AI & SOCIETY (2024) 39:597–616
https://doi.org/10.1007/s00146-022-01490-3

ORIGINAL PAPER

What do we really know about the drivers of undeclared work? An evaluation of the current state of affairs using machine learning

Josip Franic[1]

Received: 16 December 2021 / Accepted: 19 April 2022 / Published online: 23 June 2022

Abstract
It is nowadays widely understood that undeclared work cannot be efficiently combated without a holistic view on the mechanisms underlying its existence. However, the question remains whether we possess all the pieces of the *holistic puzzle*. To fill the gap, in this paper, we test if the features so far known to affect the behaviour of taxpayers are sufficient to detect noncompliance with outstanding precision. This is done by training seven supervised machine learning models on the compilation of data from the 2019 Special Eurobarometer on undeclared work and relevant figures from other sources. The conducted analysis not only does attest to the completeness of our knowledge concerning the drivers of undeclared work but also paves the way for wide usage of artificial intelligence in monitoring and confronting this detrimental practice. The study, however, exposes the necessity of having at disposal considerably larger datasets compared to those currently available if successful real-world applications of machine learning are to be achieved in this field. Alongside the apparent theoretical contribution, this paper is thus also expected to be of particular importance for policymakers, whose efforts to tackle tax evasion will have to be expedited in the period after the COVID-19 pandemic.

Keywords Undeclared work · Informal economy · Tax evasion · Machine learning · Artificial intelligence · EU

JEL Classification E26 · H26 · J46 · C45

1 Introduction

Following the seminal paper by Hart (1973) on the working unemployed in Ghana, substantial effort has been invested to conceptualise and control economic activities that are inherently legitimate but remain hidden from the authorities.[1] While there is still a lot of work to be done in other parts of this research field (e.g., the task of quantification), the quest for the factors underlying the decisions of workers and companies to go *into the shadows* seems to have reached saturation point (Dularif et al. 2019; Hofmann et al. 2017). The last 50 years have witnessed a stream of research studies either upgrading the existing theories on the drivers of informality or complementing them with new ones (Castells and Portes 1989; de Soto 1989; Feld and Frey 2007; Maloney 2004; Moser 1978; Round 2009; Sethuraman 1976). This went in parallel with discoveries of novel forms of noncompliance, which enabled continual refinements with respect to definitions and scope of activities in the focus[2] (Franic 2020a; Pfau-Effinger 2009; Williams and Horodnic 2017a).

A cursory insight into the publicly available literature exposes hundreds of research papers striving to explain why some market participants consciously circumvent legislation on taxes, employment relations, and/or various administrative obligations (see Hofmann et al. 2017). The vast amount of qualitative, quantitative, and experimental evidence from all parts of the world has, in fact, stimulated the emergence

✉ Josip Franic
josip.franic@ijf.hr

[1] Institute of Public Finance, Smiciklasova 21, 10000 Zagreb, Croatia

[1] Despite some earlier research on this topic (e.g., Geertz, 1963; Lewis, 1954), Hart's presentation to the Conference on urban unemployment in Africa is considered to be the turning point in this respect. His terminology was soon adopted and popularised by the International Labour Organization (ILO, 1972, 1993, 2002).

[2] As a side effect, more than 40 different names have been used to denote unregulated work, each of them representing slightly different (but to a large extent overlapping) group of activities (Williams, 2004). Besides introducing additional confusion for practitioners dealing with this complex phenomenon, the terminological disunity also poses significant challenges for academics in their search of robust estimation methods.

of a novel research paradigm grounded on re-usage of the results from available studies in yet another round of quantitative pursuit for undisputable determinants of concealed economic activities. The so-called *meta-analytic* approach assumes a systematic evaluation of the significance, direction, and (where possible) the magnitude of effect for a set of chosen indicators so as to understand whether, and in what circumstances, they indeed influence the behaviour of economic agents (Blackwell 2010; Dularif et al. 2019; Hofmann et al. 2017).

However, in spite of being invaluable for broadening horizons about the roots of noncompliance and the most common offenders, neither the existing core nor meta-studies have managed to provide an answer to the fundamental question: have we really gained a holistic picture of the forces behind unregistered activities or are there still some missing pieces of the puzzle? In other words, it is unclear whether the list of factors identified so far as drivers of undeclared work is complete or not.

This paper seeks to fill the gap by addressing an equivalent problem. Suppose that for a certain person/company, we possess all the information on the features hitherto known to shape the decision on (non)declaration of activities. If the list of causal factors is exhaustive, there must exist a statistical model (or a group of models) able to determine with extremely high exactness if this agent participates in the undeclared economy. Conversely, if such a model cannot be constructed, this would imply that there are some other drivers of informality which we are not aware of at this point.

This research problem was insolvable until recently due to well-known limitations of the *traditional econometric techniques* (Athey and Imbens 2019; Boulesteix and Schmid 2014; Di Franco and Santurro 2021; Mullainathan and Spiess 2017). However, recent developments in the area of machine learning not only have enabled the realisation of this *sanity check*, but also have opened the prospect for enhancements in other niches of this research field (e.g., detection and quantification). That being said, this paper intends to have both theoretical and practical contributions. Besides adding to the ongoing debate on the determinants of unregistered activities, it will also reveal the full potential of artificial intelligence in the fight against this deleterious practice. Above and beyond, the presented methodology will be highly relevant for the government officials around the world who will face the need to substantially increase public budget revenues in the post-COVID era.

To achieve these aims, the rest of the paper is organised as follows: alongside summarising the most important findings on the matter over the last 5 decades, Sect. 2 also provides a further discussion on the gaps in the literature and specifies how this paper is going to address them. Section 3 introduces the datasets and methods employed to answer the posed research question. Particular attention will be paid to deficiencies of the traditional econometric methods and the supremacy of machine learning in this respect. After presenting and discussing the most important results of the conducted analysis in Sect. 4, the paper ends with concluding remarks and suggestions for future research.

Before moving forward, it ought to be said that the focus of this paper is strictly on undeclared work, i.e., all market-oriented activities which, despite being legitimate per se, remain deliberately hidden from the authorities to evade taxes and social security contributions, to avoid compliance with labour legislation, and/or to circumvent any other administrative requirement (European Commission 1998). This would say that prohibited activities (human trafficking, prostitution, drug-smuggling, etc.), as well as self-provisioning, neighbour help, volunteering, and alike forms of unpaid work remain out of the scope of this study. The same is true for tax evasion related to activities that do not result in any added value (e.g., frauds with capital gains) and tax avoidance.

2 Literature review

The pioneering studies on the matter, based primarily on qualitative research in developing countries, described undeclared work as a leftover of the pre-capitalist period which "would disappear once these countries achieved sufficient levels of economic growth or modern industrial development" (Chen et al. 2004, p. 16). This view regarded low employment prospects and pervasive poverty among the population of fast-growing cities as the sole drivers of unregulated activities (Hart 1973; Sethuraman 1976; Tokman 1978). In line with that, the name *informal sector* was used to reflect the idea of two distinct and autonomous realms, namely "a dynamic, profit-making modern sector and everything else—a vast sponge of surplus labor" (Peattie 1987, p. 852). The latter was believed to consist mostly of low-skilled individuals who emigrated from rural parts of the country in search of any income opportunity (Hart 1973).

Following the increasing interest in this topic, it was soon realised that reality is actually far more complex. Not only was the existence of undeclared work evidenced in developed countries, as well, but it quickly became apparent that this harmful practice was going to stay ingrained in economies around the world (Moser 1978; Rakowski 1994). Explicitly, a number of studies conducted during the late 1970s found that a considerable part of concealed market-oriented activities in advanced economies was, in fact, a by-product of capitalism (see for instance Moser 1978). Owing to the combination of economic turbulences and accelerated globalisation, many formal firms were forced to decrease production costs to survive on the market. Alongside the automation of production processes and transfers of

plants to less-developed countries, subcontracting of work to small unregistered firms and hiring workers off-the-books also emerged as pleasing strategies in the *race to the bottom* (Castells and Portes 1989; Davis 2006).

These findings gave rise to the so-called *structuralist* school of thought, whose proponents saw the roots of undeclared work in weak and inefficient state institutions unable to prevent the exploitation of the impoverished masses. Structuralists thus called for better protection of workers' rights, more stringent regulation of businesses, and strengthening the rule of law (Portes and Sassen-Koob 1987; Rakowski 1994).

The turbulent 1980s put the state once again in the centre of discussion on the flourishing informalities. However, this time, the focus shifted to raising tax burdens, overcomplicated registration procedures, and amplified income inequalities (Annis and Franks 1989; de Soto 1989). Introducing the concept of *emotional agents*, research studies from the period showed that many individuals and firms decided to operate on an undeclared basis out of defiance. This *legalist* interpretation of the state of affairs portrayed tax evaders as the democratic force that openly stood against an unfair and intrusive state (Rakowski 1994).

The emotional agent approach was complemented during the 1990s and 2000s with the theories of rational and quasi-rational agents. The concept of rational voluntarism arose from studies which revealed that some workers and companies freely chose to operate off-the-books after assessing the costs and benefits of such behaviour (Fields 1990; Maloney 2004).[3] While acknowledging the driving forces identified by the legalists, this view on the roots of noncompliance introduced additional elements to the equation, namely the risk of being detected, plausible sanctions, the quality of pension and welfare systems, and the difference in pay rates between declared and undeclared work. However, the most important novelty brought by this school of thought was the recognition of the *upper tier informal economy*—part of the undeclared sphere attracting affluent individuals eager to increase their wealth (Fields 1990).

