TALLINN UNIVERSITY OF TECHNOLOGY

School of Business and Governance

Alice Mikk

FIRM-LEVEL PREDICTORS OF LABOUR TAX EVASION

Master's thesis

Program Economic Analysis

Supervisor: Karsten Staehr, PhD

Tallinn 2024

I hereby declare that I have compiled the thesis independently
and all works, important standpoints and data by other authors
have been properly referenced and the same paper
has not been previously presented for grading.

The document length is 14 889 words from the introduction to the end of conclusion.

Alice Mikk 07.05.2024

TABLE OF CONTENTS

ABSTRACT

The objective of this master's study is to analyse the relationship between different firm-level financial and non-financial indicators and the probability of a firm being engaged in labour tax evasion. Numerous studies have investigated tax evasion and the shadow economy in Estonia, however, little attention has been paid to the connection between labour tax evasion and patterns in financial reports. Considering that labour tax evasion is a difficult challenge for governments due to which a large share of revenue is lost, it is important to investigate it further.

The thesis aims to answer which firm-level predictors contribute to the probability of being engaged in labour tax evasion, as well as estimating the proportion of labour tax evading firms. The methodological approach uses Mincer wage regressions to find the firms paying "suspiciously low wages" to employees and logistic regression to analyse the relationship between different firm-level indicators and the probability of the firm being a tax evader. Administrative matched employer-employee wage data, population data and firms' annual reports are used for the purpose of this thesis.

The results show that smaller companies (in terms of number of employees) are more likely to evade labour taxes. What is more, labour tax evasion is more prevalent in the construction sector. The results regarding financial ratios reveal that the probability to evade labour taxes decreases as turnover, debt to assets, or cost of goods sold to assets increases. However, the predicted probability to evade labour taxes increases as short-term debt to assets or turnover to assets increases. The prediction of out-of-sample probability of being engaged in labour tax evasion suggested that 51% of companies in 2021 and 54% in 2022 are classified as labour tax evading.

Keywords: tax evasion, wage regression, logistic regression.

INTRODUCTION

Tax evasion is a key problem for many governments as taxation is one of the primary sources of government revenue. The importance of efficient and fair tax system has been discussed by Schumpeter (1991), Musgrave (1959), Levi (1988), and Brennan & Buchanan (1980). Any inefficiencies in the design of tax system and the administration, or unfair treatment of taxpayers can result in a decision to shift to shadow economy. Revenue lost due to tax evasion affects the ability of the government to fund essential public services and infrastructure such as education, healthcare, defence and more. This can lead to budget deficits, increased debt or governments may respond by increasing tax rates to compensate the losses. However, higher tax burden can lead to deadweight losses as higher tax rates change the behaviour of compliant taxpayers who perceive the tax burden distribution as unfair.

Estonian government receives 80-90% of the revenues from taxation, and labour taxes account for approximately 50% of the tax revenues, being therefore directly affected by compliance or resistance of the individuals and companies to report wages to the tax authorities (Statistics Estonia, table RR057). Therefore, investigating labour tax evasion is crucial for understanding its economic, social, and ethical implications of the issue. Uncovering determinants and patterns of labour tax evasion can inform effective policy measures, which may lead to improved government revenue, more accurate fiscal planning and resource allocation as well as ensure a fair tax burden for participants of the economy. In addition, it may aid to design targeted interventions to promote corporate transparency and ethical financial practices.

Numerous studies have investigated the dynamics of tax evasion and also estimated the extent of the shadow economy (Putniņš & Sauka, 2015; Kukk & Staehr, 2014; Schneider, 2016; Tafenau *et al.,* 2010). What is more, different approaches to detect financial statement fraud have been provided and tied together with labour tax evasion as it constitutes a form of financial statement manipulation resulting in specific patterns in the balance sheet and profit and loss statement (Cecchini *et al.,* 2010; Hajek & Henriques, 2017; Gavoille & Zasova, 2023; Benkovskis &

Fadejeva, 2022). Applying the novel approach on Estonian data could therefore provide exciting insights into the labour tax evasion in Estonian firms.

The aim of this thesis is to analyse the relationship between different firm-level financial and non-financial indicators and the probability of a firm being engaged in labour tax evasion. Two research questions covered in thesis are as follows:

1) Which firm-level predictors contribute to the probability of being engaged in labour tax evasion?
2) What is the proportion of labour tax evading companies?

To answer the research questions raised, a database comprising three merged datasets is utilized, including administrative data on wages, population as well as annual reports. The analysis employs Mincer wage regression to find the firms paying "suspiciously low wages" to employees and logistic regression to analyse the relationship between different firm-level indicators and the probability of being a tax evader.

The master's thesis is structured into three main parts. The first chapter gives an overview of theoretical and empirical background of tax evasion. This chapter explains the institutional context of tax system, its importance and more precisely, taxation in Estonia. What is more, emphasis is placed on explaining the concept of shadow economy and connections with tax evasion as well as the dynamics of tax evasion. Lastly, an overview of tax evasion prevention and detection opportunities is given. The second chapter presents the data and the methodology employed in this thesis to analyse the different company-level indicators and the relationship with the probability of being a tax evader. The third chapter gives an overview of the findings, providing an explanation of the results as well as discussion on weaknesses of the analysis and suggestions for future research.

1. THEORETICAL AND EMPIRICAL BACKGROUND

The following chapter presents the theoretical and empirical background of taxation, shadow economy, tax evasion and tax fraud detection. Section 1.1. explains the fundamentals of taxation, gives an overview of the Estonian tax system and more precisely, labour taxation. Section 1.2. aims to describe the concept of the shadow economy and its connection with tax evasion, focusing on labour tax evasion. Section 1.3 discusses the motivation behind the decision to be compliant or to evade on the firm or individual level. Section 1.4. explains the potential approaches and tools to be used for identifying tax fraud on company level as well as characterizes the firms engaged in fraudulent activities based on previous research in the field.

1.1. Institutional background

The tax system plays a relevant role in shaping the economic landscape of a country. The importance of an effective tax system cannot be overstated as it serves as a primary source of government revenue funding essential public services and infrastructure such as education, healthcare, defence and more. In addition to funding the operation of public institutions, the tax system is a tool to redistribute wealth. Taxation directly impacts individuals and businesses in their decisions on establishing business, conducting trade, employment, investments and more.

The tax system being one of the cornerstones of all political regimes has already been recognized by Schumpeter (1991), Musgrave (1959), Levi (1988), Brennan & Buchanan (1980) and others. Schumpeter (1991) viewed taxation as a tool that could, in addition to generating revenue for the government, either support or hinder economic development and emphasized the importance of the government as a designer of tax policies. Musgrave (1959) proposed three main functions of taxation to be revenue collection, redistribution and macroeconomic stabilization. He also stressed the importance of equity, efficiency and feasibility of the administrative side of policies. Levi (1988) contributed by showing that taxation was not only a means of generating revenue but also a mechanism through which states assert their power, authority and legitimacy over a territory or a population. He also discussed the factors that influenced tax compliance and resistance among taxpayers. Brennan & Buchanan (1980) presented taxation from a public choice perspective, viewing tax policies as a social contract between citizens and the state as well as elected officials making decisions about taxation as an extension of the hand of the public. Just as Musgrave (1959), Brennan & Buchanan (1980) examined the trade-off between equity and efficiency in tax policy as the desire is to raise revenue for public goods and services but keep the burden fair for taxpayers.

A study published by the World Bank found that countries with simpler and more efficient tax systems had higher rates of economic growth and businesses were more eager to make investments and create new jobs, compared to countries with more complicated tax systems (Dom *et al.*, 2022). Kenny & Winer (2006) also discussed that democracy is followed by higher cooperation with tax authorities, thus the government receives higher tax revenues in case substantial degree of voluntary tax compliance is required (such as self-reporting).

Estonia has ranked first in OECD countries in the International Tax Competitiveness Index Ratings published by Tax Foundation for the last 10 years (Mendgen, 2023). The four main considerations have been 20% tax rate on corporate income applying only to distributed profits, flat 20% tax on individual income, property tax covering only the value of the land, rather than the value of property or capital and territorial tax system that exempts foreign profits earned by domestic firms from domestic taxation in full (with few exemptions) (*Ibid.*).

The Estonian tax system consists of national taxes established by tax laws and local taxes established by the city or municipality council in its administrative territory. State taxes include income tax, social tax, land tax, gambling tax, sales tax, customs duty, excise duty, heavy truck tax and business income tax. Local taxes include advertising tax, road and street closure tax, motor vehicle tax, livestock tax, entertainment tax and parking fees. (Rahandusministeerium, 2024)

According to Statistics Estonia (Statistikaamet, hereinafter SA) data on quarterly consolidated revenue and expenditure of general government, the three main contibutors to the revenues of the Estonian state budget are taxes on production and imports (mainly Value Added Tax, hereinafter VAT), social contributions (mainly employers' actual social contributions) and current taxes on income, wealth etc. (mainly personal income tax, hereinafter PIT) (Statistics Estonia, table RR057). Figure 1 presents the distribution of government revenue from abovementioned sources, illustrating the importance of revenues from different taxes for the last decade.

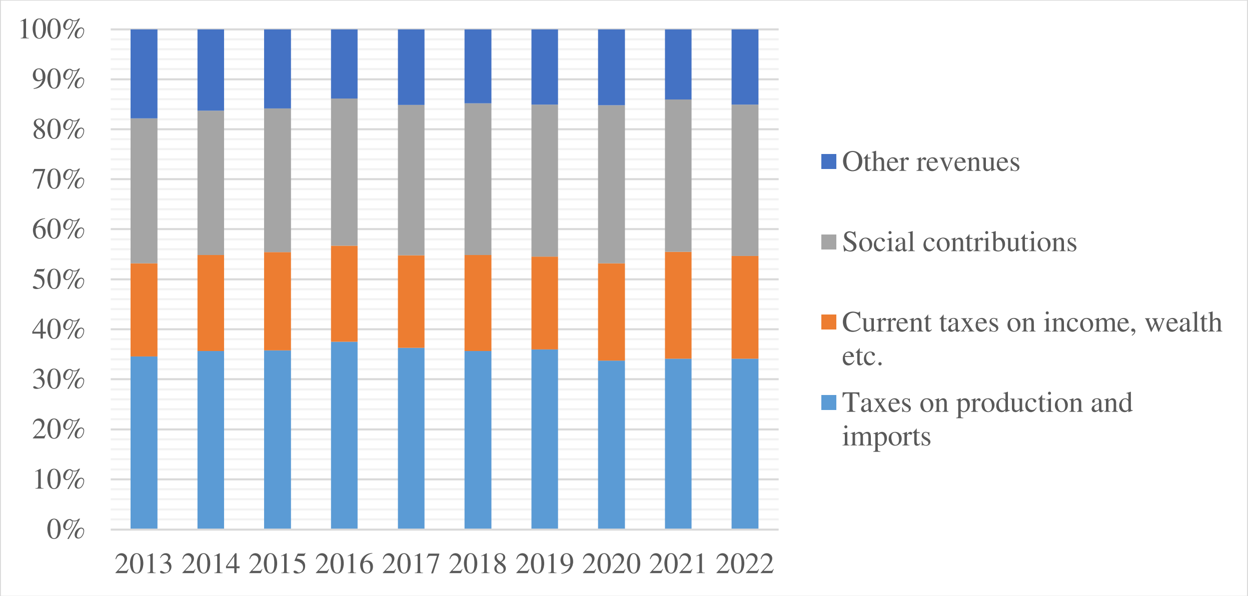

Figure 1. Distribution of government revenue from taxes and other sources
Source: Compiled by author based on Table RR057 (Statistics Estonia, table RR057).

Approximately 35% of the revenues are collected in the form of taxes on production and imports, social contributions fluctuate between 25-35% and current taxes on income, wealth etc. contribute approximately 20% of the total revenue. Other sources of revenue fluctuate between 10-20%. The distribution of government revenue from different sources has been rather stable in the last 10 years. In addition to VAT, social contributions and PIT being main sources of income, the latter two are associated with labour taxation. Individuals are obliged to pay PIT, employees' unemployment insurance premium (and if applicable, insurance premiums of mandatory funded pension) on their income and additionally, their employer is obliged to pay social tax and employers' unemployment insurance premiums on the income from employment. Therefore, government revenue arising from social contributions and current taxes on income, which in 2013-2022 fluctuated around 50% of total government revenue, is directly affected by compliance or resistance of the individuals and companies to report wages to the tax authorities.

The design of tax system can therefore significantly affect the tax revenues of a government as well as distribution of income, social equity and economic efficiency (Arsić *et al.*, 2015). Establishing the right balance in taxation is crucial – understanding and optimizing tax systems, especially in the context of labour taxes is important for implementing a fair and sustainable path for generating revenue to fund public services provided by government. Any inefficiencies in the design of tax system and the administration, or unfair treatment of taxpayers can result in a decision to shift to a shadow economy.

