, often around the square root of the total number of predictors, with a common minimum set at 4. This approach is applied to each bootstrap sample, resulting in the construction of B trees. The final prediction is made by averaging the outputs from these B trees.

The rationale behind using a limited number of predictor variables at each split is to minimize the likelihood of all trees becoming too similar, especially if one strong predictor is dominant. By restricting the set of variables at each decision point, the algorithm allows a more balanced contribution from all predictors, including weaker ones that might provide valuable information when considered together.

Without this random selection, the resulting trees would heavily favor the strongest predictors, leading to highly correlated and less diverse predictions.

Tuning the model

When running a random forest, there are several key tuning parameters to consider to ensure optimal model performance. The primary parameters include the number of trees, the number of predictors evaluated at each split, and the stopping rule for tree growth.

- **Number of trees (B):**
  - This parameter controls how many bootstrap samples are used to construct the forest. More trees generally increase model accuracy but also computational time.
- **Number of predictors per split (x):**
  - At each node, only a subset of predictors is selected for splitting. A good rule is to use the square root of the total number of predictors. For example, with 64 predictors, use around eight for each split. At least four predictors should be used.
- **Stopping rule for tree growth:**
  - Determines when to stop splitting nodes in a tree. A simple rule is to set a minimum number of observations per terminal node. Commonly, five to 20 observations are used.

Then the method looks at the combination of these three tuning parameters that produces the smallest 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.

Another important metric is the **out of bag** error abbreviated as **OOB**. This metric estimates the performance of the model. When constructing each tree in the forest, the algorithm randomly samples approximately 63.2 per cent of the data, leaving the remaining 36.8 per cent as unused or 'out of bag'. This out-of-bag data is not utilized in the construction of a particular tree, but it can be used to estimate the accuracy of that tree by testing how well the tree predicts the OOB data. Averaging these OOB errors across all trees in the forest provides a reliable estimate of the model's performance, known as the OOB error rate. This technique ensures that all data points are evaluated in the model's performance assessment, thereby offering a robust measure of accuracy without needing a separate test set (Hartshorn 2016).

Variable importance

In random forest, understanding the importance of each predictor variable is essential for interpreting the model and refining its predictive accuracy. The method uses a direct method known as permutation importance, which assesses variable importance by observing changes in prediction accuracy when the values of each predictor are randomly shuffled. The model's prediction performance is then compared using both the original and permuted values of the variable, specifically utilizing out-of-bag (OOB) data. The permutation importance is calculated by measuring the increase in prediction error—such as the mean squared error (MSE) for regression tasks or the error rate for classification tasks—when a variable's values are permuted in the OOB data. A significant increase in error indicates the variable's importance. This technique not only identifies key predictors but also captures complex interactions between variables. Since the random forest algorithm selects random subsets of predictors for each split, the algorithm can identify all correlated predictors as important if any one of them contributes significantly to the outcome (Cutler et al. 2012).

A3 Example

This section develops a simple example to clarify the characteristics of the random forest. For this purpose, we will focus on developing the model and the prediction, explaining each step but not providing empirical examples.

Let us consider a population of 100 taxpayers. Each taxpayer completes a tax declaration that includes a tax declared (the amount subject to taxation) along with complementary information. This complementary information consists of ten variables, which may include details such as employee salaries and production costs. Although these variables are not directly included in the tax calculation, they provide valuable insights for determining the appropriate tax base level.

Among the 100 taxpayers, 50 were audited. This means we have information about potential discrepancies between the tax declared and the actual amounts for these 50 taxpayers. For example, if all 50 audited taxpayers were found to have evaded taxes, the audits would allow us to gather both the misreported amounts and compare this to the declared tax.

The first step is to recognize that we only have information about misreporting for these 50 audited taxpayers. Therefore, we can only evaluate the accuracy of our prediction model using this subsample. This is why we will partition the sample and concentrate on the audited taxpayers.

The audited taxpayer sample is divided into two groups: a training sample of 25 taxpayers and a testing sample of 25 taxpayers. We will use the training sample to build the random forest model and utilize the testing sample for tuning the model. Our focus will be on two critical parameters: the number of iterations (or trees) and the number of predictors considered at each split. The model aims to estimate the amount of misreporting based on the ten additional variables provided by the taxpayers.

We will use all available variables for two primary reasons. First, these variables help to accurately characterize the tax declared and are relevant in determining its level. Second, as these variables are available, they are crucial for deciding which taxpayers will be audited. Including all variables helps avoid sample selection bias due to observable factors.

To determine the optimal number of trees to use in our model, we will consider the number of predictors included in each split (the variables used to estimate misreporting). For simplicity, we will assume that we are using one of the ten available variables. To decide on the number of trees, we will run the model on the training sample and evaluate its performance on the testing sample using various numbers of trees. Specifically, we will conduct N separate runs of the random forest model, each time varying the number of trees used. During each run, we will make predictions on the testing sample and compare these predicted values to the actual misreporting identified through the auditing process. This will provide us with N root mean square error (RMSE) values, one for each model run. After all runs, we will select the minimum RMSE value and identify the associated number of trees, which we will denote as B. This number of trees is considered optimal because it minimizes the prediction error measured by RMSE, resulting in the most accurate estimate of the misreporting tax base.

Now, we move to estimating the predictor used in each split. In this case, we use the optimal number of trees, B, which was obtained before. We repeat the same iterating procedure, but in this case, we run ten different random forest models, obtaining the prediction in the training sample for each one and comparing it with the real misreported value. We run ten models because we have ten variables to use. This is because the total number of variables is the maximum number of predictors for each split. Importantly, if you have ten variables but decide to use eight for the prediction model, you must run eight models. The number of models to run in this step must always be equal to the variables decided to use in the prediction model. Finally, we repeat the process, choosing the minimum RMSE and the number of predictors used, x. This number of predictor x is optimal to minimize the prediction error.

With these two steps, we find the optimal number of trees (B) and predictor per split (x) to use in the random forest. Recall that to estimate them, we use the 50 audited taxpayers, splitting this sample into training and testing sets. Now, we can make the predictions for the 50 taxpayers that were not audited. The procedure is as follows. First, run the random forest with the optimal parameters on the set of 50 audited taxpayers. Later, predict the values in the set of 50 unaudited taxpayers. Finally, you can create a variable composed of the discovered misreporting for the 50 audited taxpayers and the predicted misreporting for the 50 unaudited taxpayers.