The mismatch between the compliance rates implied by the rational-agent theory and actual compliance inspired a stream of research on the role of personal and social norms in the process (Alm et al. 2017; Frey and Torgler 2007; Torgler 2004). In the quest for the reasons why some people always comply, while others seek evasion strategies even when the potential cost outweighs the benefits, the academic community has recently put a greater emphasis on a latent construct known as *tax morale*. Defined as "individual's willingness to pay taxes, in other words, the moral obligation to pay taxes or the belief that paying taxes contributes to society" (Frey and Torgler 2007, p. 140), this attribute was found to be a compound outcome of numerous socio-economic, psychological, and demographic peculiarities. Although still not fully understood (which particularly applies to the hereditary context), evidence suggests that tax morale is a dynamic feature heavily influenced by vertical trust (i.e., trust in the state institutions), horizontal trust (trust in other taxpayers), and various personal characteristics. For instance, it was shown that men generally express lower intrinsic readiness to pay taxes than women, and the same applies to younger individuals compared to more experienced ones (Alm and Torgler 2006; Lago-Peñas and Lago-Peñas 2010). The importance of religion in one's life, marital status, size of the family, the level of education, and occupation are also some of the key facets in this respect (Benk et al. 2016; Lago-Peñas and Lago-Peñas 2010; Strielkowski and Čábelková 2015).

When it comes to trust, a number of studies revealed that some taxpayers tend to comply with tax legislation as long as they think that the authorities respect an imperceptible psychological contract between the state and citizens (Franic 2019; van Dijke and Verboon 2010). Conversely, if they believe the public funds are not spent fairly and efficiently, such people will seek strategies to reduce their tax duties (Barone and Mocetti 2011). The efficacy of the state apparatus, the quality of the services received, and the perceived prevalence of corruption in public institutions are the most important factors shaping the views of taxpayers on this matter (Alm et al. 2010; van Dijke and Verboon 2010).

Horizontal trust, on the other hand, is embodied in the concept of *conditional cooperation*. As explained by Frey and Torgler (2007), an individual's willingness to pay own taxes is strongly affected by the perception regarding the behaviour of their counterparts in this regard. If they think that others are not respecting the implicit social deal, some of the honest taxpayers will shift to the undeclared sphere simply because they feel fooled. Furthermore, the pervasive informality usually signals tacit approval of this practice in society, thus reducing the moral cost of the wrongdoing and igniting further noncompliance (Alm et al. 2017; Torgler 2004).

Even though each subsequent view on the mechanisms underlying undeclared work was grounded on criticism of the existing theories, time has shown that all of them are, in fact, valid. Indeed, the latest stream of quantitative inquiry, based on large-scale questionnaire surveys and experimental studies, has revealed that these schools of thought are complementing rather than contesting each other (Chen 2012; Williams and Round 2007). For instance, a number of recent studies have underlined limited employment prospects in the formal sector as being responsible for a large portion of modern-day undeclared

[3] The idea of undeclared work as a rational choice was first presented by Allingham and Sandmo (1972) in their theoretical paper on income tax evasion.

work, both in developed and developing countries (Elek and Köllő 2019; Williams and Efendic 2021; Williams and Horodnic 2015b). Likewise, numerous low-skilled workers are still being pulled into this sphere by reckless employers not hesitating to resort to exploitative practices for their own gain (Franic 2020b; Palumbo 2017). Finally, it appears that more market participants than ever before nowadays eagerly embrace undeclared work: while some do so because they believe that the benefits exceed perceived cost, others decide to go into the shadows as a rebellion against the inefficient and over-intrusive state and/or due to feeling deceived by their fellow taxpayers (Alm et al. 2010, 2017; Kogler et al. 2013; van Dijke and Verboon 2010). For the majority of taxpayers, their (non)participation in the undeclared economy is, however, a complex outcome of a number of intertwined factors (Franic 2020b; Franic and Cichocki 2021).

Since each of the aforementioned theorizations represents one piece of the key for the decision-making riddle on the part of economic agents, the term *holistic approach* now dominates the literature on this phenomenon (see Chen 2012; Williams 2016). However, while academics and professionals do agree that a holistic view is needed, it is not clear whether all parts of the puzzle are already on the table. In the rest of the paper, we seek to provide an answer precisely to this question. This will be done by applying the latest machine learning techniques on an all-inclusive dataset derived from various sources, as explained in the following section.

3 Data and methods

A careful reader might ask themselves why this essential matter has not been evaluated so far. To understand the reasons, one must be familiar with the limitations of the traditional econometric approach. First of all, the research questions and hypotheses were until recently driven strictly by data availability, given that quantitative researchers interested in the determinants of undeclared work were forced to rely solely on the variables available as part of a particular questionnaire survey or experimental study (Grimmer et al. 2021). In line with that, most research articles published before the 2010s provide quite a narrow insight into the matter.

This issue was somewhat mitigated by the emergence of multilevel techniques, which made it possible to combine the data from questionnaire surveys with external macroeconomic and akin figures. However, analyses based on a multilevel approach commonly follow the sequential modelling strategy, meaning that in practice researchers are not able to test all the required variables in parallel (see for instance Franic and Cichocki 2021; Williams and Horodnic 2016). This is a direct consequence of the susceptibility of traditional econometric models to multicollinearity, as well as of requirements with respect to the ratio of a sample size to the number of covariates (Athey and Imbens 2019; Grimmer et al. 2021).

Given these constraints, classical methods commonly require data selection procedure and/or feature engineering to be applied, which frequently leads to biased results (Athey and Imbens 2019; Boulesteix and Schmid 2014). Nevertheless, even if all steps are taken to obtain unbiased estimates, not much can be done to address the overfitting to the sample (Molina and Garip 2019; Mullainathan and Spiess 2017). Consequentially, the still-prevailing statistical methods are in practice mainly used for the exploration of causality, as they seldom show robustness in predictive modelling.

These shortcomings led to the development of a range of machine learning techniques, which were designed specifically for predictive purposes. As a result of tremendous enhancement in this area, many supervised machine learning models are nowadays able to reach predictive accuracy close to 100% (see for instance Li et al. 2020).[4] The secret of success resides in a combination of iterative estimation approach (which eliminates the concern with multicollinearity and selection bias), flexible model architectures, ability to detect latent mechanisms, and, consequently, a huge number of model coefficients (Ghoddusi et al. 2019; Mullainathan and Spiess 2017). These features make machine learning a perfect choice for our task.

Before proceeding with the details on the exact models employed in the analysis, it is first necessary to introduce the datasets used for this purpose. The starting point in this respect was the latest wave of the Special Eurobarometer on undeclared work. Being one of the most comprehensive sources of data on this topic, this survey conducted in September 2019 provides an insight into the experience, views, and attitudes of 26,514 EU citizens regarding undeclared activities.[5] Akin to the previous two waves (from 2007 and 2013), this one also contained the following question:

> Have you yourself carried out any undeclared paid activities in the last 12 months, either on your own account or for an employer?

[4] This applies primarily to deep learning, but other branches of this field also show remarkable achievements.

[5] Approximately 1,000 persons aged 15 or more were interviewed in each member state following a multi-stage random (probability) sampling procedure. For more details on methodology, see European Commission (2020).

The resulting variable is binary, which means we are faced with a standard classification task.[6] In line with the findings of existing studies on the drivers of undeclared work and reflecting the discussion from earlier parts of the paper, the following three sets of features directly available from the survey are used as input variables in the modelling:

Socio-demographic: gender, age, marital status, and household size.

Socio-economic: type of community (urban/rural), country of residence, migrant status, education, occupation, size of the company, financial difficulties, and social class.

Perceptions and attitudes: perceived detection risk, expected sanctions, having undeclared workers among friends and relatives, estimated percentage of the population engaged in undeclared activities, trust in tax authorities, trust in labour inspectorate, and tax morale index.

Details on coding and research papers finding a significant effect of these features are given in Table 1. Since these 19 variables do not represent a complete list of the factors so far known to influence the behaviour of economic agents, we also add a range of country-level determinants compiled from various sources. These variables can be roughly divided into four groups as follows:

Economic constraints: unemployment rate, the implicit tax rate on labour, income inequality index, and relative median income ratio for persons above the age of 65 (a proxy for the quality of pension systems).

State intervention: the size of the government, stringency of labour market regulations, and stringency of business regulations.

Quality of formal institutions: government effectiveness, rule of law, regulatory quality, and corruption perceptions index.

Informal institutions: trust in government, trust in other people, religiosity, and the average level of tax morale.

With the exception of the average level of tax morale, which was devised directly from the Special Eurobarometer on undeclared work, all other figures were collected from credible international institutions (see Table 2 for details). To ensure compatibility with the baseline dataset, values for 2019 were taken whenever possible. In case of data for 2019 not being available, the most recent figures were considered. These variables were then incorporated into the Eurobarometer dataset, which gave a total of 34 features whose predictive power was to be tested simultaneously.

At this point, it is important to underline yet another difference between the traditional econometric approach and machine learning. While the former requires special techniques to deal with variables collected at different levels, the latter is robust to the violation of the independence of observations assumption. That is to say, since they only care about whether and to what extent a certain feature can help in predicting the modelled variable, machine learning methods do not make any distinction between individual-level and country-level variables.

Given a relatively small number of input variables, a total of 26,514 sampled units would in most cases be sufficient to construct and train well-performing machine learning models. However, things are slightly complicated in our case due to the imbalanced nature of the target variable. Explicitly, only 950 Eurobarometer survey participants admitted their participation in the undeclared economy from the supply side, which introduces a substantial risk of models being biased towards the negative outcome.[7] To address this issue, during the training phase, we applied the random oversampling scheme with weights inversely proportional to class frequencies (see Fernández Hilario et al. 2018; Viloria et al. 2020).

As our emphasis is first and foremost on the training set accuracy (owing to the nature of the research question), this procedure elegantly resolves the problem. However, it does not help much when it comes to the generalisability of the results.[8] Given the complexity of the researched matter, information on as few as 950 undeclared individuals from across the EU is simply not enough to achieve high classification accuracy on unseen cases with a single model, regardless of its achievement on the training set.