1.2. Shadow economy and tax evasion

The shadow economy, also known as the informal, underground, cash or black economy, refers to economic activities that are not regulated, monitored, or taxed by the government. It operates outside the official channels of the economy and includes various forms of unreported or underreported transactions. The shadow economy is characterized by a lack of formal oversight, taxation, and compliance with labor laws. (Lippert & Walker, 1997)

In economics, three characteristics are used to classify the acitivities in shadow economy (*Ibid.*):

1. Are the activities market-based (monetary) or non-market based (non-monetary)?
2. Are the activities legal or illegal?
3. Are the activities carried out for tax evasion or tax aversion?

Table 1 illustrates non-exhaustive list of different activities included in the classifications based on the three characteristics listed above that individuals, households and firms can engage in. Firstly, the activities are allocated based on the existence of a financial aspect. Drug trafficking, prostitution and undeclared work expect a transfer of money from the service recipient to the service provider, which does not occur in barter trade, household work or production of drugs for personal use. Secondly, the legality of the activities is considered. Drug trafficking and prostitution or drug handling are considered illegal, working or household work are considered legal activities. Finally, the intentions behind the legal activities can be either to evade taxes or avoid taxes. Tax avoidance entails finding loopholes in tax systems such as deductions and exemptions to reduce the tax obligation and is seen as legitimate act (Alm, 1988). Tax evasion on the other hand involves purposeful deception of tax authorities to evade tax obligations (*Ibid.*).

Table 1. Parts of shadow economy

Type of activity	Monetary transactions		Non-monetary transactions	
Illegal activities	Drug trafficking, prostitution, smuggling, scams, etc		Barter trade of illegal goods and services, production of drugs for personal use, etc	
	Tax evasion	Tax avoidance	Tax evasion	Tax avoidance
Legal activities	Undeclared income and work	Fringe benefits	Barter trade of legal goods and services	Household work

Source: Lippert & Walker (1997); Müürsepp (2015)

When comparing monetary and non-monetary transactions, it is more difficult to estimate the extent of non-monetary transactions. For example, individuals' earnings reported to tax authorities in the form of tax returns can be compared to individuals' answers in the Labour Force Survey (hereinafter LFS) regarding income or the Household Budget Survey (hereinafter HBS) regarding expenditure. Large discrepancies between administrative data and survey data can indicate unreported income. On the other hand, it is much harder to analyse the prices of the non-monetary transactions taking place in the form of barter trade of goods and services and to estimate the size of the non-market based activities in the shadow economy.

Even though shadow economy can be defined as a part of the total economy that is unobserved due to the households and businesses keeping their activities undetected, it is important to note that there are alternative definitions that may significantly affect the estimates of the size of the shadow economy (Lippert & Walker, 1997). There are also several approaches to measure the extent of the shadow economy in different countries, using either questionnaires, administrative data, surveys or modelling. For example, the size of shadow economy differs drastically if comparing the figures estimated by SA, Putniņš & Sauka (2015), Schneider (2016), Tafenau *et al.* (2010), or Pissarides & Weber (1989). Also, Turu-uuringute AS (on behalf of Estonian Tax and Customs Board or Maksu- ja Tolliamet, hereinafter EMTA) and Estonian Institute of Economic Research (Eesti Konjunktuuriinstituut, hereinafter EKI) have investigated the extent of shadow economy in Estonia using surveys.

SA assesses the shadow economy following exhaustiveness principle as one of the components of the national accounts when calculating the gross domestic product (GDP). Based on official data, the financial transactions in the shadow economy are captured, the main elements of which are unreported employees, illicit trade, envelope salary and tax fraud (Müürsepp, 2015). However, the size of the shadow economy estimated by SA is relatively low compared to estimates using other estimation approaches, varying between 3-4% of GDP during 2009-2015 (*Ibid.*). As GDP is explicitly a measure of market-based output, several transactions are not a part of domestic production definition and are therefore not accounted for (Lippert & Walker, 1997).

Putniņš & Sauka (2015) used surveys of company managers to measure the extent of the shadow economy. They argue that due to the unique position, the managers, viewed as experts, ought to know how much of the business income and wages go unreported in the company. Their method presents estimates of misreported business income, unregistered employees and unreported wages as well as an estimate of the size of a shadow economy as a percentage of GDP. Putniņš & Sauka (2023) have estimated that the three main components of the shadow economy in Estonia in 2022 are underreporting of salaries (44.5%), underreporting of employees (28.0%) and underreporting of income (27.5%). What is more, they estimate, that approximately 16.8% of salaries paid by the employers are concealed from the government. The percentage of envelope salaries received has fluctuated between 11.5-18.1% in the last decade (2013-2022) according to their estimations.

Schneider (2016) utilizes the multiple-indicators-multiple-causes (MIMIC) procedure as a latent estimator to measure the extent of shadow economy in 25 EU countries. The MIMIC procedure is based on the statistical methodology of unobserved variables, which allows to investigate the relationships between observed variables (indicators) and unobserved variables (latents). The results suggest that Estonia belongs among the countries with largest shadow economies, accounting to 25.4% in 2016. Schneider (2016) observes that the size of shadow economy increases from Northern Europe to Southern Europe and from Western Europe to Eastern Europe. Comparing the results from 2016 to 2015, the size of the shadow economy decreased in most countries. Tafenau *et al.* (2010) have also estimated the extent of the shadow economy in the regions of the European Union using MIMIC approach on the NUTS 2 level regions and have found that the extent of the shadow economy varies a lot within several European countries. However, Estonia is viewed as a single region and the national average of the shadow economy in 2004 as a percentage of the reported GDP is estimated to be 16.3-16.6%.

Pissarides & Weber (1989) employ an expenditure-based approach to estimate the size of the black economy in Britain. The method involves analysing discrepancies between reported income and expenditure. The methodology has been also applied Kukk & Staehr (2014) who found that Estonian households with business income underreport 62% of their actual total income, while Kukk *et al.* (2019) find income underreporting to be more than 40% of self-employed household income on average.

Turu-uuringute AS (2023) estimates the shadow economy for 2022 by conducting a population survey and the main areas of interest are additional income sources, envelope salaries and consumption of illegal tobacco and alcohol. The main findings of the survey regarding envelope salaries is that nearly 25% of the population knows someone who earns envelope salary, however, 5% of individuals have received an envelope wage on a regular basis or from time to time in 2022, making up 34% of the individuals' total salary. The construction sector stood out the most with the share of envelope wages. The same approach was used by EKI (Josing, 2016) and according to their report, 10% of the respondents received envelope salary in 2015, showing a decline in envelope salaries compared to previous years. Individuals active in the construction sector made up 33% of the individuals earning envelope wage in 2015, supporting even more the existence of unreported income in constrution.

The Eurobarometer (2020) questionnaire on Estonia has concluded that 27% of the respondents have said that they personally know someone who works without declaring all or part of their income to tax or social security authorities while 6% of the respondents state that they have carried out undeclared paid acitivites themselves. In total, 20% of respondents claimed to be open to the idea of receiving payment from their employer that they knew would not be declared to the tax authorities. Considering the importance of revenue from labour taxation for Estonian government and the estimations for shadow economy and underreporting of income, analysing labour tax evasion is of great importance.

It should be noted that labour tax evasion can take place both at the extensive margin and at the intensive margin. The extensive margin is considered to encompass undeclared employees – individuals who work for the firm but are not registered in the working registry and therefore do not receive income reported to the tax authorities at all. The intensive margin is considered to be underreporting of labour income, meaning that the individual is registered in working registry, but only part of the income is reported to the tax authorities and rest is received in cash or in kind, in other words as an envelope wage. (Alm & Malezieux, 2021)

Mineva & Stefanov (2018) have observed that non-declared cash payments in addition to the reported income are one of the most complex tax fraud issues. Compared to undeclared employment, where the person conducts lawful but undeclared activities which can be rather easily traced, intensive margin could be executed either in form of under-declaring the actual working time or under-declaring the full-time salary, making it difficult to assess if the administrative wage is accurate or not.

1.3. Dynamics of tax evasion

Tax evasion by firms and employees is complex and multifaceted, involving various personal, economic, psychological, and institutional factors. The attitude towards tax evasion or the decision to engage in illicit practices is a combination of individual and corporate views and incentives. Privitera *et al.* (2021) have pointed out that psychological motives are more important than the economic incentives. Pickhardt & Prinz (2014) find that personal traits and attitudes toward tax evasion and compliance are rather unchangeable but can be influenced by interactions with other individuals. Levenko & Staehr (2023) also find that personal norms and perceived norms of the peers are key predictors of tax compliance, which is supported by Hashimzade *et al.* (2013), who find that the tax compliance decision is based on the context of the social environment of the taxpayer.

Different psychological factors involved include perception of fairness and risk aversion. If the tax system is perceived as unfair or the tax burden deemed disproportionately high, the individuals and companies may incline towards tax evasion. What is more, trust in government and courts is negatively related to tax evasion. From an enterprise point of view, corporate governance and culture play a huge role as companies understanding the social responsibility and ethics are less likely to engage in tax evasion. Also, internal controls and governance impact the possibilities to engage in misconduct. (Abdixhiku *et al.,* 2017)

Nevertheless, firms and individuals are rational thinkers and engage in tax evasion to gain economic benefit, either by minimizing tax liabilities or maximizing after-tax profits. Tax evasion engagement is seen as a cost-benefit analysis, where perceived benefits from tax evasion should outweigh the risk of being caught and applied costs, fines and legal consequences (Allingham & Sandmo, 1972; Hashimzade *et al.,* 2010). Putniņš & Sauka (2023) find that greater probability of being caught and more serious consequences for not paying taxes discourage entrepreneurs to get involved in such practices. On the other hand, compliance costs also matter and too much time and money spent on reporting and adhering to regulations is burdensome to companies. Slemrod (2007) also discusses that tax evasion itself imposes efficiency costs as non-compliance must be camouflaged, however if they do not exceed compliance costs, tax evasion is perceived more attractive.

Another important factor to explain the decision is the legal and regulatory environment of an economy. Complex tax regulations create opportunities for exploitation and evasion and weak enforcement mechanisms or inadequate penalties may encourage non-compliance (Musgrave 1959; Brennan & Buchanan, 1980). Business legislation, tax policy and performance of tax authorities are seen as important determinants of involvement in shadow economy with dissatisfaction encouraging shadow acitivity (Putniņš & Sauka, 2023). Abdixhiku *et al.* (2017) find a positive relationship between perceived tax burden and tax evasion. Globalisation and increasing cross-border trade have also enabled international tax planning by profit shifting, using tax havens and regulatory arbitrage by using differences in different tax regulations across jurisdictions. Even though international tax planning is rather seen as tax avoidance, it might also encompass tax evasion.

More importantly, a firm's characteristics are related to the extent of tax evasion. Abdixhiku *et al.* (2017) have found that firm's size matters and the larger the firm (in terms of the number of employees), the smaller the extent of tax evasion, also pointing out that firms applying international accounting standards are more likely audited and therefore less likely to engage in tax evasion. Kukk & Staehr (2014) and Kukk *et al.* (2019) find that self-employed households in Estonia are greatly underreporting their earnings, supported by Slemrod (2007) who points out that absence of third-party reporting of wages and salaries facilitates underreporting of income. Putniņš & Sauka (2023) have also found that smaller firms engage in tax evasion more often than larger firms. They have observed that younger firms engage in more shadow activity than older firms. Industry differences may also explain the decision to be engage in tax evasion. Highly competitive industries may experience greater pressure to minimize costs, for example construction sector firms participating in tenders. Generally, the results also support higher tax evasion in sectors that involve higher cash transactions, such as hotels and restaurants, construction and wholesale and retail (Abdixhiku *et al.,* 2017).

As discussed above, numerous studies have analysed the attitudes towards tax evasion, mostly relying on survey data and investigating individual attitudes towards tax evasion. However, it is also relevant to understand how the theory of tax evasion regarding individual decision makers is associated with firm compliance and which are the indicators to be analysed on firm-level to identify labour tax evasion.

1.4. Detecting tax evasion by firms

There are different ways to uncover tax evasion at the firm level. Audit data from tax authorities may represent the most accurate way to find a list of companies who have been engaged in illicit tax practices (Benkovskis & Fadejeva, 2022). However, the data on tax disputes solved in-house is not public and up-to-date tax debt information on a particular firm can be obtained by submitting a debt inquiry on the webpage of EMTA. The audit data from the tax authorities can also be biased towards larger companies or greater benefit as conducting a tax audit is costly and the potential increase in tax revenue should cover the cost of conducting an audit.

This is supported by Bobbio (2017), who has shown using Italian firm-level data that smaller firms also tend to spend less on innovation to remain under the radar and enjoy the cost advantage of evading taxes, resulting in unfair competition. Braguinsky & Mityakov (2015) also present in their results that small enterprises are more eager to engage in tax evasion. Kukk & Staehr (2014) and Kukk *et al.* (2019) find that self-employed household income is greatly unreported. Therefore, it could be assumed that self-employed individuals have the possibility and interest to evade taxes. Kleven *et al.* (2011) also suggest that self-reported income is increasing tax evasion and third-party reporting on the other hand is an effective enforcement device to decrease evasion.

