FREE NETWORK
#AcademicsStandWithUkraine

POLICY
BRIEF SERIES

Nicolas Gavoille, SSE Riga, BICEPS
Anna Zasova, BICEPS
May 2022

Detecting Labor Tax Evasion Using Administrative Data and Machine-Learning Techniques

Labor tax evasion is a major policy issue that is especially salient in transition and post-transition countries. In this brief, we use firm-level administrative data, tax authorities' audit data and machine learning techniques to detect firms likely to be involved in labor tax evasion in Latvia. First, we show that this approach could complement tax authorities' regular practices, increasing audit success rate by up to 35%. Second, we estimate that about 30% of firms operating in Latvia between 2013 and 2020 are likely to underreport the wage of (some of) their employees, with a slightly negative trend.

Acknowledgment: The authors gratefully acknowledge funding from the Latvian State Research Programme "Reducing the Shadow Economy to Ensure Sustainable Development of the Latvian State", Project "Researching the Shadow Economy in Latvia (RE:SHADE)"; project No VPP-FM-2020/1-0005.

Introduction

Tax evasion is a major policy issue that is especially salient in transition and post-transition countries. In particular, "envelop wage", i.e., an unofficial part of the wage paid in cash, is a widespread phenomenon in Eastern Europe (European Commission, 2020). Putnins and Sauka (2021) estimate that the share of unreported wages in Latvia amounts to more than 20%. Fighting labor tax evasion is a key objective of tax authorities, which face two main challenges. The first is to make the best use of their resources. Audits are costly, so the choice of firms to audit is crucial. The second challenge is to track the evolution of the prevalence of labor tax evasion. For this purpose, most of the existing literature relies on survey data.

In our forthcoming paper (Gavoille and Zasova, 2022), we propose a novel methodology aiming at detecting tax-evading firms, using administrative firm-level data, tax authorities' audit data and machine learning techniques.

This study provides two main contributions. First, this approach can help tax authorities to decide which firms to audit. Our results indicate that the audit success rate could increase by up to 20 percentage points, resulting in a 35% increase. Second, our methodology allows us to estimate the *share* of firms likely to be involved in labor tax evasion. To our knowledge, this paper is the first to provide such estimates, which are however of primary importance in guiding anti-tax evasion policy. We estimate that over the 2013-2020 period, about 30% of firms operating in Latvia are underreporting (at least some of) their workers' wages.

Methodology

The general idea of our approach is to train an algorithm to classify firms as either compliant or tax-evading based on observed firm characteristics. Tax evasion, like any financial manipulation, results in artifacts in the balance sheet. These artifacts may be invisible to the human eye, but machine learning algorithms can detect these systematic patterns. Such methods have been applied to corporate fraud detection (see for instance Cecchini et al. 2010, Ravisankar et al. 2011, West and Bhattacharya 2016).

The machine learning approach requires a subsample of firms for which we know the "true" firm behavior (i.e., tax-evading or compliant) in order to train the algorithm. For this purpose, we propose to use a dataset on tax audits provided by the Latvian State Revenue Service (SRS), which contains information about all personal income tax (PIT) and social security contributions (SSC) audits carried out by SRS during the period 2013-2020, including the outcome of the audit. The dataset also contains a set of firm characteristics and financial indicators, covering both audited and non-audited firms operating in Latvia (e.g., turnover, assets, profit). Assuming that auditors are highly likely to detect misconduct (e.g., wage underreporting) if present, audit outcomes provide information about a firm's tax compliance. Firms sanctioned with a penalty for, say, personal income tax fraud are involved in tax evasion, whereas audited-but-not-sanctioned firms can be assumed compliant. The algorithm learns how to disentangle the two types of firms based on the information contained in their balance sheets. Practically, we randomly split the sample of audited firms into two parts, the training and the testing subsamples. In short, we use the former to train the algorithm, and then evaluate its performance on the latter, i.e., on data that has not been used during the training stage. If showing satisfying performance on the training sample, we can then apply it to the whole universe of firms and obtain an estimate of the share of tax-evading firms.

In this study, we successively implement four algorithms that differ in the way they learn from the data: (1) Random Forest, (2) Gradient Boosting, (3) Neural Networks, and (4) Logit (for a review of machine learning methods, see Athey and Imbens, 2019). These four data mining techniques have previously been used in the literature on corporate fraud detection (see Ravisankar et al. 2011 for a survey). Each of these four algorithms has specific strengths and weaknesses, motivating the implementation and comparison of several approaches.