To mitigate this problem, the decision was made to train seven different models from a wide palette of supervised machine learning techniques and combine their results through the ensemble voting approach. The first and the simplest of models chosen for this purpose was the decision tree which is, despite being useful for grasping the relative importance of examined causal factors, highly prone to overfitting (Hastie et al. 2008; Mitchell 1997). For this reason, it was complemented with the Adaptive Boosting (AdaBoost)

[6] It is important to note that only the labour supply side of the phenomenon will be evaluated here. However, the approach applied in the rest of this paper can be easily adapted to the demand for unregistered employees/subcontractors, purchase of undeclared goods and services, and alike forms of violation.

[7] Since the goal of a typical supervised learning model is to maximise accuracy, the computer can easily figure out that predicting majority-class outcome for each and every training set member is the most elegant solution (see for instance Johnson & Khoshgoftaar, 2019). In our case, this means that the model would be correct in 100*(1-(950/26,514))=96.42% of cases if always foreseeing nonparticipation in undeclared work, which might appear satisfactory from the perspective of this artificial intelligence and prevent it from further learning.

[8] Synthetic minority oversampling (SMOTE) and dimensionality reduction using principal component analysis, which often yield models with better out-of-sample predictions, were also considered as possible solutions to this problem. However, the results were inferior to the ones based on reweighting.

Table 1 Overview of individual-level features. Source: Author's own work

	Variable name	Type	Original coding	Nbr. of missing values	Recent studies finding a significant effect of this determinant
Demographic	Gender	Binary	0: male; 1: female	0	Elek and Köllő (2019), Franic and Cichocki (2021), Gregorio and Giordano (2016), Hofmann et al. (2017), Popescu et al. (2016), van Dijke and Verboon (2010), Williams and Horodnic (2015a, b)
	Age	Interval	Values representing exact age	0	Elek and Köllő (2019), Gregorio and Giordano (2016), Hofmann et al. (2017), Popescu et al. (2016), Williams and Horodnic (2015a, b), Windebank and Horodnic (2017)
	Marital status	Categorical	1: (re-)married without children; 2: (re-)married with children from this marriage; 3: (re-)married with children from a previous marriage(s); 4: (re-)married with children from this and previous marriage(s); 5: cohabiting without children; 6: cohabiting with children from this union; 7: cohabiting with children from previous union(s); 8: cohabiting with children from this and previous union(s); 9: single without children; 10: single with children; 11: divorced/separated without children; 12: divorced/separated with children; 13: widowed without children; 14: widowed with children	164	Alm et al. (2016), Arendt et al. (2020), Franic and Cichocki (2021), Gregorio and Giordano (2016), Popescu et al. (2016), Strielkowski and Čábelková (2015), Williams and Horodnic (2015a, b)
	Household size	Interval	Values representing the exact number of persons in a household	2	Arendt et al. (2020), Williams and Horodnic (2015a, b), Williams et al. (2015a, b)

Table 1 (continued)

	Variable name	Type	Original coding	Nbr. of missing values	Recent studies finding a significant effect of this determinant
Socio-economic	Type of community	Categorical	1: rural area; 2: town or suburb/small urban area; 3: city/large urban area	0	Boone et al. (2013), Popescu et al. (2016), Williams and Efendic (2021)
	Country of residence	Categorical	Values designating in which of 27 member states an individual lives	0	Franic and Cichocki (2021), Kayaoglu and Williams (2017), Williams et al. (2015a)
	Migrant worker	Categorical	0: individual does not work abroad; 1: individual works in another EU member state; 3: individual works outside EU	0	Gregorio and Giordano (2016), McKay (2014), Porthé et al. (2010), Rodgers et al. (2019), Williams and Efendic (2020)
	Age when finished education	Categorical	1: up to 15 years; 2: 16–19; 3: 20 years and older; 4: still studying; 5: no full-time education	409	Arendt et al. (2020), Boone et al. (2013), Elek and Köllő (2019), Gregorio and Giordano (2016), Hofmann et al. (2017), van Dijke and Verboon (2010), Williams and Horodnic (2015a, b)
	Occupation	Categorical	1: houseperson; 2: student; 3: unemployed, temporary not working; 4: retired, unable to work; 5: farmer; 6: fisherman; 7: professional (lawyer, etc.); 8: owner of a shop, craftsman, etc.; 9: business proprietor; 10: employed professional (doctor, etc.); 11: general management; 12: middle management; 13: employed position, et desk; 14: employed position, travelling; 15: employed position, service job; 16: supervisor; 17: skilled manual worker; 18: unskilled manual worker	0	Franic and Cichocki (2021), Gregorio and Giordano (2016), Kayaoglu and Williams (2017), Strielkowski and Čábelková (2015), Williams et al. (2015a), Windebank and Horodnic (2017)
	Size of the company	Categorical	0: not working or self-employed without workers; 1: 1–4 employees; 2: 5–9 employees; 3: 10–19 employees; 4: 20–49 employees; 5: 50–99 employees; 6: 100–499 employees; 7: 500 + employees	772	Elek and Köllő (2019), Franic and Cichocki (2021), Popescu et al. (2016), Vallanti and Gianfreda (2020), Williams et al. (2015a)
	Financial difficulties	Categorical	1: most of the time; 2: from time to time; 3: almost never/never	388	Arendt et al. (2020), Boone et al. (2013), Hofmann et al. (2017), Popescu et al. (2016), Williams and Efendic (2021), Williams and Horodnic (2015a, b), Williams et al. (2015a, b)
	Social class (self-assessed)	Categorical	1: working class; 2: lower middle class; 3: middle class; 4: upper-middle class; 5: higher class	868	Williams et al. (2015a, b), Williams and Horodnic (2015a), Williams and Horodnic (2017b)

Table 1 (continued)

	Variable name	Type	Original coding	Nbr. of missing values	Recent studies finding a significant effect of this determinant
Perceptions and attitudes	Perceived detection risk	Categorical	1: very high; 2: fairly high; 3: fairly small; 4: very small	2685	Arendt et al. (2020), Elek and Köllő (2019), Franic and Cichocki (2021), van Dijke and Verboon (2010)
	Expected sanction if caught in undeclared work	Categorical	1: normal tax or social security contributions due; 2: normal tax or social security contributions due, plus a fine; 3: prison	3408	Fegatilli (2009), Feld and Larsen (2012), van Dijke and Verboon (2010)
	Any undeclared workers in social circle	Binary	0: No; 1: Yes	1172	Horodnic and Williams (2019), Kayaoglu and Williams (2017), Williams and Öz-Yalaman (2021)
	Estimated % of population engaged in undeclared work	Categorical	1: less than 1%; 2: 1–5%; 3: 6–10%; 4: 11–20%; 5: 21–30%; 6: 31–40%; 7: 41–50%; 8: more than 50%	5203	Alm et al. (2017), Franic and Cichocki (2021), Jimenez and Iyer (2016), Williams (2019), Gërxhani and Wintrobe (2021)
	Trust in tax authorities	Binary	0: No; 1: Yes	2640	Alm et al. (2010), Kogler et al. (2015), Rodrigues (2020), van Dijke and Verboon (2010)
	Trust in labour inspectorate	Binary	0: No; 1: Yes	2894	Kogler et al. (2013), van Dijke and Verboon (2010)
	Tax morale	Interval	Values from 1 to 10, where larger numbers represent lower tax morale	1424	Franić (2020), Franic and Cichocki (2021), Jimenez and Iyer (2016), Williams and Horodnic (2015a, b), Windebank and Horodnic (2017)

(i) To enhance training and increase predictive power, all categorical variables were recoded into a set of binary indicators, while interval variables were normalised

(ii) Missing values were imputed through an iterated round-robin procedure based on Bayesian ridge regression

Table 2 Overview of country-level variables. Source: Author's own work

	Variable name	Type	Coding details	Source	Studies finding a significant effect of this or a closely related variable
Economic constraints	Unemployment rate	Interval	The number of unemployed persons as a percentage of the active population	Eurostat (2021d)	Elek and Köllő (2019), Williams and Efendic (2021), Williams and Horodnic (2015a, b), Williams et al. (2015a, b)
	Implicit tax rate on labour	Interval	The sum of all direct and indirect taxes and employees' and employers' social contributions levied on employed labour income divided by the total compensation of employees working in the economic territory increased by taxes on wage bill and payroll	Eurostat (2021a)	Ameyaw and Dzaka (2016), Clotfelter (1983), Dreher et al. (2009), Kayaoglu and Williams (2017), Rei and Bhattacharya (2008), Kassa (2021), Kuehn (2014), Pommerehne (1996), Yitzhaki (1987)
	Income inequality index	Interval	The ratio of total income received by the 20% of the population with the highest income to that received by the 20% of the population with the lowest income	Eurostat (2021b)	Bloomquist (2003), Christie and Holzner (2006), Engel et al. (2020), Kayaoglu and Williams (2017), Shafer et al. (2020)
	Relative median income ratio for persons 65 +	Interval	The ratio of the median equivalised disposable income of people aged above 65 to the median equivalised disposable income of those aged below 65	Eurostat (2021c)	Franic and Cichocki (2021), Franic (2020b), Katnic and Williams (2018), Williams (2007)
State intervention	Size of the government	Interval	Values indicating the extent to which countries rely on the political process to allocate resources and goods and services (values given on the scale between 0 and 10)	Fraser Institute (2021)	Arendt et al. (2020), Li and Ma (2015), Picur and Riahi-Belkaoui (2006), Rei and Bhattacharya (2008)
	Labour market regulations	Interval	The extent to which various restraints (e.g., minimum wages, dismissal regulations, centralized wage setting, extension of union contracts to non-participating parties, and conscription) upon economic freedom are present (values given on the scale between 0 and 10)	Fraser Institute (2021)	Rei and Bhattacharya (2008), Bíró et al. (2020), Loayza et al. (2005), Vorley and Williams (2012), Williams (2015)
	Business regulations	Interval	The extent to which regulations and bureaucratic procedures restrain entry and reduce competition (values given on the scale between 0 and 10)	Fraser Institute (2021)	Goel and Saunoris (2017), Islam et al. (2020), Loayza et al. (2005), Popescu et al. (2018), Riahi-Belkaoui (2004), Vallanti and Gianfreda (2020)