EMTA also monitors the average salary and requests additional checks by the companies as significantly lower salary paid to the employee compared to average salary for the same position in Estonia may indicate the payment of an envelope salary and therefore labour tax evasion (Lepassar, 2024). However, the ratings are not public if not shared or given access to by the company itself. The more fine-tuned approach, taking into account individual characteristics would be estimating a wage regression (Mincer, 1975). The Mincer wage regression is a widely used model to analyse the relationship between an individual's earnings and factors such as education, work experience, and other demographic characteristics. The model is commonly employed in labour economics to understand how human capital influences wages. Gavoille & Zasova (2023) and Benkovskis & Fadejeva (2022) apply wage regression to spot firms with "suspiciously low wages".

What is more, Gavoille & Zasova (2023) suggest there is a link between labour tax evasion and financial fraud. Labour tax evasion is considered one form of financial manipulation and results in particular balance sheet and profit and loss statement patterns such as understatement of revenue, assets, expenditure, or liabilities. The approach to analyse relationship between the probability of financial manipulation and financial statement variables relies on previous accounting research by Massod Beneish. Beneish (1999) presents a financial model designed to detect the manipulation of financial statements by assessing the likelihood of earnings manipulation or financial fraud by a firm. The M-score consists of eight financial ratios that are combined to form a single score. These ratios focus on various aspects of firms' financial statements, such as profitability, cash flows, and accounting quality.

The approach is further developed by Dechow *et al.* (2009), Cecchini *et al.* (2010), Hajek & Henriques (2017), showing that different variables from firms' annual financial reports provide a good indication of whether a company is engaged in financial fraud. Dechow *et al.* (2009) further assure that financial statement information is useful for identifying misreporting and earnings manipulation. Cecchini *et al.* (2010) provide a methdology for detecting management fraud based on support vector machines, which is a machine learning algorithm using supervised learning models to solve complex classification problems. Their approach correctly labels 80% of the fraudulent cases and 90.6% non-fraudulent cases, exceeding the performance of probit method applied by Beneish (1999) detecting 56% of fraudulent cases and logistic regression applied by Dechow *et al.* (2009) detecting 64.5% fraudulent firms.

Hajek & Henriques (2017) examine whether a financial fraud detection model could be developed, combining financial information and linguistic analysis of managerial comments (positive, negative, neutral or other words) from financial reports. They applied different machine learning methods and found that ensemble methods are best at classifying fraudulent companies as fraudulent and Bayesian belief networks work well for classyfying non-fraudulent firms as non-fraudulent.

Following the assumption of tax evasion being associated with financial fraud, Gavoille & Zasova (2023) use a set of balance sheet and income statement items as predictors to detect tax evading firms and implement gradient boosting decision trees for classification purposes. However, they point out the black box nature of this method as one of the main drawbacks. Benkovskis & Fadejeva (2022) also use different explanatory variables derived from financial reports and estimate a probit model for classification purposes.

Analyzing the relationship between firm-level financial and non-financial variables and the probability of engaging in labour tax evasion provides relevant insights for tax authorities to determine whether the allocation of additional resources for tax audits or more profound investigations is necessary. Identifying firms possibly engaged in labour tax evasion by conducting a wage regression is a novel methodology, but as wage data is usually confidential and not available for the wider audience, detecting patterns in publicly available financial reports suggesting tax evasion could be a widely used approach.

2. DATA AND METHODOLOGY

This chapter presents an overview of the data used for thesis purposes as well as the methodological approach applied to analyse firm-level predictors related to labour tax evasion. Section 2.1. provides overview of data source and discussion of variables included in the analysis. In section 2.2. explanations regarding the choice of methods as well as the limitations are provided.

2.1. Data

The paper relies on a matched employer-employee wage dataset with a monthly frequency, population data and annual report data. This dataset is collected by SA, the main data competence centre in Estonia. Data on wages are collected from the employment register (TÖR) and Annexes 1 and 2 of the tax declaration form TSD (declaration of income and social tax, unemployment insurance premiums, and contributions to mandatory funded pension). Data on wages are administrative data, therefore representing the general population and not a sample. Wage data are anonymised[1], meaning that the personal identification code of an individual is replaced by a SA personal identification number to make it impossible to identify individuals from the dataset. Data regarding annual reports originates from the e-Business Register. Data generated and analysed during the thesis are not publicly available due to confidentiality concerns, and sharing the data is not permitted.[2]

Figure 2 illustrates the connections between the three datasets. The anonymised identification numbers of the employees allow combining population and wage datasets into a subsequently detailed dataset of individual characteristics of the employees. The dataset provides information on gross wages, personal income tax payments, social security payments, working time, as well as gender, date of birth, education, date of employment, economic activity of the employer,

[1] The author did not perform anonymisation but received the dataset already anonymised. Therefore, the author had no possibility to identify anyone personally.

[2] Access to various datasets owned by SA for research purposes can be obtained by submitting a corresponding request, if not published on the webpage of SA.

occupation, location of workplace and more. Additionally, the company registry code connects the wage data and the annual report data.

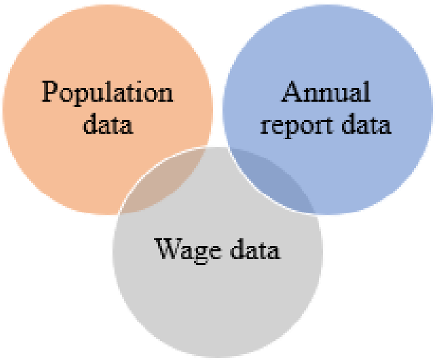

Figure 2. Data map
Source: Composed by author
Notes:

1. Wage and population datasets can be merged by anonymised identification numbers of the employees.
2. Wage and annual report datasets can be merged by company registry code.

The analysis focuses on four Statistical Classification of Economic Activities (NACE) sectors only: manufacturing (NACE 1-digit level code C), construction (NACE 1-digit level code F), wholesale and retail trade; repair of motor vehicles and motorcycles (NACE 1-digit level code G), transportation and storage (NACE 1-digit level code H), following Gavoille & Zasova (2023). Regarding occupations, 1-digit level code of International Standard Classification of Occupations (ISCO) was included, omitting employees in occupation code zero, referring to armed forces occupations. For location of workplace, the Nomenclature of Territorial Units for Statistics (hereinafter NUTS) is derived from Estonian Administrative and Settlement Classificator (hereinafter EHAK), dividing the country into five different regions, i.e. Northern Estonia, Central Estonia, North-Eastern Estonia, Western Estonia and Southern Estonia.

2.1.1. Wage and population data

Several steps were performed in the preprocessing of data. The following calculations were performed in order to obtain variables for analysis purposes and to implement the necessary trimming of data:

1) monthly average wage was calculated by adding up all monthly payments and diving the sum by the count of months engaged in employment by employer;

2) age in years was calculated by subtracting date of birth from year end date[3];
3) work experience in years was calculated by subtracting the date of employment from year end date.

Step 1) was necessary due to the nature of monthly wage data. This calculation smoothed the fluctuations in monthly salary compared to using only working time rate adjusted wage from one particular month as the selected month could have employees receiving abnormally high salary due to bonuses or receiving no salary at all. Incorporating a month with extremely high salaries due to bonuses incorrectly reflect a person's income and contribute to upward bias in the results. For example, year-end or Christmas bonuses are common among businesses. On the other hand, choosing a month where individuals report no adminstrative salary, the individuals are omitted from the analysis. This could happen, for example, with many construction workers as execution of various construction stages is highly seasonal.

Therefore, in order to take into account all employees in the dataset without their average wage being dependent on how many months they were engaged in employment, the annual summarised payments were divided by the months engaged in work. Otherwise, if divided by twelve months for all employees, the monthly average salary could be underestimated for those who were not engaged in the labour market for one or more months of the year. Calculating the average working time rate adjusted monthly salary based on annual income eliminates these issues.

The calculation was done separately for different employers, and therefore accounting for individuals engaged in several employment contracts during a year. As presented in table 2, individuals were mainly working for one employer only, but there were also individuals who worked for 2-8 different employers during a year. These individuals are not omitted from the analysis, but the wage calculations take into account their rather "jumpy employment" as the main focus is on the labour tax evasion of the employer. Therefore, including only one employment and omitting others would not be reasonable and could potentially exclude tax evading employers from analysis.

[3] Year end date refers to the year for which the regression analysis was carried out (Dec. 31, 2021 or 2022).

Table 2. Unique employer-employee combinations

Year	Number of unique employer-employee pairs	Number of unique employees	Number of employees employed by several employers
2021	316 679	289 502	27 177
2022	318 542	292 305	26 237

Source: Author's calculations
Note: The total number of unique employer-employee pairs in the dataset presents the observations from the four NACE sectors selected for analysis purposes.

The age calculation in step 2) resulted in values ranging from 11 to 91 years in the dataset. Only individuals aged 16-65 years are kept in the analysis. Individuals younger than 16 are omitted due to serving the mandatory minimum of general education requirement and individuals older than 65 are omitted due to old-age pension. The retirement age does not automatically result in complete exclusion from the labour market, but rather part-time participation or irregular activities. Post state pension age workforce works fewer hours than younger workers and the gap in hours is greater for men than women (Smeaton & McKay, 2003). The Estonian labour market policies are encouraging the elderly to stay longer in paid employment, increasing the labour market participation after they reach the state pension age. A total of 18 066 individuals were omitted in 2021 and 14 752 individuals were omitted in 2021, mainly elderly.

Only full-time employees according to working time rate were included in the analysis. However, this does not account for individuals who may have worked for only part of the month (for example quit work in the middle of a month) and it is also not visible in the data source. One indicator suggesting a shorter span of work is receiving working time adjusted monthly salary lower than official minimum wage. Additionally, receiving markedly less than the minimum wage could suggest to data irregularities or errors. To omit individuals whose salary is downward biased due to their irregular inclusion in employment during one or several months, all individuals receiving average working time rate adjusted monthly salary less than the minimum wage by law for that year (584 euros for 2021 and 654 euros for 2022), are omitted. As a result, 31 927 individuals were omitted in 2021 and 35 330 individuals were omitted in 2022.

It is important to note that this exlusion could possibly omit individuals who have received benefits for temporary work incapacity from the Social Insurance Board, such as sickness benefit or care benefit. Individuals earning the mininum wage are more sensitive to a decrease in income as experiencing loss of income in even one month in period under investigation would shift them to earning less than minimum wage by law.

Calculation of step 3) considers an exemption regarding these employees whose employment contract had ended during the observed timespan and in this case, the work experience in years was calculated by subtracting the date of employment from the end of employment. In other cases, the experience is calculated by subtracting the date of the employment from the end of the year for which the analysis was carried out. The experience is rounded down to a full year.

The variable for occupation is not available for all employees in TÖR; the missing observations are however only a small proportion from all observations, amounting to 982 observations in 2021 and 676 observations in 2022. Due to the low magnitude, the observations with missing occupation were dropped. Moreover, data on education matched from population dataset is not available for all observations, 72 910 observations were missing for 2021 and 81 212 observations were missing for 2022, making up approximately a quarter of the observations.

To understand whether the issue of missing educational attainment is systematic or random, observations with missing values were further investigated. The average salary of the individuals in 2021 was 1368 euros and the median salary accounted to 1193 euros. Compared to the dataset including observation with and without educational attainment variable, the average salary was 1445 euros and the median salary was 1213 euros. The share of different occupation, region, NACE and gender categories follows the distribution of the initial dataset. Therefore, observations with missing educational attainment should not affect the results of the analysis severely, if omitted. However, robustness checks are carried out to confirm that the results are not compromised by omitting observations with missing education.

Table 3 represents the number of unique employer-employee pairs and omitted variables in each steps as well as the final sample used for thesis purposes.

Table 3. Observations dropped and final sample

Year	Number of unique employer-employee pairs	Younger than 16 or older than 65	Wage below minimum wage	No occupation	Armed forces	No educational attainment	Final sample size
2021	316 679	18 066	31 927	982	3	72 910	192 791
2022	318 542	14 752	35 330	676	0	81 212	186 572

Source: Author's calculations
Note: The montly minimum wage was 584 euros for 2021 and 654 euros for 2022.

The final sample includes 192 791 observations for 2021 and 186 572 observations for 2022, so a total of 123 888 observations were dropped for 2021 and 131 970 observations were dropped for 2022. The descriptive statistics of the final sample variables are included in table 4.

Table 4. Descriptive statistics

Year	2021				2022			
Variable	mean	st.dev	min	max	mean	st.dev	min	max
Wage	7.16	0.50	6.37	10.89	7.27	0.49	6.48	11.20
Woman	0.40	-	0	1	0.40	-	0	1
Age	46.80	10.20	17	65	46.60	10.30	16	65
Education	-	-	1	4	-	-	1	4
Experience	8.20	6.67	0	40	7.60	6.86	0	40
NACE	-	-	1	4	-	-	1	4
Occupation	-	-	1	8	-	-	1	8
Region	-	-	1	5	-	-	1	5

Source: Author's calculations
Note: NACE, Statistical Classification of Economic Activities.