Results

Predictive Performance

Table 1 provides the out-of-sample performance of the four different algorithms. In other words, it shows how precise the algorithm is at classifying firms based on data that has not been included during the training stage. Accuracy is the percentage of firms correctly classified (i.e., the model prediction is consistent with the observed audit's outcome). In our sample, about 44% of audited firms are required to pay extra personal income tax and social security contributions. This implies that a naive approach predicting all firms to be evading would be 44% accurate. Similarly, a classification predicting all firms to be tax compliant would be correct in 56% of the cases. This latter number can be used as a benchmark to evaluate the performance of the algorithms. ROC-AUC (standing for Area Under the Curve – Receiver Operating Characteristics) is another widespread classification performance measure. It provides a measure of separability, i.e., how well is the model able to distinguish between the two types. This measure is bounded between 0 and 1, the closer to 1 the better the performance. A score above 0.8 can be considered largely satisfying.

Table 1. *Performance measures*

	Random Forest	Gradient Boosting	Neural Network	Logit
Accuracy	75.3%	72.0%	66.4%	63.3%
ROC-AUC	0.823	0.790	0.719	0.653

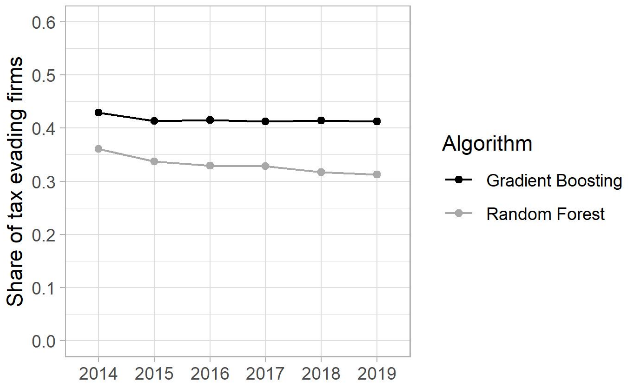
Random Forest is the algorithm providing the best out-of-sample performance, with more than 75% of the observations in the testing set correctly classified. Random Forest is also the best performing model according to the ROC-AUC measure, with performance slightly better than Gradient Boosting.

Our results imply that a naive benchmark prediction is outperformed by almost 20 percentage points by Random Forest and Gradient Boosting in terms of accuracy. It is important to emphasize that this improvement in performance is achieved using a relatively limited set of firm-level observable characteristics that we obtained from SRS (which is limited compared to what SRS has access to), and that mainly come from firms' balance sheets. This highlights the potential gain of using data-driven approaches for the selection of firms to audit in addition to the regular practices used by the fiscal authorities. It also suggests a promising path for further improvements, as in addition to this set of readily available information the SRS is likely to possess more detailed limited-access firm-level data.

Share of Tax-Evading Firms Over Time and Across NACE Sectors

We can now apply these algorithms to the whole universe of firms (i.e., to classify non-audited firms). Figure 1 shows the share of firms classified as tax-evading over the years 2014 to 2019 for our two preferred algorithms - Gradient Boosting and Random Forest. Random Forest (the best performing algorithm) predicts that 30-35% of firms are involved in tax evasion, Gradient Boosting predicts a slightly higher share (around 40%). Both algorithms, especially Random Forest, suggest a slight reduction in the share of tax-evading firms since 2014.

Figure 1. *Share of tax-evading firms over time*

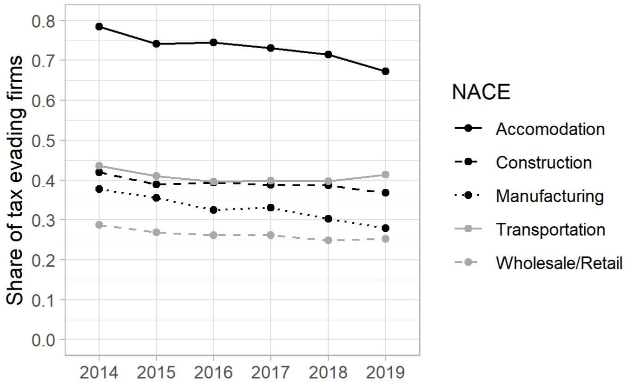


Source: authors' calculations

The identified reduction, however, does not necessarily imply that the overall share of unreported wages has declined. In fact, existing survey-based evidence (Putnins and Sauka, 2021) indicate that the size of the shadow economy as a share of GDP remained roughly constant over the 2013-2019 period, and that there was no reduction in the contribution of the "envelope wages". With our method, we are estimating the share of firms likely to be involved in labor tax evasion. Unlike the survey approach, our methodology does not allow measurement of tax-evasion intensity. In other words, the share of non-tax compliant firms may have decreased, but the size of the envelope may have increased in firms involved in this scheme.