Table 2 (continued)

	Variable name	Type	Coding details	Source	Studies finding a significant effect of this or a closely related variable
Quality of formal institutions	Government effectiveness	Interval	Perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies (values in the range between − 2.5 and 2.5)	World Bank (2021)	Dreher et al. (2009), Kayaoglu and Williams (2017), Islam et al. (2020), Kuehn (2014), Rashid et al. (2021), Yamen et al. (2018)
	Rule of law	Interval	Perceptions of the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, property rights, the police, and the courts, as well as the likelihood of crime and violence (values in the range between − 2.5 and 2.5)	World Bank (2021)	Christie and Holzner (2006), Islam et al. (2020), Kogler et al. (2013), Kayaoglu and Williams (2017), Rashid et al. (2021), Richardson (2006), Yamen et al. (2018)
	Regulatory quality	Interval	Perceptions of the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development (values in the range between − 2.5 and 2.5)	World Bank (2021)	Loayza et al. (2005), Popescu et al. (2018), Rashid et al. (2021), Riahi-Belkaoui (2004), Richardson (2006), Yamen et al. (2018)
	Corruption perceptions index	Interval	Perceptions by business people and country experts of the level of corruption in the public sector (values given on the scale between 0 and 100)	Transparency International (2021)	Christie and Holzner (2006), Kayaoglu and Williams (2017), Khlif et al. (2016), Kogler et al. (2013), Picur and Riahi-Belkaoui (2006)

Table 2 (continued)

	Variable name	Type	Coding details	Source	Studies finding a significant effect of this or a closely related variable
Informal institutions	Trust in government	Interval	Percentage of people that tend to trust the national government	European Commission (2019)	Gërxhani and Wintrobe (2021), Forteza and Noboa (2019), Jimenez and Iyer (2016), Richardson (2006), Strielkowski and Čábelková (2015), van Dijke and Verboon (2010)
	Trust in other people	Interval	The average level of trust in other people (values given on the scale between 0 and 10)	ESS (2021)	Alm et al. (2017), Bobek et al. (2013), Engel et al. (2020), Frey and Torgler (2007), Gërxhani and Wintrobe (2021)
	Religiosity	Interval	The share of the population attending religious services (apart from weddings, funerals, and christenings) at least once a week	EVS (2021)	Alm et al. (2016), Benk et al. (2016), Boone et al. (2013), Islam et al. (2020), Rashid et al. (2021), Richardson (2006), Strielkowski and Čábelková (2015)
	Average level of tax morale	Interval	Average values by country of the tax morale index	Values constructed from individual-level variable	Franic and Cichocki (2021), Kemme et al. (2020), Riahi-Belkaoui (2004), Richardson (2006)

(i) All values refer to 2019, except for state intervention variables (2018), an indicator of trust in other people (2018), and figures on religiosity (2017)

model, which combines a sequence of *weak classifiers* into a single *strong classifier* in an iterative ensemble-learning fashion (Hastie et al. 2008; Wyner et al. 2017). To compensate for the limitations of these two methods, we further constructed a random forest model. Also belonging to a family of ensemble-learning techniques, random forest combines the results of multiple decision trees constructed using randomly drawn subsamples of the training set (Breiman 2001; Genuer and Poggi 2020).

Turning to the linear classifiers, which are commonly less prone to a high variance problem, the two most obvious choices were support vector machine (SVM) and logistic regression. However, while expected to outperform the non-linear classifiers on out-of-sample cases, these two methods usually exhibit higher bias (i.e., lower accuracy on the train set). Their limited efficiency in situations where data are not linearly separable is envisaged to specifically come to the fore in our case, given the likely existence of multiple latent mechanisms underlying decisions of labour suppliers with respect to (non)declaration of activities.

To account for the presence of multifaceted interactions between causal factors, two artificial neural networks were also devised. A deep learning approach commonly beats other classifications techniques, but it is not entirely immune to overfitting, especially when handling imbalanced data (Dabare et al. 2018; Johnson and Khoshgoftaar 2019; Mitchell 1997). For this reason, alongside a deep artificial neural network with five hidden layers, we also defined a *shallow* version, which incorporated only one hidden layer. The exact details on the design of these two artificial neural networks, as well as of the remaining five models are given in Appendix 1.

Before moving to the results, it should be mentioned that the data preparation phase was done in STATA, while Python modules TensorFlow and Scikit-learn were used to construct, train, and test the models introduced above. To make hyperparameter tuning and evaluation of performance possible, the original dataset was split into the training, validation, and test sets, whereby 20,000 individuals (approximately 75% of the sample) were randomly allocated to the training set, and the remaining ones were evenly split into the other two sets.

4 Findings and discussion

Results of the conducted analysis, presented in Table 3, provide strong evidence for the completeness of the list of causal factors. Five out of seven models managed to classify training examples with an accuracy above 99.5%, which indicates that all the building blocks for the holistic approach are already there.

Table 3 Performance metrics of the trained machine learning models. Source: Author's own work

		Decision tree	Random forest	AdaBoost	SVM	Logistic regression	Neural network (one hidden layer)	Neural network (five hidden layers)
Train set	Accuracy	1.0000	0.9999	1.0000	0.8012	0.7929	0.9997	0.9955
	Precision	1.0000	0.9972	1.0000	0.1250	0.1194	0.9902	0.8879
	Recall	1.0000	1.0000	1.0000	0.7720	0.7635	1.0000	0.9986
	F1 score	1.0000	0.9986	1.0000	0.2151	0.2066	0.9958	0.9698
	AUC	1.0000	0.9999	1.0000	0.7871	0.8643	1.0000	0.9998
Validation set	Accuracy	0.9416	0.9622	0.9410	0.7982	0.7869	0.9481	0.9490
	Precision	0.1650	0.4375	0.1553	0.1288	0.1205	0.2391	0.2078
	Recall	0.1405	0.0579	0.1322	0.7686	0.7521	0.1818	0.1322
	F1 score	0.1518	0.1022	0.1429	0.2206	0.2078	0.2066	0.1616
	AUC	0.5565	0.5275	0.5522	0.7840	0.8620	0.7502	0.6464
Test set	Accuracy	0.9358	0.9638	0.9380	0.8001	0.7940	0.9484	0.9515
	Precision	0.1293	0.6923	0.1441	0.1323	0.1267	0.2527	0.2464
	Recall	0.1220	0.0732	0.1301	0.7724	0.7561	0.1870	0.1382
	F1 score	0.1255	0.1324	0.1368	0.2259	0.2170	0.2150	0.1771
	AUC	0.5449	0.5359	0.5499	0.7868	0.8577	0.7252	0.6430

(i) Precision denotes the share of true positives in total predicted positives; recall is the share of true positives in total actual positives; F1 score is the harmonic mean of the precision and recall; area under the curve (AUC) measures the ability of a classifier to distinguish between classes (on a scale from 0 to 1, with larger values signalising better performance)

Nonetheless, a more thorough insight into the key performance metrics reveals that it is not individual drivers, but rather complex interactions of theirs what shapes the behaviour of labour suppliers. This can be easily concluded by inspecting the figures for support vector machine and logistic regression, which were strongly outperformed by the remaining models on the training set. Being able to assess only a direct effect of the examined features on the target variable, these two models managed to correctly classify just 8 out of 10 individuals from the training set.

On the other hand, the shallow artificial neural network with a total of 625,001 parameters capturing latent interconnections between the explanatory variables achieved a remarkable accuracy of 99.97%. A similar result was obtained with the five-layer network, which comprised 749,901 parameters.[9] The importance of interdependence between causal factors is further evidenced by evaluation statistics of the remaining three models. Decision tree and AdaBoost managed to correctly classify all training examples, while random forest made only a few mistakes (see Table 3).

Nevertheless, an extraordinary performance on the training set does not imply that these models can be straightforwardly applied in practice. As a matter of fact, results on the unseen data are also quite misleading in this respect. Although the performance metrics for the test set, which are also presented in Table 3, reveal that five models were able to correctly classify at least nine out of ten previously unseen individuals (with accuracy rates ranging from 93.58% for decision tree and 96.38% in case of the random forest), this is not as satisfying as one might assume on a first glance. Given that only 3.58% of the survey respondents admitted working on an undeclared basis, a naive model predicting a negative outcome for each and every worker would achieve higher accuracy than any of our seven models (it would be correct in 96.42% of cases). This implies that accuracy is not a particularly informative performance measure for models dealing with *rare events* if the primary interest lies in the correct identification of positive cases.

Recall, F1 score, and area under the curve (AUC), which are better indicators of the classification power in our situation, shed completely new light on the models. In spite of much lower overall accuracy, logistic regression and support vector machine would actually perform best in practice, as each of them was able to correctly flag three out of four workers receiving undeclared income. The remaining models exhibit extremely low ability to identify true violators, with recall rates ranging from 0.0732 to 0.1870.

Although seeming contradictory, these results are fully reasonable and in line with the aforementioned caveat regarding a small number of undeclared workers in the sample. Besides modelling an imbalanced target variable, we also deal with the situation in which hidden interrelations of explanatory factors are vital for the segregation of positive

[9] In fact, unregularized versions of these two models were able to perfectly separate positive and negative training examples.