The mean wage is higher in 2022 compared to 2021 which is expected as average salary has increased. Mean of 0.4 for gender shows that there are slightly more men in the dataset than women. The average age of the individuals is 46.8 years for 2021 and 46.6 years for 2022. The average experience of the individuals is 8.20 years in 2021 and 7.6 years in 2023. Regarding categorical variables, the distribution to different categories is visible in Appendix 1.

2.1.2. Annual report data

Due to the matched employer-employee wage data, referring to availability of the registry code of the employer for each employee, wage data can later be linked to various firm-level financial and non-financial variables from the annual reports. For each employer balance sheet and profit and loss statement as well as annexes are available. Therefore, different balance sheet and profit and loss statement values are retrieved, such as assets, liabilities, revenues, expenditures and profit. Also, more detailed allocation of balance sheet and profit and loss statement items in annexes, average number of employees reduced to full-time, as well as NACE is available.

Following previous studies on tax evasion and financial fraud by Hajek & Henriques (2017), Gavoille & Zasova (2023) and Benkovskis & Fadejeva (2022), different ratios are calculated using various annual report items. Preprocessing of data also included removing observations with missing values as not all companies have all balance sheet or profit and loss statement values available. Benkovskis & Fadejeva (2022) keep the set of independent variables short and include ratios based on the most common financial indicators. The same approach is followed in this thesis as the aim is to also investigate tax evasion in micro enterprises. Due to their simplified reporting, not too many financial indicators are included in the annual reports and inclusion of very specific indicators in the analysis could result in exclusion of a large proportion of small firms, as only observations where all the variables included in the final model were present, were kept. The number of unique firms in the dataset, omitted observations and final dataset is presented in the table 5.

Table 5. Observations dropped and final sample size

Year	Number of unique firms	No employees	No profit or turnover	No cash or assets	No liabilities	No COGS	Infinite values	Final sample
2021	65 187	32 255	139	5673	242	2553	534	23 791
2022	66 207	32 721	152	5783	202	2519	516	24 314

Source: Author's calculations

Notes:

1. Missing values in assets include both total assets and current assets.
2. Missing values in liabilities include both total liabilities and short-term liabilities.
3. COGS, cost of goods sold.
4. Infinite values arose from division by zero.

Approximately half of the companies in the annual report dataset reported having zero employees. Looking into these firms, several patterns can be detected. Part of the firms could have been inactive during the year as turnover is zero or relatively small. Others seem to have zero employees due to error or potential engagement in labour tax evasion at the extensive margin (unreported employees). Taking into account the field of activity of firms and turnover arising from acitivites, a firm should have employees on their payroll. The companies for which the number of employees is zero are cross-checked with data on administrative wages to see how many employees have received reported salary in the corresponding year. As a result 893 firms (868 in 2022) paid and reported salaries to employees in 2021 which show zero employees in annual report data and 32 255 firms (32 721 in 2022) did not report any salary payments in 2021. For the firms for which the zero employees was erroneous, the number of different employees who received salary according to wage data, is used as a proxy. The firms who did not report any salary payments are omitted from the analysis, as the aim is to investigate labour tax evasion on the intensive margin and not on the extensive margin. The same approach is applied on data from 2022.

Additionally, some ratio calculations produced infinite values (dividing with zero). Cecchini *et al.* (2010) tackle the issue of division by zero by replacing zero values with 0.001; however it should be approached with caution and the effect on the analysis results should be clear. Therefore, to avoid biases and unclear effect on interpretation, the infinite values were omitted from the analysis. The descriptive statistics of the final sample variables is included in table 6.

Table 6. Descriptive statistics

Year	2021				2022			
Variable	Mean	St.dev	Min	Max	Mean	St.dev	Min	Max
NACE	-	-	1	4	-	-	1	4
Size	8.30	42.74	1	3 115	8.20	42.50	1	3010
Log of turnover	12.39	1.77	3.85	20.22	12.47	1.77	4.09	20.79
Debt to assets	1.08	51.65	0	7 661.33	0.65	10.70	0	1377.00
Short-term debt to current assets	2.29	66.38	0	6 663.40	1.68	39.80	0	4711.00
Cash to assets	0.28	0.27	-2.33	1.00	0.28	0.28	-0.06	4.26
Turnover to assets	3.25	34.70	0	4 404.00	3.57	35.4	0	3270.00
COGS to turnover	0.66	6.34	0	926.79	0.65	4.83	0	458.56

Source: Author's calculations

Notes:

1. NACE, Statistical Classification of Economic Activities.
2. COGS, cost of goods sold.

The descriptive statistics show that enterprise data exhibits a lot of variance. The mean size of the company is 8.30 employees in 2021 and 8.20 employees in 2022, showing that the number of employees has slightly decreased. The mean of log turnover increased when comparing 2022 to 2021. Debt to assets or short-term debt to assets is extremely high for example in case where a short-term loan liability has been taken on but there are very few assets owned by the firm. The same situation arises regarding turnover to assets. The cash to assets ratio is negative in case the firm has a credit account in use. COGS is high when a company spends more on the intermediate consumption to provide products and services than it receives taxable revenue. It is important to note that turnover accounts to sales revenue of a company only, and the high ratio could be due to the fact that company has reported other revenue. Regarding the categorical variable NACE, the number of observations falling into each category is presented in table 7.

Table 7. Number of observations per category for NACE

NACE	2021	2022
Manufacturing	4375	4421
Construction	7066	7479
Wholesale and retail trade; repair of motor vehicles and motorcycles	9116	9151
Transportation and storage	3234	3263
Total	23 791	24 314

Source: Author’s calculations

The number of observations falling into each NACE is relatively similar for both, 2021 and 2022. Approximately 38% of firms are active in NACE wholesale and retail, 30% in construction, 18% in manufacturing and 14% in transportation and storage.

2.2. Methodology

This section will give an overview of the methodology used in the empirical analysis to obtain subsets of tax evading and tax compliant firms as well as analyse the firm-level predictors contributing to the probability of being engaged in labour tax evasion. The methodology of the empirical analysis is based on previous research in this field, implementing Ordinary Least Squares (hereinafter OLS) and logistic regression. In subsection 2.2.1, the methodology to obtain samples of tax compliant or tax evading firms is described in detail. In subsection 2.2.2, the methodology to model the relationship between firm-level financial and non-financial variables and labour tax evasion is explained. What is more, classification of the remaining firms as tax evading or compliant is discussed.

2.2.1. Obtaining subsets of tax evading and tax compliant firms

Defining the treated group of firms could be approached differently. Kukk & Staehr (2014) find that households with business income underreport 62% of their actual total income and Kukk *et al.* (2019) find income underreporting to be more than 40% of self-employed household income on average. Braguinsky & Mityakov (2015) show that small enterprises are more eager to engage in tax evasion. Gavoille & Zasova (2023) and Benkovskis & Fadejeva (2022) apply wage regression to pinpoint firms with "suspiciously low wages".

To obtain a subset of firms for which the classification is tax evading, the Mincer (1975) wage regression model is employed as a starting point in developing an empirical earnings model, following Gavoille & Zasova (2023) and Benkovskis & Fadejeva (2022). Mincer wage regression model is an extensively used model employing OLS method to analyse the relationship between an individual's earnings and various other human capital variables, primarily education and experience. The natural logarithm of earnings is used to address issues related to the distributions of earnings, and the model aims to estimate how changes in education, experience and other individual characteristics affect the individual's earnings. To spot firms paying suspiciously low wages to their employees, Gavoille & Zasova (2023) regress the log of wage for an individual employee against characteristics such as age, experience, gender, field of activity, location of workplace, occupation and educational attainment. They consider employees in the bottom 10% of the residual distribution receiving abnormally low wages as the wage regression predicts significantly higher salary taking into account the individual characteristics of the employee. They classify firms with at least one employee in the bottom of the residual distribution in a given year as tax evader.

However, as the size and therefore individuals employed by a company varies a lot, the bottom 10% of a residual distribution should be approached with caution. For example, if one employee from a company with more than 100 employees falls in the bottom 10% of the residual distribution, the firm would be classified as tax evading. On the other hand, a firm with ten employees who all fall into the bottom 10% of the residual distribution is classified tax evading as well. The scale of beforementioned cases is not readily comparable and an individual falling in the bottom due to large actual and predicted salary discrepancy could be there for other reasons than employer being engaged in labour tax evasion. Benkovskis & Fadejeva (2022) also disuss that for some individuals seemingly low wages could be due to unobserved worker characteristics. They require two conditions to be met in order to be classified as a treated group. Firstly, the share of employees with "suspiciously low wages" in a given year is equal to 50% or more and secondly occupation data should be available for one third or at least 10 employees for that firm. Therefore it is beneficial to take into account the population of employees and share of employees falling to the bottom 10% of the distribution. It is therefore assumed that firms for which 50% or more of employees fall into the bottom 10% of the residual distribution pay a suspiciously low wage and are therefore classified as tax evading.

Harmon & Walker (1995) also investigate whether instrumental variables (hereinafter IV) approach should be preferred over OLS due to endogeneity and biases. They conclude that, even though IV estimates are nearly double the estimates for OLS, IV estimation provides less precise estimates, and the differences from OLS are not statistically significant. Therefore, IV is discarded for the purpose of this analysis and OLS is chosen. The standard errors are assumed to be heteroscedastic, and robust standard errors are therefore applied. The final wage regression model is presented as follows:

$$\ln(wage) = \alpha + \beta_1 gndr + \beta_2 age + \beta_3 age^2 + \beta_4 educ + \beta_5 exp + \beta_6 exp^2 + \beta_7 nace + \beta_8 occup + \beta_9 reg + \varepsilon \qquad (1)$$

The list of variables is included in Appendix 2. The dependent variable is the natural logarithm of the average monthly wage and the independent variables included in the wage regression are:

1. Individual characteristics such as age (*age*), age², experience (*exp*), experience² (expressed in years) and a dummy variable taking the value 1 if the respondent is a woman (*gndr*);
2. A set of dummy variables (*reg*) indicating the NUTS region in which the respondent works (reference group is Northern Estonia);
3. A set of dummy variables (*nace*) indicating the NACE of the employer of the respondent (reference group is manufacturing);
4. A set of dummy variables (*occup*) indicating the type of occupation of the respondent (reference group is managers);
5. A set of dummy variables (*educ*) indicating the education of the respondent (reference group is pre-school education);
6. Error term ε.

Regarding tax-compliant firms to be used as a control group, Gavoille & Zasova (2023) assume that firms owned by Nordic companies are less likely to engage in illicit corporate activities. They have obtained similar results as Braguinsky & Mityakov (2015) in their previous research regarding the transparency and law-abiding cultural norms (Gavoille & Zasova, 2021). Benkovskis & Fadejeva (2022) have identified “definitely compliant” companies to be state-owned firms and companies whose owners are located in low-corruption countries. DeBacker *et al.* (2015) have found that firms owned by parents operating in countries with higher corruption levels, evade more tax in the USA, supporting the findings and approach. It could be also argued, whether wage regression could be used to distinguish both, tax evading and tax compliant companies. However, companies not falling into the bottom 10% of the residual distribution can still be engaged in labour tax evasion and should not be therefore classified as “definitely compliant”. The aim is to find patterns indicating labour tax evasion also for those who are not in the bottom 10% of the residual distribution.

This thesis follows Gavoille & Zasova (2021, 2023), considering companies whose owners reside in Nordic countries (Iceland, Norway, Sweden, Finland or Denmark) as tax compliant. If there are differences in the country the parent company and group parent company are registered in, parent company is considered more influential in importing corporate culture and conduct than group parent company. However, it should be noted that even though findings by Gavoille & Zasova (2021, 2023) and Braguinsky & Mityakov (2015) support the approach, selecting Nordic owned companies as tax compliant relies on strong assumptions.

2.2.2. Firm-level predictors of tax evasion

After obtaining a sample of firms for which the true and false type (classification) is known, it is important to distinguish other firms between tax compliant and tax evading using different firm-level non-financial and financial indicators. The literature on accounting and computer science on fraud detection has shown good prediction performance using different variables from firms’ annual financial reports although no formal model in economic theory has been provided and the approach rather lies on identifying the patterns (Gavoille & Zasova, 2023). Therefore, different balance sheet and income statements values have been used to calculate ratios that could indicate patterns associated with labour tax evasion. The chosen variables mainly follow Benkovskis & Fadejeva (2022) and are presented in Appendix 3.