Next, we disaggregate the share of tax-evading firms by NACE sector. Figure 2 displays the results obtained with Random Forest, our best performing algorithm.

Figure 2. *Share of tax-evading firms by NACE, based on Random Forest*



Source: authors' calculations

First, the sector where tax evasion is the most prevalent is the accommodation/food industry, where the predicted share of tax-evading firms is 70-80%. Second, our results indicate that the overall decrease in the share of firms likely to evade is not uniform. It is mostly driven by the accommodation/food and manufacturing sectors. Other sectors remain nearly flat. This highlights the fact that labor tax evasion varies both in levels and in changes across sectors.

Conclusion

We show that machine learning techniques can be successfully applied to administrative firm-level data to detect firms that are likely to be involved in (labor) tax evasion. Machine learning techniques can be used to improve the selection of firms to audit in order to maximize the probability to detect tax-evading firms, in addition to the regular practices already used by SRS. Our preferred algorithms – Random Forest and Gradient Boosting – outperform the naive benchmark classification by almost 20 percentage points, which is a substantial improvement. Once implemented, the use of these tools can improve the audit effectiveness at virtually no extra cost.

Our findings also suggest a promising path for further improvements in the application of such methods. The improvement in predictive power achieved by our proposed algorithm is attained by using a limited set of variables readily available from the firms' balance sheets. Given that SRS is likely to have access to more detailed firm-level information that cannot be provided to third parties, there is clear room for improving the performance of the algorithms by using such limited-access data.

References

Athey, Susan, and Guido Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics 11*: 685–725.

Cecchini, Mark, and Haldun Aytug, and Gary J. Koehler, and Praveen Pathak, 2010. "Detecting management fraud in public companies". *Management Science* 56, 1146-1160.

European Commission, 2020. "Undeclared Work in the European Union. Special Eurobarometer 498" (Report)

Gavoille, Nicolas and Anna Zasova, 2022. "Estimating labor tax evasion using tax audits and machine learning", *SSE Riga/BICEPS Research papers*, forthcoming.

Putnins, Talis, and Arnis Sauka, 2021. "Shadow Economy Index for the Baltic Countries 2009–2020" (Report), SSE Riga

Ravisankar, Pediredla, and Vadlamani Ravi, and Gundumalla Raghava Rao, and Indranil Bose, 2011. "Detection of financial statement fraud and feature selection using data mining techniques". Decision Support Systems, 50(2), 491-500.

West, Jarrod, and Maumita Bhattacharya, 2016. "Intelligent financial fraud detection: a comprehensive review". *Computers & security*, 57, 47-66

Nicolas Gavoille

Stockholm School of Economics in Riga (SSE Riga), Baltic International Centre for Economic Policy Studies (BICEPS)
Nicolas.Gavoille@sseriga.edu
www.sseriga.edu

Nicolas Gavoille is an Associate Professor at the Stockholm School of Economics in Riga and a Research Fellow at BICEPS. He holds a PhD in Economics from the University of Rennes 1, France, and is a member of the European Public Choice Society, of the French Economic Association and of the Condorcet Center for Political Economy. Nicolas' main research interests are in the field of public economics, labour economics, and political economy. He published articles in peer-reviewed journals such as the European Economic Review, the European Journal of Political Economy, IMF Economic Review, Public Choice, Economics Letters, the International Review of Law and Economics, and the Journal of Institutional and Theoretical Economics.

Anna Zasova

Baltic International Centre for Economic Policy Studies (BICEPS)
Anna@biceps.org
www.biceps.org

Anna Zasova is a Research Fellow at the Baltic International Centre for Economic Policy Studies (BICEPS). She received her PhD degree in Economics from the University of Latvia. Her main research interests include public economics and labour economics.

freepolicybriefs.com

The Forum for Research on Eastern Europe and Emerging Economies is a network of academic experts on economic issues in Eastern Europe and the former Soviet Union at BEROC (Minsk), BICEPS (Riga), CEFIR (Moscow), CenEA (Szczecin), KEI (Kiev) and SITE (Stockholm). The weekly FREE Network Policy Brief Series provides research-based analyses of economic policy issues relevant to Eastern Europe and emerging markets. Opinions expressed in policy briefs and other publications are those of the authors; they do not necessarily reflect those of the FREE Network and its research institutes.

FREE NETWORK