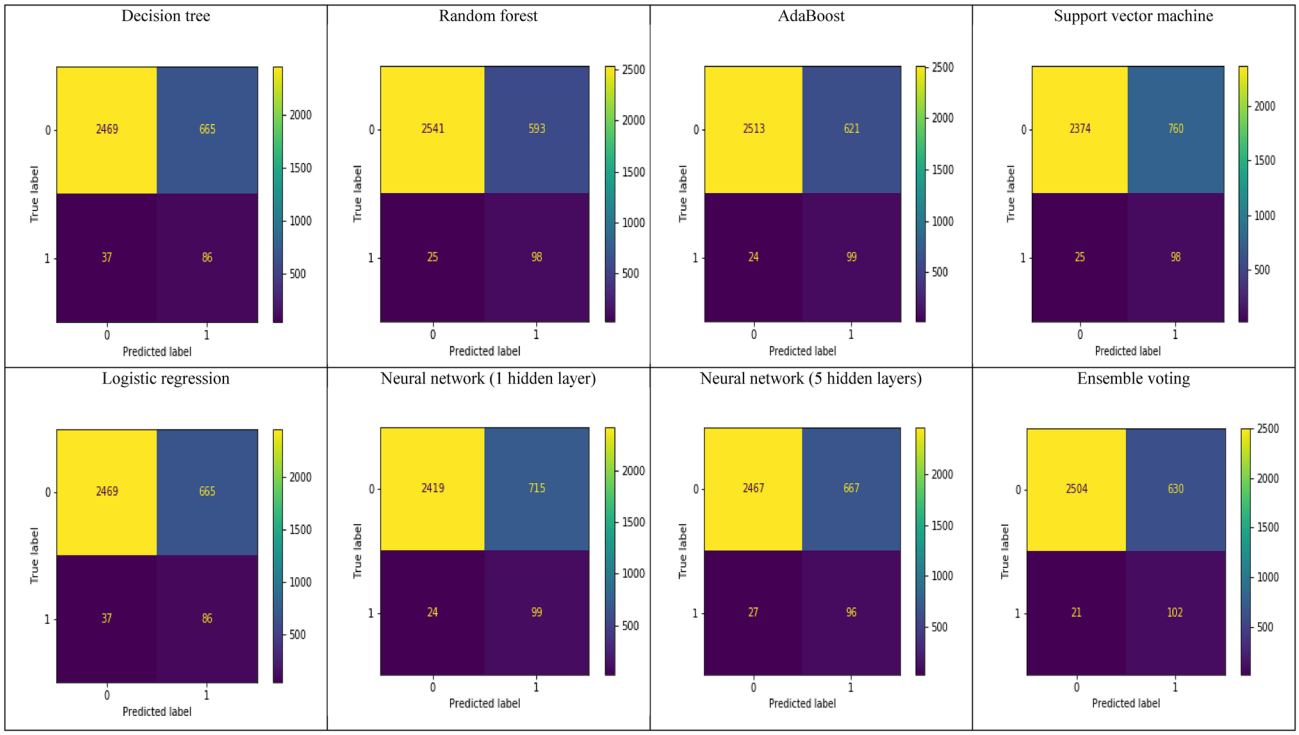

Fig. 1 Confusion matrices on the test set for models maximising recall. Source: Author's own work

and negative cases. At the same time, our models utilised information from 20,000 training examples, whereby only 728 of them were engaged in undeclared work (the remaining ones were allocated to the validation and test sets). Given such a small number of positive cases, five models suffer from high variance, meaning that they overfit to the training set.

This implies that practical applications of machine learning models on the identification of undeclared workers would require far larger datasets. It is hard to speculate on the exact size at this point, but no model would likely be able to generalise well unless there is at least a four-digit number of positive cases in the sample.

However, not everything is lost even with the existing dataset. To exemplify this, we modified hyperparameters of the models so as to maximise the accurateness in identifying true positives.[10] As can be seen from Fig. 1, which presents the accompanying confusion matrices, the majority of these *adjusted models* were able to correctly identify at least three out of four out-of-sample violators. AdaBoost and artificial neural network with one hidden layer were most successful in this respect, with a hit rate of 80%. The worst achievers, on the other hand, are decision tree and logistic regression, which managed to recognise 69.9% of violators.

To further increase recall, an additional logistic regression model was constructed and trained with input variables being predictions of individual models (see Appendix 2). The resulting coefficients, which are essentially the importance weights attached to each model based on their credibility, were then applied to the test set. This ensemble voting strategy indeed provided more accurate estimates, as can be seen from the last panel of Fig. 1. Specifically, the seven models were collectively able to spot 102 out of 123 previously unseen undeclared workers from the test set. Of course, this came with the cost of a substantial number of false positives, i.e., fully compliant workers who were incorrectly labelled as offenders. Adding more models to the ensemble voting scheme would certainly reduce this number and yield better results on recall. Nevertheless, due to the insufficient size of the training set, the final product would still fall below the level required for practical application.

Before finalising our discussion, it is beneficial to scrutinise the relative importance of individual drivers. This can be easily done by inspecting the results of the decision tree model. Due to the greedy variable selection approach and clear visualisation, decision trees provide a straightforward insight into the hierarchy of explanatory variables in terms of their contribution to the segregation of the target concept. In line with this, Fig. 2 unfolds the structure of the first four layers of the original decision tree as defined in Appendix 1 and elaborated in Table 1.

[10] Specification details are given in Appendix 2.

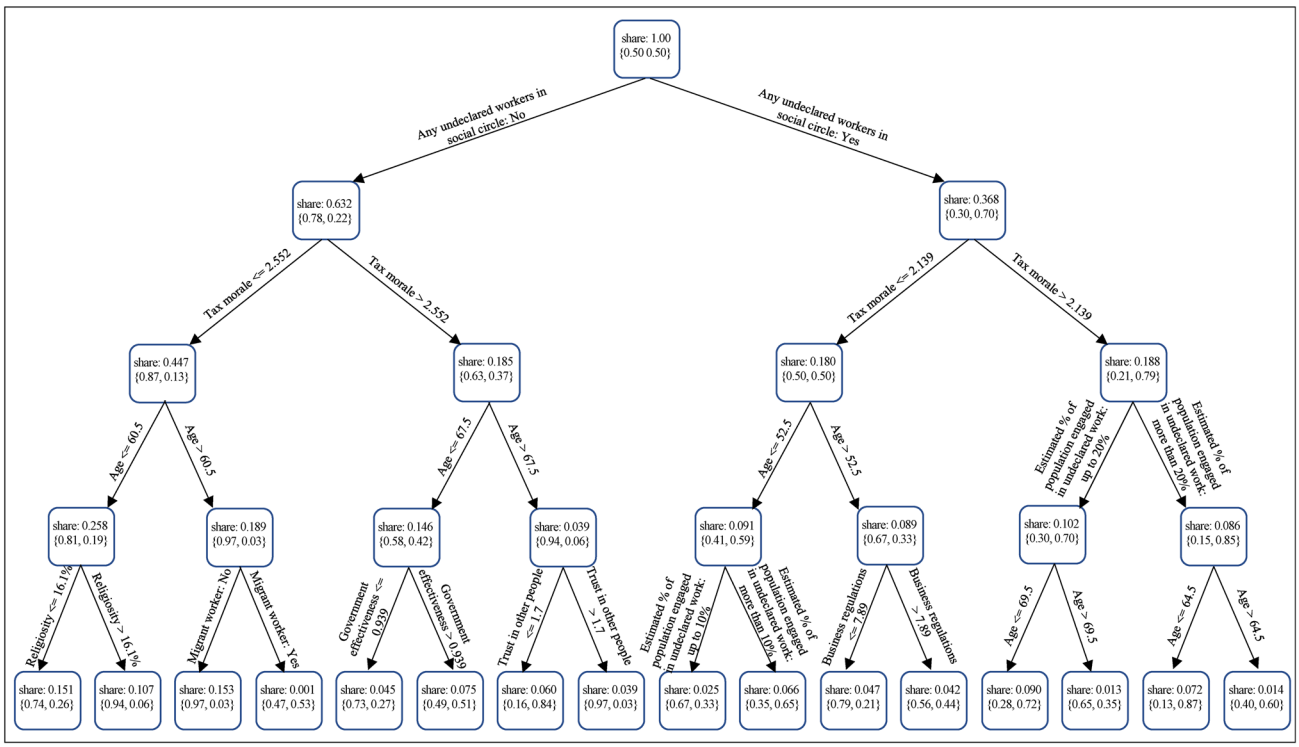

Fig. 2 Results of the decision tree. (i) Numbers in parentheses denote the split for target variable. Source: Author's own work

The visualisation exposes socio-psychological factors as the dominant forces driving one's decision whether or not to participate in the undeclared economy. Specifically, the existence of direct social contact with undeclared workers was identified as the best discriminant between the two cohorts of taxpayers. According to the results, individuals knowing somebody operating off-the-books are almost 2.5 times more likely to do the same. A note of caution is, however, necessary here due to a bidirectional causality: from the very nature of their activities, it follows that people who participate in the undeclared economy are more likely to have friends and acquaintances who also work in this sphere.

Further endorsing the findings from the recent stream of research on this matter, the decision tree highlighted tax morale as the second most important causal factor (see Franic 2020a; Frey and Torgler 2007; Williams and Bezeredi 2017). This variable, which emerged as the best separator in both branches of the second layer, also appears to be correlated with the previously mentioned determinant. This follows directly from the cut-off points of the two sub-trees, which are somewhat lower on the right side.

Things get more interesting when going deeper into the tree. In most cases, the third-best choice for filtering wrongdoers was their age, which exerts a substantial negative effect on the propensity to participate in the undeclared sphere (as already evidenced by, e.g., Arendt et al. 2020; Hofmann et al. 2017; Popescu et al. 2016). The only exception was individuals with low tax morale who have an undeclared worker in their surroundings (the rightmost rectangle in layer 3). For them, it is the perception regarding the pervasiveness of undeclared work in a society that matters most. Those convinced that at least every fifth citizen is hiding activities were found to have a 15% higher probability of doing the same than their counterparts who are more optimistic regarding the prevalence of undeclared work.

The subsequent layer of the tree further accentuates the relevance of subjective perceptions, given that trust in other people and religiosity were identified as the next best discriminant variables for a great many taxpayers. On the other hand, economic factors, such as government effectiveness, the stringency of business regulations, and the migrant status of a worker only at this stage do come to the fore.