Gavoille & Zasova (2023) proceed by splitting the sample of firms for which they know the true type to training (80%) and test (20%) sample. They then train a gradient boosting algorithm to distinguish between tax compliant and tax evading firms using the training sample and financial variables of the firms as input variables. Afterwards, the model is applied to the firms in the test sample and as a final step, they classify all the firms in the analysis. Benkovskis & Fadejeva (2022) on the other hand employ a probit model to model the relationship and predict the probability that each firm is engaged in tax evasion. They discuss that despite potential losses in predictive power, using a probit model is transparent and allows to report the sign and significance for coefficients on the firm-level predictors. What is more, probit results in an estimate of the probability of a firm being involved in tax evasion and not a binary classification. First, they estimate a probit model on firms for which the true type is known. After this, the probit model is used to predict the out-of-sample probability of being engaged in tax evasion and finally, the goodness of the model is evaluated.

To analyse the factors that contribute to probability of tax evasion, a probit or a logit model is considered as both are designed for dependent variables taking on values between 0 and 1, being therefore suitable for analysing the relationship between different firm-level financial and non-financial predictors and the event of tax evasion occuring. What is more, the interpretability of the model is important to further analyse the relationship between labour tax evasion and firm-level predictors. The choice between probit and logit model is dependent on the characteristics of the data and underlying assumption, however, both models produce similar results in many cases. As per interpretability, logit model coefficients are often found more straightforward to interpret due to the simplicity of log-odds scale compared to changes in the standard normal distribution (Gujarati, 2003). Therefore, the logistic regression model is employed for the purpose of this thesis.

After combining the two subsets of firms for which the binary classification assumption of tax compliant and tax evading was done and merging it with firm's financial data, logistic regression model is used to model the relationship between binary outcome and predictor variables. The final logistic regression model takes the following form:

$$Y_i = \ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta X_i + \varepsilon_i \qquad (2)$$

Y_i is the dependent variable of respondent *i*, representing the probability of being engaged in labour tax evasion. β_0 is the intercept, β is the parameter estimate. X_i is a vector of explanatory variables of respondent i, including different firm-level financial and non-financial variables that may correlate with tax evasion. Lastly, ε_i is the error term.

Logistic distribution of the errors is assumed and maximum likelihood estimation method is used in parameter estimation to obtain most accurate estimates. The standard errors are assumed to be heteroscedastic, and robust standard errors are therefore applied. As the logit model only allows determining the direction of the effect of independent variables on the dependent variable, the marginal effects are also calculated to better understand the extent of the impact of independent variables on the dependent variable. However, it is important to note that the estimated coefficients from logistic regression present conditional correlations between the dependent and independent variables and should not be considered as causal relationships when interpreting and discussing the results. (Gujarati, 2003)

There are several ways to measure the prediction performance of logistic regression model. Following Hajek & Henriques (2017), confusion matrix and different performance metrics are presented, i.e. recall, type I error, specificity, type II error, accuracy, F-measure and AUC-ROC (Area Under the Receiver Operating Characteristic Curve).

Recall or true positive rate is the number of firms correctly classified as evading as a percentage of all evading companies (*Ibid.*):

$$TP\ rate = \frac{TP}{P} \tag{3}$$

Type I error or false positive rate is the number of firms incorrectly classified as evading as a percentage of all compliant companies (*Ibid.*):

$$FP\ rate = \frac{FP}{N} \tag{4}$$

Specificity or true negative rate is the number of firms correctly classified as compliant as a percentage of all compliant firms (*Ibid.*):

$$TN\ rate = \frac{TP}{N} \tag{5}$$

Type II error or false negative rate is the number of firms incorrectly classified as compliant as a percentage of all evading firms (*Ibid.*):

$$FN\ rate = \frac{FN}{P} \tag{6}$$

Accuracy is defined as a percentage of observations correctly classified (*Ibid.*):

$$Accuracy = \frac{(TP+TN)}{P+N} \tag{7}$$

F-measure is the mean of precision and TP rate (*Ibid.*):

$$F\text{-}measure = 2 * \frac{Precision * TP\ rate}{Precision + TP\ rate} \tag{8}$$

The AUC-ROC score is used to evaluate the performance of binary classification of the logistic model. The AUC-ROC score ranges from 0 to 1, where 0 indicates a poor model and 1 indicates a perfect model that makes all predictions correctly. (*Ibid.*)

Lastly, the logit model is employed to predict the out-of-sample probability of being engaged in tax evasion for all firms:

$$\hat{p} = \gamma(X_i\hat{\beta}) \tag{9}$$

Here, $\hat{p}$ on the left-hand side denotes the predicted probability of labour tax evasion for firm i. X_i is a vector of explanatory variables for respondent i, including various firm-level financial and non-financial variables. $\hat{\beta}$ signifies the estimated coefficients, γ is representing the logistic function. The probability is computed separately for each year.

Setting the probability threshold is crucial for the outcome, as the absolute share of the firms classified as tax evading is heavily dependent on the subjective threshold. Benkovskis & Fadejeva (2022) set the probit estimation threshold to be at 0.84, therefore firms with predicted probability above 84% are classified as evading, and the others as compliant. The threshold is rather high and it can result in a model that is conservative in its true positive predictions. This leads to fewer false positives, but could potentially miss many true positives. For the purpose of this thesis, the predicted probability threshold is set to be at 0.65 and a robustness analysis is done using the probability threshold of 0.84, following Benkovskis & Fadejeva (2022).

3. EMPIRICAL ANALYSIS

This chapter presents the main results from the wage regression and logistic regression and provides an overview of the robustness checks carried out to confirm the reliability of the results. Furthermore, discussion regarding results is presented in the section 3.3. as well as shortcomings of the analysis and suggestions for improvements and further research.

3.1. Main results

This section will give an overview of the main results of the empirical analysis to obtain subsets of tax evading and tax compliant firms as well as analyse the firm-level predictors contributing to the probability of being engaged in labour tax evasion.

3.1.1. Wage regression

To obtain the set of tax evading firms, i.e. the firms which pay "suspiciously low wages" to their employees, wage regression is performed. The data used is combined from matched employer-employee wage data and population data, as wage data do not include a variable of educational attainment. The reference group for NACE is manufacturing, for education is pre-school education, for region is Northern Estonia and for occupation is managers. The results of the wage regression for 2021 and 2022 separately are presented in table 8.

Table 8. Wage regression

	2021 ln(wage)	2022 ln(wage)
Intercept	7.073*** (0.020)	7.270*** (0.018)
Gender	-0.221*** (0.002)	-0.220*** (0.002)
Age	0.021*** (0.001)	0.021*** (0.001)
Age²/100	-0.028*** (0.001)	-0.029*** (0.001)

Experience	0.026***	0.023***
	(0.001)	(0.001)
Experience^2/100	-0.063***	-0.058***
	(0.002)	(0.002)
Construction	-0.133***	-0.132***
	(0.003)	(0.003)
Wholesale and retail trade	-0.066***	-0.050***
	(0.003)	(0.003)
Transportation and storage	-0.101***	-0.080***
	(0.003)	(0.003)
Basic education	0.045*	0.014
	(0.015)	(0.013)
Secondary education	0.071***	0.044**
	(0.015)	(0.013)
Tertiary education	0.142***	0.113***
	(0.015)	(0.013)
Central Estonia	-0.109***	-0.117***
	(0.003)	(0.003)
North-Eastern Estonia	-0.265***	-0.255***
	(0.003)	(0.003)
Western Estonia	-0.158***	-0.158***
	(0.003)	(0.003)
Southern Estonia	-0.120***	-0.127***
	(0.002)	(0.003)
Professionals	0.119***	0.101***
	(0.006)	(0.006)
Technicians and associate professionals	-0.053***	-0.078***
	(0.005)	(0.005)
Clerical support workers	-0.250***	-0.281***
	(0.005)	(0.005)
Services and sales workers	-0.438***	-0.451***
	(0.005)	(0.005)
Skilled agricultural, forestry and fishery workers	-0.482***	-0.451***
	(0.033)	(0.032)
Craft and related trades workers	-0.390***	-0.413***
	(0.005)	(0.005)
Plant and machine operators and assemblers	-0.382***	-0.420***
	(0.005)	(0.005)
Elementary occupations	-0.498***	-0.530***
	(0.005)	(0.005)
Observations	192 791	186 572
R^2	0.316	0.336

Note: Results are based on Eq.1. Significance level * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust stardard errors in parentheses.

The R^2 is 31.6% for 2021 and 33.6% for 2022 and is therefore resembling that of previous studies estimating wage equations. The wage equation by Harmon & Walker (1995) explains 27% of the variance in the depenent variable and Gavoille & Zasova (2023) account to 25%. What is more, the results of the wage regression for 2021 and 2022 are relatively similar.

The results of the wage regression suggest that women experience a wage penalty, which is in line with the gender pay gap estimations for Estonia. Meriküll & Tverdostup (2023) discuss that there has been certain intertia to the wage gap after Estonia transitioned from communism to capitalism and estimated the gap to have declined from 34% in 1989 to around 19% in 2019. SA estimated the wage gap to be 14.9% in 2021 and 17.7% in 2022 with highest gap in financial and insurance activities (Statistics Estonia, table PA5355). According to the regression, being a woman compared to being a man results in 22% lower wage, holding other factors fixed. The results also suggest the variables age and experience have a non-linear relationship with wage.

Additionally, basic education, secondary and tertiary education impact wage positively compared to having pre-school education. The effect of basic education compared to pre-school education is statistically significant for 2021 but not for 2022. This could be due to fact that basic education serves as the mandatory minimum of general education requirement in Estonia and individuals who have only the pre-school education are not differentiated by the employers. Another explanation could be that the base group is relatively small and therefore the effect of additional years in basic education is underestimated. The number of observations falling into binary and categorical variable groups is presented in Appendix 1. Interestingly, professionals seem to have higher wage expectancy compared to managers. The model suggests that professionals earn 10-12% higher salary compared to managers. The wage expectancy of other occupation groups compared to managers is negative, with elementary occupations earning 39-41% less than managers, holding other variables constant.

As far as NACE dummy is concerned, working in the manufacturing sector results ceteris paribus in higher wages than others, the lowest being construction. Regional dummies also present the expected results, with the highest wages for Northern Estonia and lowest for North-Eastern Estonia. Holding other factors fixed, working in North-Eastern Estonia compared to working in Northern Estonia results in 23% lower wage. This is in line with SA estimations as average gross salary for Harju county in 2021 was from 1593 to 1730 euros (2022: 1751-1946 euros), compared to 1102-1184 euros (2022: 1176-1334 euros) in Ida-Viru county (Statistics Estonia, table PA117).

Next, the distribution of residuals is inspected. The bottom 10% of the residual distribution is considered suspiciously low-paid meaning that the wage regression, taking into account all individual characteristics, estimated the wage of the individual to be significantly higher than the actual wage reported to the tax authorities. This could indicate that the employee could be receiving a part of its salary in an envelope to evade labour taxes. Table 9 presents the total number of firms (employers), the firms for which one employee was present in the bottom 10% of the residual distribution. If looking at 2021 or 2022 individually, the wage regression suggests for both years that nearly 50% of the companies have at least one employee who receives a salary drastically lower than the wage regression would estimate. However, if looking at companies for which 50% or more of the employees fell into the bottom 10% of the residual distribution, the result is slightly above 30% for both, 2021 and 2022. What is more, 3552 companies have 50% or more of employees in bottom 10% of the residual distribution for two consecutive years.

Table 9. Share of tax evading firms

Year	Total firms	One employee in bottom 10% of residual distribution	Bottom 10% firms %	≥50% of employees in bottom 10% of the residual distribution	Tax evading firms %
2021	22 672	11 207	49.4%	5997	26.45%
2022	21 463	10 488	48.9%	5666	26.40%

Source: Author's calculations

After investigating the individuals and firms in the bottom 10% of the residual distribution, the pattern shows that 9.8% of the identified evasion is due to individuals being paid the minimum wage, whereas the wage regression suggests a much higher salary. Tonin (2011) suggests that tax evasion among employees is concentrated at the lower end of productivity distribution and lower wages. More than a third earn less than 110% of the minimum wage and the administrative average salaries falling into the distribution do not exceed 2000 euros. What is more, 75.2% of the individuals falling into the bottom are male. Out of all individuals working in the construction sector, 14.3% fall into the bottom, followed by 12.4% in transportation and storage. According to the survey by EKI (Josing, 2016), the characteritics of an individual receiving envelope wages are young, male, lower educational attainment, lower salary, living outside of the city and working in smaller firms active in construction, service or agricultural sectors.

3.1.2. Logistic regression

After combining the two subsets of firms for which the binary classification assumption of tax compliant and tax evading was done and merging it with firm's financial data, logistic regression is used to model the relationship between binary outcome and predictor variables. Additonally, the out-of-sample probability of tax evasion is predicted for companies with unknown classification.

After the data availability for all necessary variables was assessed for 2021, a subset of 23 791 (2022: 24 314) firms remain in the dataset, of which 5195 (2022: 4860) firms are available for training and testing purposes. From the latter, 529 (2022: 520) firms are presumed to be tax compliant and 4666 (2022: 4340) are presumed to be tax evading firms. The subset of firms is randomly split into training and test set to model the outcome of labour tax evasion, training subset accounting to 80% of the observations and test subset accounting to 20% of the observations. Due to fact that one class is significantly more prevalent in the training set, weighted loss function is applied. The logistic regression model coefficients represent the log odds of the binary outcome changing by one unit for each unit increase in the predictor variable, ceteris paribus. The logit model only makes it possible to determine the direction of the effect of the independent variables on the dependent variable, so marginal effects are also calculated to better understand the impact. The marginal effects of the regression are presented in table 10.

Table 10. Marginal effects

	Tax evading 2021	Tax evading 2022
Size	-0.015*** (0.001)	-0.017*** (0.001)
Construction	0.114*** (0.016)	0.196*** (0.017)
Wholesale and retail trade	0.043* (0.015)	0.057*** (0.016)
Transportation and storage	0.166*** (0.017)	0.183*** (0.018)
Turnover	-0.097*** (0.004)	-0.077*** (0.004)
Debt to assets	-0.035*** (0.006)	-0.023* (0.007)
Short-term debt to assets	-0.001 (0.001)	0.011** (0.004)
Cash to assets	-0.067 (0.022)	-0.041 (0.021)
Turnover to assets	0.016*** (0.002)	0.005** (0.001)
COGS to turnover	0.012* (0.006)	-0.028*** (0.005)
Observations	4 156	3 888

Note: Results are based on Eq. 2. Significance level * p < 0.05, ** p < 0.01, *** p < 0.001. Stardard errors in parentheses.

The results of the logistic regression suggest that size of the company is negatively associated with the probability to evade taxes, which is in line with previous papers by Beneish (1999) and Putniņš

& Sauka (2015). For 2021, an one-unit increase in the size of the company (one additional person employed) is associated with the probability of tax evasion decreasing by 0.015 (2022: 0.017) on average, holding other variables constant. What is more, compared to manufacturing, operating in construction, wholesale and retail trade or transportation and storage results in higher probability to be engaged in tax evasion. In 2021, transportation and storage present higher probability to be engaged in labour tax evasion than construction, but in 2022, construction is showing the highest impact.

The results of the logistic regression also suggest that higher turnover indicates a decrease in the predicted probability of tax evasion. This is in line with previous studies by Putniņš & Sauka (2015), Abdixhiku *et al.* (2017) and Benkovskis & Fadejeva (2022). If the turnover exceeds certain thresholds, a firm is required to meet additional reporting obligations as well as officiate audits. Therefore, the higher the turnover, the higher is the propensity to comply to rules. Benkovskis & Fadejeva (2022) also find that higher debt to assets ratio and short-term debt to assets ratio is associated with more probable tax evasion, however, the results on the Estonian dataset suggest otherwise. Beneish (1999) also reports that evading firms tend to be more leveraged. Hajek & Henriques (2017) suggest that higher leverage incentivises the firm to boost financial performance. The opposing result on Estonian data could be due to inclusion of a large number of micro enterprises in the analysis as smaller companies are less likely to have substantial loan liabilities on their balance sheet. The same applies to short-term liabilities.

Increase in cash to assets ratio results in decrease in the predicted probability of tax evasion, however the variable is not statistically significant at the 5% level for 2021 and 2022. Benkovskis & Fadejeva (2022) also report a negative relationship, suggesting that firms with relatively high cash holdings are less likely to be involved in labour tax evasion. Regarding turnover to assets, the coefficient is positive and statistically significant for both years (on 0.1% level for 2021 and on 1% level for 2022). The higher probability of labour tax evasion for firms with a higher turnover to assets ratio could be due to overreporting of revenue in tax evading firms, a line of reasoning supported by Benkovskis & Fadejeva (2022). The coefficients for COGS to turnover suggest positive effect on the predicted probability of labour tax evasion for 2021 and negative for 2022, however not statistically significant for 2021. Benkovskis & Fadejeva (2022) also report only marginal statistical significance for intermediate inputs to turnover.

The logistic regression model is tested on test set and the goodness of the model is evaluated. The probability threshold is set to 65%. Table 11 presents the confusion matrix to evaluate the performance of a classification model on a test sample using a probability threshold 65% for 2021 and 2022. For 2021, the model correctly predicted 568 instances (2022: 510) as belonging to class 1 (true positives) and incorrectly predicted 16 instances (2022: 13) as belonging to class 1 (false positives), when they actually belong to class 0. The model correctly predicted 92 instances (2022: 94) as belonging to class 0 (true negatives) and incorrectly predicted 363 instances (2022: 355) as belonging to class 0 when they actually belong to class 1 (false negatives).

Table 11. Confusion matrix

	2021		2022	
	False	True	False	True
0	92	16	94	13
1	363	568	355	510

Source: Author's calculations

To measure the prediction performance of the logistic regression model on test data, accuracy, TP rate, FP rate, TN rate, FN rate, F-score and AUC are presented in Table 12.

Table 12. Prediction performance ratios

	Accuracy	TP rate	FP rate	TN rate	FN rate	F-score	AUC
2021	63.52	61.01	14.81	85.19	38.99	0.75	0.84
2022	62.14	58.96	12.15	87.85	41.04	0.73	0.88

Source: Author's calculations

The prediction performance of the logistic regression model suggests that for 2021, 61% (2022: 58%) of the companies are correctly predicted as fraudulent companies and 85% (2022: 88%) of companies are correctly predicted as non-fraudulent. This is in line with previous studies, suggesting slightly lower TP rate for logit and probit compared to other machine learning approaches. Cecchini *et al.* (2010) compared the prediction performance of different approaches on their dataset and found that logistic regression following paper by Dechow *et al.* (2009) correctly predicted 64.5% of the fraudulent companies and 66.4% of non-fraudulent companies while support vector machines using the financial kernel correctly predicted 80.0% of the fraudulent and 90.6% of the non-fraudulent companies, being therefore better at predicting than logistic regression model.

However, as the recall is lower than specifity, the model could be better at classifying compliant firms than non-compliant. The F-score indicates that the classifier has archieved moderate precision and TP rate, AUC indicates a relatively good performance of the classifier. Figure 3 presents the ROC curve for 2021 on the left-hand side and for 2022 on the right-hand side. The ROC curve shows the trade-off between sensitivity and specificity and the closer the curve is to the top-left corner, the better the performance of the classifier. Therefore, the classifier performs better compared to a random classifier representing the 45-degree line on the graph.

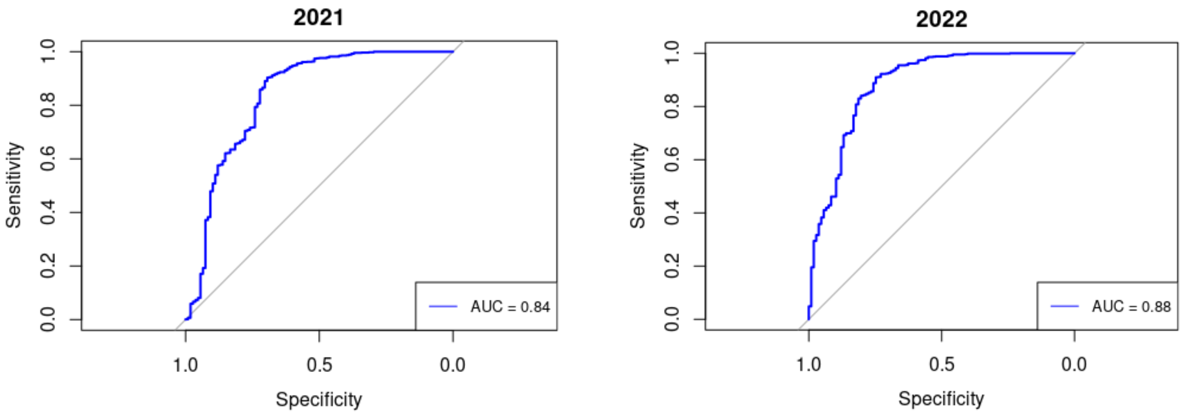

Figure 3. ROC curve
Source: Compiled by author
Note: This figure displays the ROC curve based on the test set.

Thirdly, the logit model is used to predict the out-of-sample probability of being engaged in tax evasion for all firms for which the labour tax evasion was unknown (based on Eq. 9). For the given threshold of 65%, the model classifies 9534 firms (2022: 10 536) as tax evading and 9062 firms (2022: 8918) as tax compliant for 2022. Gavoille & Zasova (2023) find that 37% of the firms are engaged in tax evasion, compared to Benkovskis & Fadejeva (2022) who report that 75-80% of Latvia's firms are evading labour taxes. The percentage of companies classified as tax evading is 51% (2022: 54%) which is in between the estimates of previous studies.

To understand the distribution of tax evading firms better, the evasion according to size and sector of the firm is shown for 2022. Table 13 presents the share of evading firms by size and the share of employees working in evading firms from the size group. The results reveal that tax evasion is more prevalent in smaller enterprises, peaking to 84% in firms with one employee only.

Table 13. Evading firms by firm size in 2022

Size class	Number of evading	Share of evading	Share of employees
1 employee	5765	84.31%	84.31%
2 to 4 employees	4071	62.75%	60.30%
5 to 9 employees	693	23.15%	20.50%
10 to 19 employees	7	0.42%	0.37%
20 or more employees	0	0%	0%

Source: Author's calculations

Table 14 presents the share of evading firms by sector and the share of employees working in evading firms from the size group. The share of evading firms is highest in construction and transportation and storage. However, the share of employees working in tax evading firms is much smaller for transportation and storage compared to construction, the latter accounting to nearly 36%. The higher share of employees working in tax evading firms, as can be seen in construction sector, is in line with previous findings that highly competitive and cash-intensive industries tend to have higher share of labour tax evasion.

Table 14. Evading firms by sectors in 2022

Industry	Number of evading	Share of evading	Share of employees
Manufacturing	981	28.43%	2.54%
Construction	4752	77.03%	35.50%
Wholesale and retail trade	3027	41.66%	7.61%
Transportation and storage	1776	69.16%	16.50%

Source: Author's calculations

Therefore, despite the arguably subjective choice of probability threshold, the results are in line with previous findings by Putniņš & Sauka (2015), Gavoille & Zasova (2021), Benkovskis & Fadejeva (2022).

3.2. Robustness checks

To ensure the reliability and validity of the regression results, robustness checks were conducted for both the wage regression and logistic regression.

3.2.1. Wage regression

The robustness checks carried out for wage regression were exclusion of a variable and cross-validation. Firstly, the educational attainment variable was excluded from the model as there were missing values for approximately 25% of the observations. Even though the distribution of individuals with missing educational attainment seemed random and not systematic, it could still cause issues and influence wage regression results. The wage regression without educational attainment showed relatively similar results compared to the final model and is presented in Appendix 4. The signs of the coefficients remain unchanged, however, the explanatory performance of the model decreases. This is expected, as education is an important factor impacting the job and salary prospects (Mincer, 1975).

Secondly, cross-validation was implemented on the final model. The data was randomly partitioned into training (70%) and test (30%) sets, and an OLS regression model was fitted using the training data. Afterwards, the values of the dependent variable were predicted using the fitted model and testing data. The performance was evaluated by calculating Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R^2. The results from five random partitions are presented in table 15.

Table 15. Cross-validation results for OLS model

Iteration	MSE	RMSE	R^2
1	0.170	0.423	0.316
2	0.169	0.421	0.314
3	0.169	0.411	0.316
4	0.170	0.412	0.315
5	0.170	0.412	0.316
Average	0.170	0.416	0.315

Source: Author's calculations
Note: MSE, Mean Squared Error. RMSE, Root Mean Squared Error. R^2, R-squared.

The values suggest that the model has consistent MSE, RMSE and R^2 values across iterations. The deviation from average is relatively low, and the proportion of variance in the dependent variable that is explained by the independent variables is similar to the final model. Therefore, the cross-validation results suggest that the OLS model is stable.

3.2.2. Logistic regression

To validate and assess the model's generalizability and performance on unseen data, cross-validation was applied. Over all samples, 10-fold cross-validation suggests an accuracy of 92.57% for 2021 (2022: 92.30%). The kappa statistic measure value of 0.45 for 2021 (2022: 0.44) suggests a moderate agreement beyond chance. Therefore, the model has a good overall performance, but the predictive performance could be improved by, for example, adjusting model specifications.

As logistic regression is sensitive to probability threshold selection, the probability threshold of 84% following Benkovskis & Fadejeva (2022) is applied and the results are compared. The probability threshold of 65% presented conservative results, therefore it is expected that higher probability threshold results in a lower TP rate. The confusion matrix in table 16 presents results for the comparative analysis. The model is good at finding compliant companies but performs poorly in finding non-compliant companies.

Table 16. Confusion matrix

	2021		2022	
	False	True	False	True
0	102	6	106	1
1	774	157	721	144

Source: Author's calculations

Therefore, increasing the probability threshold to 84% following Benkovskis & Fadejeva (2022) results in lower accuracy and more conservative results (table 17). The TN rate for 2021 in this case would be 94% (2022: 99%), but TP rate is 18% (2022: 17%), which is low. The relatively high AUC shows that the model is able to distinguish between true and false type well, but the chosen threshold might be inaccurate.