5 Conclusion

A 5-decade-long endeavour to comprehend the fundamentals of undeclared work has brought to light a range of economic and socio-psychological factors influencing the behaviour of taxpayers in this respect. To underpin recent calls for an all-inclusive approach towards the eradication of the phenomenon (see Franic 2020b; Williams 2016), in this paper, we assessed the exhaustiveness of the list of known driving forces. The conducted analysis, which represents one of the

very first applications of up-to-date machine learning techniques to this research field, verified that all components necessary for a holistic approach are already there.

Specifically, five out of seven supervised machine learning models exhibited remarkable classification accuracy (99.5% or more) on the training set. Alongside demonstrating that the existing causal factors jointly provide sufficient information to fully segregate wrongdoers, our findings also support a growing body of research on the dominance of personal norms, beliefs, and values over mere economic constraints in the modern-day reasoning of workers (e.g., Alm et al. 2017; Franic 2020b; Williams and Yang 2018). As shown, intrinsic willingness to pay taxes, exposure to information about concealed activities, and their perception about the pervasiveness of such activities in society are nowadays crucial for one's decision whether to choose the same path. The quality of the state institutions and rigidness of regulations, on the other hand, are less important from the perspective of labour suppliers.

This, however, does not mean that the same is true for companies, buyers of undeclared goods and services, afternoon moonlighters, and akin offenders. Furthermore, given the economic, cultural, and political specificities of the European Union, it would be overly optimistic to claim that the same hierarchy of causal factors applies to workers from other parts of the world. If this study encourages similar research on the motives of other players within this realm and on other geographical areas, then it will have fulfilled one of its broader aims.

Besides the apparent theoretical contribution, the paper also paves the way for a wide usage of artificial intelligence in the fight against illegitimate economic activities. Yet, successful practical applications of machine learning models require not only much larger but also more credible datasets. This brings us to the main limitation of our study, which is closely linked to the nature of the information collected during the fieldwork. Due to the sensitivity of the researched matter, it is very likely that many survey respondents consciously denied their involvement in undeclared work. While imputation techniques helped to reduce the bias arising from a considerable number of missing responses, not much could be done in case of deliberate misreporting.

Given this, official records of labour inspectorates and tax administrations appear to be a much better source of information for training efficient violator-detecting machines. Although inferior to questionnaire surveys from the standpoints of the traditional econometric approach, data held by enforcement authorities are quite appealing in our case given the robustness of machine learning to the non-randomness of the sample. Typical examples of the existing sources that can be used for this purpose are the large-scale audit campaigns by the Internal Revenue Service (IRS 2016) and Her Majesty's Revenue and Customs (HMRC 2018). These and alike datasets already make it possible to train models which not only could detect noncompliance, but also would be able to provide more details on the type of offense in place and the magnitude of the violation.

That being said, machine learning might be a promising avenue for solving the most challenging task within this research field, which is the issue of quantification. As a matter of fact, recent advances in the fields of unsupervised learning and reinforcement learning, coupled with the introduction of sophisticated transaction-monitoring procedures on the part of surveillance bodies, open the prospects for real-time detection of noncompliance in a near future. If this paper encourages research studies heading in this direction, then it will have fulfilled its main goal.

6 Appendix 1: Details on the design of supervised machine learning models maximising accuracy

Decision tree	Criterion: Gini; splitter: best; max depth: 187; minimum samples at a leaf node: 1; maximum number of features: none; class weights: balanced; random state: 1
Random forest	Number of trees: 5000; criterion: Gini; maximum depth: 20; minimum samples at a leaf node: 2; minimum samples to split: 2; maximum number of features: auto; class weights: balanced; random state: 1
AdaBoost	Maximum number of estimators: 5000; base estimator: decision tree (class weights: balanced); learning rate: 1; random state: 1
Support vector machine	Regularization parameter: 2; kernel: linear; class weights: {0:1,1:26}; random state: 0
Logistic regression	Loss: binary cross-entropy; kernel initializer: Glorot uniform; kernel optimiser: Adam; kernel regularizer: l2 ($\lambda = 1e{-}2$); learning rate: 1e−4 for epochs 1–150, 1e−5 for epochs 151–200, 1e−9 for epochs > 200; class weights: {0: 1, 1: 26}; batch size: 64; number of epochs: 210

Neural network (1 hidden layer)	Number of neurons: 5000; activation: relu; loss: binary cross-entropy; kernel initializer: Glorot uniform; kernel optimiser: Adam; kernel regularizer: l2 ($\lambda = 1e{-}4$); learning rate: 1e−4 for epochs 1–180, 1e−5 for epochs > 180; class weights: {0: 1, 1: 26}; batch size: 64; number of epochs: 300
Neural network (5 hidden layers)	Number of neurons: 1000–500–200–100–50; activation: relu; loss: binary cross-entropy; kernel initializer: Glorot uniform; kernel optimiser: Adam; kernel regularizer: dropout (rate = 1e−1); learning rate: 1e−4 for epochs 1–100, 1e−5 for epochs 101–150, 1e−6 for epochs > 150; class weights: {0: 1, 1: 26}; batch size: 64; number of epochs: 200

Source: Author's own work

7 Appendix 2: Details on the design of supervised machine learning models maximising recall

Decision tree	Criterion: Gini; splitter: best; max depth: 8; minimum samples at a leaf node: 2; maximum number of features: none; class weights: balanced; random state: 1
Random forest	Number of trees: 500; criterion: Gini; maximum depth: 5; minimum samples at a leaf node: 2; minimum samples to split: 2; maximum number of features: auto; class weights: balanced; random state: 1
AdaBoost	Maximum number of estimators: 350; base estimator: decision tree (class weights: balanced; max depth: 1); learning rate: 1; random state: 1
Support vector machine	Regularization parameter: 2; kernel: polynomial; degree: 2; class weights: {0:1,1:26}; random state: 0
Logistic regression	Loss: binary cross-entropy; kernel initializer: Glorot uniform; kernel optimiser: Adam; kernel regularizer: l2 ($\lambda = 1e{-}5$); learning rate: 1e−4 for epochs 1–50, 1e−5 for epochs 51–100, 1e−6 for epochs > 100; class weights: {0: 1, 1: 26}; batch size: 64; number of epochs: 700; callbacks: early stopping (monitor: validation loss; patience: 4; restore best weights: true)
Neural network (one hidden layer)	Number of neurons: 5000; activation: relu; loss: binary cross-entropy; kernel initializer: Glorot uniform; kernel optimiser: Adam; kernel regularizer: l2 ($\lambda = 1e{-}4$); learning rate: 1e−5 for epochs 1–180, 1e−6 for epochs > 180; class weights: {0: 1, 1: 26}; batch size: 64; number of epochs: 300; callbacks: early stopping (monitor: validation loss; patience: 4; restore best weights: true)
Neural network (five hidden layers)	Number of neurons: 1000–500–200–100–50; activation: relu; loss: binary cross-entropy; kernel initializer: Glorot uniform; kernel optimiser: Adam; kernel regularizer: dropout (rate = 1.5e−1); learning rate: 1e−5 for epochs 1–100, 1e−6 for epochs 101–150, 1e−7 for epochs > 150; class weights: {0: 1, 1: 26}; batch size: 64; number of epochs: 200; callbacks: early stopping (monitor: validation loss; patience: 4; restore best weights: true)

Source: Author's own work

Data availability statement The datasets generated during and/or analysed during the current study are available in the GESIS repository, https://www.gesis.org/en/eurobarometer-data-service/search-data-access/data-access.

Declarations

Conflict of interest The author has no competing interests to declare that are relevant to the content of this article.

References

Allingham MG, Sandmo A (1972) Income tax evasion: a theoretical analysis. J Public Econ 1(2):323–338

Alm J, Torgler B (2006) Culture differences and tax morale in the United States and in Europe. J Econ Psychol 27(2006):224–246