Table 17. Prediction performance ratios

	Accuracy	TP rate	FP rate	TN rate	FN rate	F-score	AUC
2021	24.93	17.86	5.56	94.44	83.14	0.29	0.84
2022	25.72	16.65	0.93	99.07	83.35	0.29	0.88

Source: Author's calculations

As a result, probability threshold of 84% results in 15 466 firms (2022: 16 021) classified as tax compliant and 3130 firms (2022: 3433) classified as tax evading. Therefore, 17% of firms (2022: 18%) firms are classified as evading labour taxes.

3.3. Discussion

The results of the analysis suggest that larger companies (in terms of number of employees) are less likely to evade labour taxes and labour tax evasion is highest among self-employed. The results regarding financial ratios suggest that the probability to evade labour taxes decreases as turnover, debt to assets, or cost of goods sold to assets increases. Conversely, the predicted probability to evade labour taxes increases as short-term debt to assets or turnover to assets increases. What is more, labour tax evasion is more prevalent in construction sector. The prediction of out-of-sample probability of being engaged in labour tax evasion suggested that in 2021 51% of companies (54% in 2022) are classified as labour tax evading in the four NACE sectors under investigation.

Even though using administrative data from SA has advantages of being representative, there are also some limitations to be considered in the interpretation and discussion of the results. Administrative data on wages is only available for years 2021 and 2022, which on the positive side is the most recent data but on the negative side limits the time frame to be used for analysis purposes. More comprehensive results could be obtained if investigating longer time series and implementing a panel data approach.

From the methodological perspective, logistic regression is suitable if the main focus is not solely on prediction power but rather on interpretation of the results. Cecchini *et al.* (2010) also compare the performance of other fraud detection methods applied in previous papers and find that support vector machines using the financial kernel performs best for recall. Gavoille & Zasova (2023) apply gradient boosting decision trees for better prediction performance. However, these models offer less opportunities for interpretation and could be used if predictive power is of importance.

It is also important to acknowledge that the wage regression model and logistic regression model do not capture all the individual or firm-level heterogeneity that could explain the decision to be engaged in tax evasion, but only factors available in the database. There are many other unobserved factors, such as the personal views of the key personnel and employees, overall company culture, understanding the tax system and experience with tax authorities that can have significant effect on the decision of tax evasion or tax compliance.

Combining together survey results, for example evasion behaviour in company managers (or personal views of employees) with financial and non-financial variables of firms could give a more comprehensible overview of the topic. Finally, using administrative data combined together with surveys such as HBS, offering information on consumption for validating the results of the labour tax evasion, as relatively smaller income compared to consumption can refer to envelope salary. Hajek & Henriques (2017) note that text in annual reports can be used to detect fraudulent firms. They found that using linguistic variables (frequency count of different positive, negative, uncertain or other words) in addition to financial variables improved the performance of logistic regression compared to including only financial variables, resulting in improved accuracy, higher true positive rate and true negative rate.

The findings may also suggest that measures to motivate labour tax compliance for self-employed should be improved. Personal views on taxation largely determine the decision to comply or evade, however, nudging techniques could improve tax compliance among self-employed. Therefore, the effect of tax compliance ratings displayed to taxpayers by EMTA on wages could be further investigated to see, whether the tool has been effective in increasing labour tax compliance. Additionally, it would be important to measure the revenue lost due to non-compliance, for example, by evaluating the volume of unreported wage, following Benkovskis & Fadejeva (2022). What is more, it could be investigated how the minimum wage hike effects the compliant and non-compliant firms differently, following Gavoille & Zasova (2023).

CONCLUSION

The aim of this thesis was to analyse the relationship between firm-level financial and non-financial indicators and the probability of a firm being engaged in labour tax evasion. The analysis aimed to answer two main research questions:

1) Which firm-level predictors contribute to the probability of being engaged in labour tax evasion?
2) What is the proportion of labour tax evading companies?

To answer the research questions raised, a database comprising three merged datasets was utilized including administrative data on wages, population as well as annual reports. The analysis employs Mincer wage regression to find the firms paying "suspiciously low wages" to employees and logistic regression to analyse the relationship between different firm-level indicators and the probability of being a tax evader.

The analysis findings indicate that larger companies, as measured by their number of employees, exhibit lower tendencies to evade labor taxes. Conversely, labor tax evasion is most prevalent among the self-employed, reaching 84% in 2022. Consequently, the results support the need to implement strategies aimed at fostering labor tax compliance among the self-employed. While individual attitudes towards taxation significantly influence compliance decisions, employing various cost-effective nudging techniques could enhance tax compliance among the self-employed. Therefore, further investigation into the impact of tax compliance ratings displayed to taxpayers by EMTA on wages is required to ascertain the tool's effectiveness in increasing labor tax compliance. Additionally, it is crucial to quantify the revenue loss resulting from non-compliance, such as by assessing the volume of unreported wages, following the methodology outlined by Benkovskis & Fadejeva (2022).

What is more, labour tax evasion is more prevalent in the construction sector. The results for 2022 suggest that 77% of companies active in the construction sector are engaged in labour tax evasion; however they employ 36% of the employees in the sector. Therefore, more attention should be paid to the construction sector, which is by nature competitive and cash-intensive. The prediction of the out-of-sample probability of being engaged in labour tax evasion suggested that in 2021 51% of companies (2022: 54%) are classified as labour tax evading in the four NACE sectors under investigation.

The results regarding financial ratios suggest that the probability of evading labour taxes decreases as turnover, debt to assets, or cost of goods sold to assets increases. Conversely, the predicted probability of evading labour taxes increases as short-term debt to assets or turnover to assets increases.

There are opportunities for improvements as well as for future research on this topic. The results suggest that a more refined model is required or another machine learning approach should be applied to improve the prediction performance of the model. Additionally, the topic could be further developed to examine the dynamics and patterns in employee salaries, changes in minimum wage and financial statements. For example, it could be investigated how the minimum wage hike affects the compliant and non-compliant firms differently, following Gavoille & Zasova (2023). Lastly, besides the extensive margin, the intensive margin of labour tax evasion could be further investigated to understand the features of companies with unreported employees.

KOKKUVÕTE

TÖÖJÕUMAKSUDEST KÕRVALEHOIDUMIST ENNUSTAVAD TEGURID ETTEVÕTTE TASANDIL

Alice Mikk

Käesoleva magistritöö eesmärk oli hinnata erinevate finants- ja mittefinantsnäitajate seost tööjõumaksudest kõrvale hoidumise tõenäosusega ettevõtte tasandil. Analüüsi käigus keskenduti põhiliselt kahele uurimisküsimusele:

1) Millised ettevõttespetsiifilised tegurid panustavad tööjõumaksudest kõrvale hoidumise tõenäosuse suurenemisse?
2) Kui suur on tööjõumaksudest kõrvale hoiduvate ettevõtete osakaal?

Küsimustele vastamiseks kasutati kolme ühendatud andmestikku, mis hõlmas endas nii palkade, rahvastiku kui ka ettevõtete majandusaasta aruannete andmeid. Analüüsis kasutati Minceri palgaregressiooni, et leida ettevõtted, kes maksavad töötajatele "kahtlaselt madalat palka" ning logistilist regressiooni, et analüüsida erinevate ettevõtte finants- ja mittefinantsnäitajate seost palgamaksudest kõrvale hoidumise tõenäosusega.

Analüüsi tulemused viitavad sellele, et suuremad ettevõtted (töötajate arvu mõistes) hoiduvad väiksema tõenäosusega tööjõumaksude tasumisest. Tööjõumaksude maksmisest kõrvale hoidumine on kõrgeim ühe töötajaga ettevõtete hulgas, moodustades 84% kõigist ühe töötajaga ettevõtetest. Tulemused ilmestavad, et tööjõumaksude laekumise parandamiseks tuleb just mõelda ühe töötajaga firmade võimalikult kuluefektiivsele motiveerimisele, näiteks nügimismeetodeid (inglise keeles *nudging*) kasutades. Näiteks võiks analüüsida EMTA poolt maksumaksjatele kuvatava maksukuulekuse reitingu mõju palkade deklareerimisele ja tööjõumaksudele. Antud analüüsi laiendusena võiks leida saamata jäänud tööjõumaksude ulatuse kõigi ettevõtete osas järgides Benkovskis & Fadejeva (2022) lähenemist.

Tööjõumaksudest hoidumine on kõige prevalentsem ehitussektoris, moodustades analüüsi tulemuste põhjal 2022. aastal koguni 77% kõigist ehitussektori ettevõtetest ning pakkudes tööd 36% ehitussektori tööjõule. Seetõttu tuleb tööjõumaksudest hoidumise vaates rohkem tähelepanu pöörata ehitussektorile kui kõrge konkurentsiga ja sularahaintensiivsele sektorile. Magistritöö tulemusel hinnati, et uuritud neljal tegevusalal kokku hoidus 2021. aastal tööjõumaksudest 51% ettevõtetest ning 2022. aastal 54% ettevõtetest.

Finantssuhtarvude ja tööjõumaksudest hoidumise tõenäosuse uurimisel selgus, et tööjõumaksudest kõrvalehoidumise tõenäosus väheneb käibe, võlakordaja ja müüdud toodangu (kaupade, teenuste) kulu suhe varadesse kasvades. Vastupidiselt kasvab tõenäosus tööjõumaksudest kõrvale hoidumiseks kui suureneb lühiajalise võla kordaja või varade käibekordaja.

Antud analüüsi puhul on mitmeid võimalusi täiustusteks ja edaspidiseks uurimiseks. Tulemused viitavad ka sellele, et mudeli ennustusvõime parandamiseks on vaja mudeli spetsifikatsioone täpsustada või rakendada mõnd teist masinõppe lähenemist. Lisaks võiks analüüsi edasiarendusena uurida töötajate palkade, miinimumpalga muutuste ja raamatupidamisaruannete dünaamikat ja mustreid. Näiteks võiks Gavoille & Zasova (2023) lähenemist järgides hinnata, kui palju erineb miinimupalga tõusu mõju maksukuulekate ja maksudest hoiduvate ettevõtete finantsnäitajatele. Täiendavalt võiks analüüsida ka tööjõumaksudest kõrvale hoidumist ettevõtetes registreerimata töötajate vaatest.

LIST OF REFERENCES

Abdixhiku, L., Krasniqi, B., Pugh & G., Hashi, I. (2017). Firm-level determinants of tax evasion in transition economies. *Economic Systems, 41*(3), 354-366. https://doi.org/10.1016/j.ecosys.2016.12.004

Allingham, M. G. & Sandmo, A. (1972). Income tax evasion: a theoretical analysis. *Journal of Public Economics, 1*(3-4), 323-338.

Alm, J. & Malézieux, A. (2021). 40 years of tax evasion games: a meta-analysis. *Experimental Economics, 24*(3), 699-750.

Alm, J. (1988). Compliance Costs and the Tax Avoidance-Tax Evasion Decision. *Public Finance Quarterly, 16*(1), 31-66. https://doi.org/10.1177/109114218801600102

Arsić, M., Arandarenko, M., Radulović, B., Ranđelović, S. & Janković, I. (2015). Causes of the Shadow Economy. Formalizing the Shadow Economy in Serbia. *Contributions to Economics*, 21-56. https://doi.org/10.1007/978-3-319-13437-6_4

Benkovskis, K. & Fadejeva, L. (2022). Chasing the Shadow: The Evaluation of Unreported Wage Payments in Latvia. *Bank of Latvia Working Paper,* 2022/01.

Bobbio, E. (2017). Tax Evasion, Firm Dynamics and Growth. *Bank of Italy Occasional Paper, 357.* http://dx.doi.org/10.2139/ssrn.2910376

Braguinsky, S. & Mityakov, S. (2015). Foreign corporations and the culture of transparency: Evidence from Russian administrative data. *Journal of Financial Economics 117*(1), 139–164. https://doi.org/10.1016/j.jfineco.2013.02.016

Brennan, G. & Buchanan, J. (1980). The Power to Tax: Analytical Foundations of a Fiscal Constitution. *Cambridge University Press.*

DeBacker, J., Heim, B. T. & Tran, A. (2015). Importing corruption culture from overseas: Evidence from corporate tax evasion in the United States. *Journal of Financial Economics, 117*(1), 122-138. https://doi.org/10.1016/j.jfineco.2012.11.009.

Dom, R., Davenport, A. & Prichard, S. R. (2022). Innovations in Tax Compliance: Building Trust, Navigating Politics, and Tailoring Reform. *World Bank.*

Eurobarometer (2020). Undeclared work in the European Union. *European Commission, Technical report, 498.*

Gavoille, N. & A. Zasova (2021). Foreign ownership and labor tax evasion: Evidence from Latvia. *Economics Letters 207*(C), 1–4. https://doi.org/10.1016/j.econlet.2021.110030.