Alm J, Cherry T, Jones M, McKee M (2010) Taxpayer information assistance services and tax compliance behavior. J Econ Psychol 31(4):577–586
Alm J, Clark J, Leibel K (2016) Enforcement, socioeconomic diversity, and tax filing compliance in the United States. South Econ J 82(3):725–747
Alm J, Bloomquist KM, McKee M (2017) When you know your neighbour pays taxes: information, peer effects and tax compliance. Fisc Stud 38(4):587–613
Ameyaw B, Dzaka D (2016) Determinants of tax evasion: empirical evidence from Ghana. Mod Econ 07(14):1653–1664
Annis S, Franks J (1989) The idea, ideology, and economics of the informal sector; the case of Peru. Grassroots Dev 13(1):9–23
Arendt L, Grabowski W, Kukulak-Dolata I (2020) County-level patterns of undeclared work: An empirical analysis of a highly diversified region in the European Union. Soc Indic Res 149(1):271–295
Athey S, Imbens GW (2019) Machine learning methods that economists should know about. Annu Rev Econ 11:685–725
Barone G, Mocetti S (2011) Tax morale and public spending inefficiency. Int Tax Public Financ 18(6):724–749
Benk S, Budak T, Yüzba B, Mohdali R (2016) The impact of religiosity on tax compliance among Turkish self-employed taxpayers. Religions 7(4):1–10
Bíró A, Prinz D, Sándor L (2020) Tax evasion and the minimum wage: evidence from Hungary, CERS-IE Working Papers, No. CERS-IE WP—2020/43. Hungarian Academy of Sciences, Institute of Economics, Centre for Economic and Regional Studies. https://scholar.harvard.edu/files/dprinz/files/double_minimum_wage_draft_february2020.pdf. Accessed 1 Dec 2021
Blackwell C (2010) Meta-analysis of incentive effects in tax compliance experiments. In: Alm J, Martinez-Vazquez J, Torgler B (eds) Developing alternative frameworks for explaining tax compliance. Routledge, London, pp 97–112
Bloomquist KM (2003) Tax evasion, income inequality and opportunity costs of tax compliance. In: Proceedings of the annual conference on taxation and minutes of the annual meeting of the National Tax Association, vol 96(2003), 91–104
Bobek DD, Hageman AM, Kelliher CF (2013) Analyzing the role of social norms in tax compliance behavior. J Bus Ethics 115(3):451–468
Boone JP, Khurana IK, Raman KK (2013) Religiosity and tax avoidance. J Am Tax Assoc 35(1):53–84
Boulesteix AL, Schmid M (2014) Machine learning versus statistical modeling. Biom J 56(4):588–593
Breiman L (2001) Random forests. Mach Learn 45:5–32
Castells M, Portes A (1989) World underneath: the origins, dynamics, and effects of the informal economy. In: Portes A, Castells M, Benton LA (eds) The informal economy; studies in advanced and less developed countries. The Johns Hopkins University Press, Baltimore
Chen MA, Vanek J, Carr M (2004) Mainstreaming informal employment and gender in poverty reduction: a handbook for policymakers and other stakeholders. The Commonwealth Secretariat, London, United Kingdom
Chen MA (2012) The informal economy: definitions, theories and policies. WIEGO working paper no. 1. WIEGO
Christie E, Holzner M (2006) What explains tax evasion? An empirical assessment based on European data. wiiw working papers. https://wiiw.ac.at/what-explains-tax-evasion-an-empirical-assessment-based-on-european-data-dlp-540.pdf. Accessed 1 Dec 2021
Clotfelter CT (1983) Tax evasion and tax rates: an analysis of individual returns. Rev Econ Stat 65(3):363–373
Dabare R, Wai Wong K, Koutsakis P, Fairuz Shiratuddin M (2018) A study of the effect of dropout on imbalanced data classification using deep neural networks. J Multidiscip Eng Sci Technol (JMEST) 5(10):2458–9403
Davis M (2006) Planet of slums. Verso
de Soto H (1989) The other path; the invisible revolution in the third world. Harper and Row, New York
Di Franco G, Santurro M (2021) Machine learning, artificial neural networks and social research. Qual Quant 55(3):1007–1025
Dreher A, Kotsogiannis C, McCorriston S (2009) How do institutions affect corruption and the shadow economy? Int Tax Public Financ 16(6):773–796
Dularif M, Sutrisno T, Nurkholis, & Saraswati, E. (2019) Is deterrence approach effective in combating tax evasion? A meta-analysis. Probl Perspect Manag 17(2):93–113
Elek P, Köllő J (2019) Eliciting permanent and transitory undeclared work from matched administrative and survey data. Empirica 46(3):547–576
Engel C, Mittone L, Morreale A (2020) Tax morale and fairness in conflict an experiment. J Econ Psychol 81(March):102314
ESS (2021) European Social Survey, ESS9–2018. http://nesstar.ess.nsd.uib.no/webview/index.jsp?v=2&submode=variable&study=http%3A%2F%2F129.177.90.83%3A-1%2Fobj%2FfStudy%2FESS9e03.1&gs=undefined&variable=http%3A%2F%2F129.177.90.83%3A80%2Fobj%2FfVariable%2FESS9e03.1_V10&mode=documentation&top=yes. Accessed 1 Dec 2021
European Commission (1998) Communication from the Commission on undeclared work (COM(1998) 219 Final)
European Commission (2020) Special Eurobarometer 498. Undeclared work in the European Union. https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/instruments/SPECIAL/surveyKy/2250. Accessed 1 Dec 2021
European Commission (2019) Standard Eurobarometer 91. Public opinion in the European Union
Eurostat (2021a) Implicit tax rate on labour. https://ec.europa.eu/taxation_customs/sites/taxation/files/implicit-tax-rates.xlsx. Accessed 1 Dec 2021
Eurostat (2021b) Inequality of income distribution. https://ec.europa.eu/eurostat/databrowser/view/tespm151/default/table?lang=en. Accessed 1 Dec 2021
Eurostat (2021c) Relative median income ratio (65 +). https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ilc_pnp2&lang=en. Accessed 1 Dec 2021
Eurostat (2021d) Unemployment by sex and age—annual data. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=une_rt_a&lang=en. Accessed 1 Dec 2021
EVS (2021) European values study. https://www.atlasofeuropeanvalues.eu/maptool.html. Accessed 1 Dec 2021
Fegatilli E (2009) Undeclared work in the European Union. What can we learn from an European survey? Centre de Recherche En Economie Publique et de La Population, CREPP working paper 2009-02
Feld LP, Frey BS (2007) Tax compliance as the result of a psychological tax contract: the role of incentives and responsive regulation. Law Policy 29(1):102–120
Feld LP, Larsen C (2012) Self-perceptions, government policies and tax compliance in Germany. Int Tax Public Financ 19(1):78–103
Fernández Hilario A, García López S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) Learning from imbalanced data sets. Springer, Cham
Fields GS (1990) Labour market modelling and the urban informal sector: theory and evidence. In: Turnham D, Salomé B, Schwarz A (eds) The informal sector revisited. Organisation for Economic Co-operation and Development, Paris, pp 49–69
Forteza A, Noboa C (2019) Perceptions of institutional quality and justification of tax evasion. Const Polit Econ 30(4):367–382

Franic J (2019) Explaining workers' role in illegitimate wage underreporting practice: Evidence from the European Union. Econ Labour Relat Rev 30(3):366–381
Franic J (2020a) Repression, voluntary compliance and undeclared work in a transition setting: some evidence from Poland. Post Communist Econ 32(2):250–266
Franic J (2020b) Dissecting the illicit practice of wage underreporting: some evidence from Croatia. Econ Res 33(1):957–973
Franic J (2020c) Why workers engage in quasi-formal employment? Some lessons from Croatia. East J Eur Stud 11(2):94–112
Franic J, Cichocki S (2021) Envelope wages as a new normal? Exploring the supply side of quasi-formal employment in the EU. Empl Relat 44(1):37–53. https://doi.org/10.1108/ER-02-2021-0073
Fraser Institute (2021) Economic freedom of the world. https://www.fraserinstitute.org/economic-freedom/approach. Accessed 1 Dec 2021
Frey BS, Torgler B (2007) Tax morale and conditional cooperation. J Comp Econ 35(1):136–159
Geertz C (1963) Peddlers and princes. The University of Chicago Press, Chicago
Genuer R, Poggi J-M (2020) Random forests with R. In Use R! Springer, Berlin
Gërxhani K, Wintrobe R (2021) Understanding tax evasion: combining the public choice and new institutionalist perspectives. In: Douarin E, Havrylyshyn O (eds) The Palgrave handbook of comparative economics. Palgrave Macmillan, Cham, pp 785–810
Ghoddusi H, Creamer GG, Rafizadeh N (2019) Machine learning in energy economics and finance: a review. Energy Econ 81:709–727
Goel RK, Saunoris JW (2017) Unemployment and international shadow economy: gender differences. Appl Econ 49(58):5828–5840
Gregorio, C. De, & Giordano, A. (2016). *The heterogeneity of undeclared work in Italy: some results from the statistical integration of survey and administrative sources*. Istituto nazionale di statistica.
Grimmer J, Roberts ME, Stewart BM (2021) Machine learning for social science: an agnostic approach. Annu Rev Polit Sci 24(1):395–419
Hart K (1973) Informal income opportunities and urban employment in Ghana. J Mod Afr Stud 2(1):61–89
Hastie T, Tibshirani R, Friedman J (2008) The elements of statistical learning: data mining, inference, and prediction. Springer, Berlin. https://doi.org/10.1007/978-1-4614-4714-6
HMRC (2018) Measuring tax gaps 2018 edition: Tax gaps estimates for 2016–17. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/715742/HMRC-measuring-tax-gaps-2018.pdf. Accessed 1 Dec 2021
Hofmann E, Voracek M, Bock C, Kirchler E (2017) Tax compliance across sociodemographic categories: meta-analyses of survey studies in 111 countries. J Econ Psychol 62:63–71
Horodnic IA, Williams CC (2019) Tackling undeclared work in the European Union: beyond the rational economic actor approach. Policy Stud 0(0):1–35
ILO (1972) Employment, incomes and equality; a strategy for increasing productive employment in Kenya. International Labour Organization, Geneva
ILO (1993) Statistics of employment in the informal sector. In: Fifteenth international conference of labour statisticians (fifteenth international conference of labour statisticians). International Labour Organization
ILO (2002) Decent work and informal economy. In: International labour conference report. International Labour Organization
IRS (2016) Tax gap estimates for tax years 2008–2010. https://www.irs.gov/pub/newsroom/taxgapestimatesfor2008through2010.pdf. Accessed 1 Dec 2021
Islam A, Rashid MHU, Hossain SZ, Hashmi R (2020) Public policies and tax evasion: evidence from SAARC countries. Heliyon 6(11):e05449
Jimenez P, Iyer GS (2016) Tax compliance in a social setting: the influence of social norms, trust in government, and perceived fairness on taxpayer compliance. Adv Account 34:17–26
Johnson JM, Khoshgoftaar TM (2019) Survey on deep learning with class imbalance. J Big Data 6(1):27
Kassa ET (2021) Factors influencing taxpayers to engage in tax evasion: evidence from Woldia City administration micro, small, and large enterprise taxpayers. J Innov Entrep 10(1):8
Katnic M, Williams CC (2018) Diagnostic report on undeclared work in Montenegro. https://www.esap.online/download/docs/ESAP_Diagnosticreportonundeclaredwork_ME.PDF/17aab8c11ceb44e4addb1dc44108ecc2.pdf
Kayaoglu A, Williams CC (2017) Beyond the declared/undeclared economy dualism: evaluating individual and country level variations in the prevalence of under-declared employment. J Econ Manag Perspect 11(4):36–47
Kemme DM, Parikh B, Steigner T (2020) Tax morale and international tax evasion. J World Bus 55(3):101052
Khlif H, Guidara A, Hussainey K (2016) Sustainability level, corruption and tax evasion: a cross-country analysis. J Financ Crime 23(2):328–348
Kogler C, Batrancea L, Nichita A, Pantya J, Belianin A, Kirchler E (2013) Trust and power as determinants of tax compliance: testing the assumptions of the slippery slope framework in Austria, Hungary, Romania and Russia. J Econ Psychol 34:169–180
Kogler C, Muehlbacher S, Kirchler E (2015) Testing the "slippery slope framework" among self-employed taxpayers. Econ Gov 16(2):125–142
Kuehn Z (2014) Tax rates, governance, and the informal economy in high-income countries. Econ Inq 52(1):405–430
Lago-Peñas I, Lago-Peñas S (2010) The determinants of tax morale in comparative perspective: evidence from European countries. Eur J Polit Econ 26(4):441–453
Lewis WA (1954) Economic development with unlimited supplies of labour. Manch Sch Econ Soc Stud 22(2):139–191
Li D, Bledso JR, Zeng Y, Liu W, Hu Y, Bi K, Liang A, Li S (2020) A deep learning diagnostic platform for diffuse large B-cell lymphoma with high accuracy across multiple hospitals. Nat Commun 11(1):6004
Li L, Ma G (2015) Government size and tax evasion: evidence from China. Pac Econ Rev 20(2):346–364
Loayza NV, Oviedo AM, Servén L (2005) The impact of regulation on growth and informality: cross-country evidence. World Bank Policy research paper no. WPS3623. https://openknowledge.worldbank.org/bitstream/handle/10986/8222/wps36230rev.pdf?sequence=1&isAllowed=y. Accessed 1 Dec 2021
Maloney WF (2004) Informality revisited. World Dev 32(7):1159–1178
McKay S (2014) Transnational aspects of undeclared work and the role of EU Legislation. Eur Labour Law J 5(2):116–131
Mitchell TM (1997) Machine learning. McGraw-Hill, New York
Molina M, Garip F (2019) Machine learning for sociology. Ann Rev Sociol 45:27–45
Moser CON (1978) Informal sector or petty commodity production: dualism or dependence in urban development? World Dev 6(9/10):1041–1064
Mullainathan S, Spiess J (2017) Machine learning: an applied econometric approach. J Econ Perspect 31(2):87–106
Palumbo L (2017) Exploiting for care: trafficking and abuse in domestic work in Italy. J Immigr Refug Stud 15(2):171–186
Peattie L (1987) An idea in good currency and how it grew: the informal sector. World Dev 15(7):851–860