Gavoille, N. & Zasova, A. (2023). What we pay in the shadows: Labor tax evasion, minimum wage hike and employment, *Journal of Public Economics, 228*. https://doi.org/10.1016/j.jpubeco.2023.105027

Gujarati, D. N. (2003). *Basic econometrics* (4th ed). McGraw Hill.

Hajek, P. & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud – a comparative study of machine learning methods. *Knowledge-Based Systems, 128*, 139–152. https://doi.org/10.1016/j.knosys.2017.05.001

Harmon, C. & Walker, I. (1995). Estimates of the economic return to schooling for the United Kingdom. *American Economic Review, 85*(5), 1278-1286.

Hashimzade, N., Myles, G. D. & Tran-Nam, B. (2013). Applications of behavioural economics to tax evasion. *Journal of Economic Surveys, 27*(5). http://dx.doi.org/10.1111/j.1467-6419.2012.00733.x

Josing, M. (2016). Varimajanduse trendid. *Eesti Konjunktuuriinstituut.*

Kenny, L. W. (2006). Tax Systems in the World: An Empirical Investigation into the Importance of Tax Bases, Administration Costs, Scale and Political Regime. *International Tax and Public Finance, 13,* 181-215. https://doi.org/10.1007/s10797-006-3564-7

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., & Saez, E. (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica, 79*(3), 651-692. https://doi.org/10.3982/ECTA9113

Kukk, M. & Staehr, K. (2014). Income underreporting by households with business income: evidence from Estonia. *Post-Communist Economies, 26*(2), 257-276. https://doi.org/10.1080/14631377.2014.904110

Kukk, M., Paulus, A. & Staehr, K. (2019). Cheating in Europe: underreporting of self-employment income in comparative perspective. *International Tax and Public Finance, 27,* 363-390. https://doi.org/10.1007/s10797-019-09562-9

Lepassar, A. (2024). Maksukäitumise hinnangud MTA e-teenus ettevõtjale “virtuaalne audiitor”. *Maksu- ja Tolliamet.*

Levenko, N. & Staehr, K. (2023). Self-reported tax compliance in post-transition Estonia. *Economic Systems, 47(3).* https://doi.org/10.1016/j.ecosys.2022.101047

Levi, M. (1988). Of Rule and Revenue. University of California Press.

Lippert, O. & Walker, M. (1997). The Underground Economy: Global Evidence of Its Size and Impact. The Fraser Institute, Vancouver.

Mendgen, A. (2023). International Tax Competitiveness Index 2023. *Tax Foundation Center for Global Tax Policy, 10.*

Meriküll, J. & Tverdostup, M. (2023). The gap that survived the transition: The gender wage gap in Estonia over three decades. *Economic Systems, 47*(3). https://doi.org/10.1016/j.ecosys.2023.101127

Mincer, J. (1975). Education, experience and the distribution of earnings and employment: An overview. *National Bureau of Economic Research*, 71–93.

Musgrave, R.A. (1959) The Theory of Public Finance. McGraw Hill, New York.

Müürsepp, R. (2015, October 12). Kui suur on Eesti varimajandus? Used 21.02.2024. https://www.stat.ee/et/uudised/2015/10/12/kui-suur-on-eesti-varimajandus

Pickhardt, M. & Prinz, A. (2014). Behavioural dynamics of tax evasion – A survey. *Journal of Economic Psychology, 40,* 1-19. https://doi.org/10.1016/j.joep.2013.08.006

Pissarides, C. A. & Weber, G. (1989). An expenditure-based estimate of Britain's black economy, *Journal of Public Economics, 39*(1), 17-32.

Privitera, A., Enachescu, J., Kirchler, E. & Hartmann, A. J. (2021). Emotions in Tax Related Situations Shape Compliance Intentions: A Comparison between Austria and Italy. *Journal of Behavioural and Experimental Economics, 92.* https://doi.org/10.1016/j.socec.2021.101698

Putniņš, T. J. & Sauka, A. (2015). Measuring the shadow economy using company managers. *Journal of Comparative Economics, 43*(2), 471-490. https://doi.org/10.1016/j.jce.2014.04.001

Putniņš, T. J. & Sauka, A. (2023). Shadow Economy Index for the Baltic Countries 2009-2022. *13th Annual Shadow Economy Conference.*

Rahandusministeerium (2024, April 3). *Maksud.* Used 06.03.2024. https://www.fin.ee/riigi-rahandus-ja-maksud/maksu-ja-tollipoliitika/maksud

Schneider, F. (2016). Estimating the Size of Shadow Economies of Highly-developed Countries: Selected New Results. *Leibniz Institute for Economic Research at the University of Munich, 14*(4), 44-53.

Schumpeter, J. A. (1991). The Economics and Sociology of Capitalism. *Princeton University Press.* https://doi.org/10.2307/j.ctv173f01t

Slemrod, J. (2007). Cheating Ourselves: The Economics of Tax Evasion. *Journal of Economic Perspectives, 21*(1), 25-48.

Smeaton, D. & McKay, S. (2003). Working after State Pension Age: Quantitative Analysis. *Department for Work and Pensions, Research Report 182.*

Statistics Estonia. (2023). PA117: Average monthly gross wages (salaries), median and number of employees by Indicator, County and Reference period. Used 23.03.2024. https://andmed.stat.ee/en/stat/majandus__palk-ja-toojeukulu__palk__luhiajastatistika/PA117

Statistics Estonia. (2023). PA5335: Gender pay gap by economic activity (EMTAK 2008), October. Used 29.03.2024. https://andmed.stat.ee/en/stat/majandus__rahandus__valitsemissektori-rahandus__valitsemissektori-tulud-kulud/RR057

Statistics Estonia. (2023). RR057: Consolidated revenue and expenditure of general government (quarters) (ESA 2010). Used 21.02.2024. https://andmed.stat.ee/en/stat/majandus__rahandus__valitsemissektori-rahandus__valitsemissektori-tulud-kulud/RR057

Tafenau, E., Herwartz, H. & Schneider, F. (2010). Regional Estimates of the Shadow Economy in Europe. *International Economic Journal, 24*(4), 629-636, https://doi.org/10.1080/10168737.2010.526010

Tonin, M. (2011). Minimum wage and tax evasion: Theory and evidence. *Journal of Public Economics, 95*(11-12), 1635-1651. https://doi.org/10.1016/j.jpubeco.2011.04.005

Turu-uuringute AS (2023, October 5). *Varimajanduse uuring 2023*. Used 20.03.2024. https://www.emta.ee/sites/default/files/documents/2023-10/varimajanduse_uuring_2023.pdf

APPENDICES

Appendix 1. Binary and categorical variables in wage regression

	2021	2022
Binary/categorical variable	No of obs	No of obs
Gender		
Male	116 695	111 854
Female	76 096	74 718
Education		
Pre-school education	452	600
Basic education	41 904	40 915
Secondary education	93 430	89 907
Tertiary education (bachelor’s, master’s and doctoral)	57 005	55 150
NACE		
Manufacturing	79 043	76 863
Construction	30 414	28 389
Wholesale and retail trade; repair of motor vehicles and motorcycles	58 152	57 195
Transportation and storage	25 182	24 125
Occupation		
Managers	20 015	19 421
Professionals	12 352	12 236
Technicians and associate professionals	20 311	20 142
Clerical support workers	16 274	16 386
Services and sales workers	22 958	22 787
Skilled agricultural, forestry and fishery workers	198	191
Craft and related trades workers	48 730	46 068
Plant and machine operators and assemblers	35 951	34 282
Elementary occupations	16 002	15 059
Region		
Northern Estonia	98 959	95 832
Central Estonia	17 197	16 478
North-Eastern Estonia	15 507	15 427
Western Estonia	15 434	14 606
Southern Estonia	45 694	44 239

Source: Compiled by author

Appendix 2. List of variables in wage regression

Type	Variable	Description	Coding
Dependent	Wage	Average monthly salary in euros, calculated by sum of monthly salaries and divided by months employed	Logarithm
Independent	Gender	Individual's gender	1 – female 0 – male
Independent	Age	Age of the individual, calculated from date of birth	Continuous, in years
Independent	Highest education	Highest level of formal education obtained, based on International Standard Classification of Education (ISCED-11).	1 – Pre-school education 2 – Basic education 3 – secondary education 4 – tertiary education (bachelor's, master's and doctoral)
Independent	Work experience	Work experience of the individual calculated from date of employment	Continuous, in years
Independent	Field of activity	Level 2 (sub-major groups) of the Estonian Classification of Economic Activities (EMTAK) following Statistical Classification of Economic Activities (NACE)	1 – Manufacturing 2 – Construction 3 – Wholesale and retail trade; repair of motor vehicles and motorcycles 4 – Transportation and storage
Independent	Occupation	Level 1 (major groups) of the Classification of Occupations 2008, partly following International Standard Classification of Occupations (ISCO-08).	1 – Managers 2 – Professionals 3 – Technicians and associate professionals 4 – Clerical support workers 5 – Services and sales workers 6 – Skilled agricultural, forestry and fishery workers 7 – craft and related trades workers 8 – Plant and machine operators and assemblers 9 – Elementary occupations

Appendix 2 continued

Type	Variable	Description	Coding
Independent	Location of work	The Nomenclature of Territorial Units for Statistics (NUTS) is derived from Estonian Administrative and Settlement Classification (EHAK).	1 –Northern Estonia: Harju county 2 – Central Estonia: Järva, Lääne-Viru, and Rapla county 3 – North-Eastern Estonia: Ida-Viru county 4 – Western Estonia: Hiiu, Lääne, Saare, and Pärnu county 5 – Southern Estonia: Jõgeva, Põlva, Tartu, Valga, Viljandi, and Võru county

Source: Compiled by author

Appendix 3. List of variables in logistic regression

Type	Variable	Description	Coding
Dependent	Labour tax evasion	Based on the result from wage regression and parent company country.	1 – tax evasion 0 – no tax evasion
Independent	Firm size	Average number of employees reduced to full-time	Continuous
Independent	Field of activity	Level 2 (sub-major groups) of the Estonian Classification of Economic Activities (EMTAK) following Statistical Classification of Economic Activities (NACE)	1 – Manufacturing 2 – Construction 3 – Wholesale and retail trade; repair of motor vehicles and motorcycles 4 – Transportation and storage
Independent	Turnover	Sales	Logarithm
Independent	Debt to asset	Total liabilities/total assets	Continuous
Independent	Short-term debt to assets	Short-term liabilities/current assets	Continuous
Independent	Cash to assets	Cash/total assets	Continuous
Independent	Turnover to assets	Turnover/total assets	Continuous
Independent	COGS to turnover	Cost of goods sold/turnover	Continuous

Source: Compiled by author

Appendix 4. Robustness checks for wage regression

	ln(wage) 2021	ln(wage) 2022
Intercept	6.974***	7.165***
	(0.011)	(0.010)
Gender	-0.204***	-0.200***
	(0.002)	(0.002)
Age	0.029***	0.028***
	(0.001)	(0.001)
Age²/100	-0.028***	-0.027***
	(0.001)	(0.001)
Experience	0.026***	0.026***
	(0.001)	(0.001)
Experience²/100	-0.060***	-0.060***
	(0.002)	(0.002)
Construction	-0.114***	-0.113***
	(0.003)	(0.002)
Wholesale and retail trade	-0.066***	-0.048***
	(0.002)	(0.002)
Transportation and storage	-0.098***	-0.070***
	(0.003)	(0.003)
Central Estonia	-0.104***	-0.108***
	(0.003)	(0.003)
North-Eastern Estonia	-0.246***	-0.234***
	(0.003)	(0.003)
Western Estonia	-0.150***	-0.151***
	(0.003)	(0.003)
Southern Estonia	-0.116***	-0.120***
	(0.002)	(0.003)
Professionals	0.122***	0.099***
	(0.005)	(0.004)
Technicians and associate professionals	-0.062***	-0.094***
	(0.005)	(0.003)
Clerical support workers	-0.262***	-0.296***
	(0.005)	(0.004)
Services and sales workers	-0.435***	-0.473***
	(0.005)	(0.004)
Skilled agricultural, forestry and fishery workers	-0.516***	-0.362***
	(0.028)	(0.023)
Craft and related trades workers	-0.397***	-0.428***
	(0.004)	(0.003)
Plant and machine operators and assemblers	-0.399***	-0.442***
	(0.005)	(0.003)
Elementary occupations	-0.499***	-0.538***
	(0.005)	(0.004)
Observations	265 604	267 458
R²	0.302	0.324

Note: Significance level * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust stardard errors in parentheses.

Appendix 5. Non-exclusive licence

A non-exclusive licence for reproduction and publication of a graduation thesis[4]

I, Alice Mikk

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "Firm-level predictors of labour tax evasion", supervised by Karsten Staehr,

1.1 to be reproduced for the purpose of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;

1.2 to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.

2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.

3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

07.05.2024

[4] *The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period*