Pfau-Effinger B (2009) Varieties of undeclared work in European societies. Br J Ind Relat 47(1):79–99
Picur RD, Riahi-Belkaoui A (2006) The impact of bureaucracy, corruption and tax compliance. Rev Acc Financ 5(2):174–180
Pommerehne WW (1996) Tax rates, tax administration and income tax evasion in Switzerland. Public Choice 88(1–2):161–170
Popescu ME, Cristescu A, Stanila L, Vasilescu MD (2016) Determinants of undeclared work in the EU member states. Procedia Econ Financ 39:520–525
Popescu GH, Davidescu AAM, Huidumac C (2018) Researching the main causes of the romanian shadow economy at the micro and macro levels: Implications for sustainable development. Sustainability (switzerland) 10(10):3518
Portes A, Sassen-Koob S (1987) Making it underground: comparative material on the informal sector in western market economies. Am J Sociol 93(1):30–61
Porthé V, Ahonen E, Vázquez ML, Pope C, Agudelo AA, García AM, Amable M, Benavides FG, Benach J (2010) Extending a model of precarious employment: a qualitative study of immigrant workers in Spain. Am J Ind Med 53(4):417–424
Rakowski CA (1994) Convergence and divergence in the informal sector debate: a focus on Latin America, 1984–92. World Dev 22(4):501–516
Rashid HU, Buhayan SA, Masud AK, Sawyer A (2021) Impact of governance quality and religiosity on tax evasion: evidence from OECD countries. Adv Tax 29:89–110
Rei D, Bhattacharya M (2008) The impact of institutions and policy on informal economy in developing countries. An econometric exploration. Working paper no. 84. International Labour Organization (ILO)
Riahi-Belkaoui A (2004) Relationship between tax compliance internationally and selected determinants of tax morale. J Int Account Audit Tax 13(2):135–143
Richardson G (2006) Determinants of tax evasion: a cross-country investigation. J Int Account Audit Tax 15(2):150–169
Rodgers P, Shahid MS, Williams CC (2019) Reconceptualizing informal work practices: some observations from an ethnic minority community in urban UK. Int J Urban Reg Res 43(3):476–496
Rodrigues MJS (2020). Power and trust as determinants of tax compliance. The slippery slope framework applied to Portugal and Switzerland. Master thesis. Lisbon School of Economics and Management
Round J (2009) The boundaries between informal and formal work. Beyond Current Horizons project. http://www.beyondcurrenthorizons.org.uk/wp-content/uploads/final_roundjohn_informalformaleconomies20090116.pdf. Accessed 1 Dec 2021
Sethuraman SV (1976) The urban informal sector: concept, measurement and policy. Int Labour Rev 114(1):69–81
Shafer WE, Wang Z, Hsieh TS (2020) Support for economic inequality and tax evasion. Sustainability 12(19):1–18
Strielkowski W, Čábelková I (2015) Religion, culture, and tax evasion: evidence from the Czech Republic. Religions 6(2):657–669
Tokman VE (1978) An exploration into the nature of informal—formal sector relationship. World Dev 6(9/10):1065–1075
Torgler B (2004) Tax morale, trust and corruption: empirical evidence from transition countries. Center for Research in Economics, Management and the Arts, Working paper no. 2004–05
Transparency International (2021) Corruption perceptions index. https://www.transparency.org/en/cpi/2019/index/nzl. Accessed 1 Dec 2021
Vallanti G, Gianfreda G (2020) Informality, regulation and productivity: do small firms escape EPL through shadow employment? Small Bus Econ 57:1383–1412
van Dijke M, Verboon P (2010) Trust in authorities as a boundary condition to procedural fairness effects on tax compliance. J Econ Psychol 31(1):80–91
Viloria A, Lezama OBP, Mercado-Caruzo N (2020) Unbalanced data processing using oversampling: machine learning. Procedia Comput Sci 175:108–113. https://doi.org/10.1016/j.procs.2020.07.018
Vorley T, Williams C (2012) Evaluating the variations in undeclared work in the European Union. J Econ Appl 2(2):20–39
Williams CC (2004) Cash-in-hand work; the underground sector and the hidden economy of favours. Palgrave Macmillan, Cham
Williams CC (2007) Tackling undeclared work in Europe: lessons from a study of Ukraine. Eur J Ind Relat 13(2):219–236
Williams CC (2015) Explaining the informal economy: an exploratory evaluation of competing perspectives. Relat Ind 70(4):741–765
Williams CC (2016) Developing a holistic approach for tackling undeclared work. Eur Platf Undeclar Work. https://doi.org/10.2139/ssrn.2937694
Williams CC (2019) Explaining and tackling the informal economy: an evaluation of competing perspectives. Open Econ 2:63–75
Williams CC, Bezeredi S (2017) Tackling the illegal practice of underreporting employees' wages: lessons from the Republic of Macedonia. UTMS J Econ 8(3):243–258
Williams CC, Efendic AS (2020) Evaluating the relationship between migration and participation in undeclared work: lessons from Bosnia and Herzegovina. Econ Altern 26(4):592–606
Williams CC, Efendic A (2021) Evaluating the relationship between marginalization and participation in undeclared work: lessons from Bosnia and Herzegovina. Southeast Eur Black Sea Stud 00(00):1–19
Williams CC, Horodnic I (2015a) Are marginalised populations more likely to engage in undeclared work in the Nordic countries? Sociol Res Online 20(3):1–15. https://doi.org/10.5153/sro.3719
Williams CC, Horodnic IA (2015b) Who participates in the undeclared economy in South-Eastern Europe? An evaluation of the marginalization thesis. South Eastern Europe J Econ 13(2):157–175
Williams CC, Horodnic IA (2016) Evaluating the illegal employer practice of under-reporting employees' salaries. Br J Ind Relat 55(1):1–29
Williams CC, Horodnic IA (2017a) Tackling bogus self-employment: some lessons from Romania. J Dev Entrep 22(2):1–20
Williams CC, Horodnic IA (2017b) Who participates in undeclared work in the European Union? Toward a reinforced marginalization perspective. Int J Sociol 47(2):99–115
Williams CC, Öz-Yalaman G (2021) Re-theorising participation in undeclared work in the European Union: lessons from a 2019 Eurobarometer survey. Eur Soc 1(1):1–25
Williams CC, Round J (2007) Re-thinking the nature of the informal economy: some lessons from Ukraine. Int J Urban Reg Res 31(2):425–441
Williams CC, Yang J (2018) Evaluating competing perspectives towards undeclared work: some lessons from Bulgaria. J Contemp Cent East Europe 26(2–3):247–265
Williams CC, Horodnic IA, Windebank J (2015a) Evaluating the prevalence and distribution of envelope wages in the European Union: lessons from a 2013 Eurobarometer survey. J Contemp Eur Res 11(2):179–195
Williams CC, Horodnic IA, Windebank J (2015b) Explaining participation in the informal economy: an institutional incongruence perspective. Int Sociol 30(3):294–313
Windebank JE, Horodnic IA (2017) Explaining participation in undeclared work in France: lessons for policy evaluation. Int J Sociol Soc Policy 37(3/4):203–217
World Bank (2021) Worldwide governance indicators. http://info.worldbank.org/governance/wgi/Home/Reports. Accessed 1 Dec 2021
Wyner AJ, Olson M, Bleich J, Mease D (2017) Explaining the success of adaboost and random forests as interpolating classifiers. J Mach Learn Res 18:1–33

Yamen A, Allam A, Bani-Mustafa A, Uyar A (2018) Impact of institutional environment quality on tax evasion: a comparative investigation of old versus new EU members. J Int Account Audit Tax 32:17–29

Yitzhaki S (1987) On the excess burden of tax evasin. Public Financ Q 15(2):123–137