Scientific Annals of Economics and Business
70 (1), 2023, 97-114
DOI: 10.47743/saeb-2023-0001

Modelling the Non-Linear Dependencies between Government Expenditures and Shadow Economy Using Data-Driven Approaches

Codruț-Florin Ivașcu*, Sorina Emanuela Ștefoni**

Abstract: This article aims to model the relationship between the size of the shadow economy and the most important government expenditures respectively social protection, health, and education, using nonlinear approaches. We applied four different Machine Learning models, namely Support Vector Regression, Neural Networks, Random Forest, and XGBoost on a cross-sectional dataset of 28 EU states between 1995 and 2020. Our goal is to calibrate an algorithm that can explain the variance of shadow economy size better than a linear model. Moreover, the most performant model has been used to predict the shadow economy size for over 30,000 simulated combinations of expenses in order to outline some possible inflection points after which government expenditures become counterproductive. Our findings suggest that ML algorithms outperform linear regression in terms of R-squared and root mean squared error and that social protection spending is the most important determinant of shadow economy size. Further to our analysis for the 28 EU states, between 1995 and 2020, the results suggest that the lowest size of shadow economy occurs when social protection expenses are greater than 20% of GDP, health expenses are greater than 6% of GDP, and education expenses range between 6% and 8% of GDP. To the best of the authors' knowledge, this is the first paper that used ML to model shadow economy and its determinants (i.e., government expenditures). We propose an easy-to-replicate methodology that can be developed in future research.

Keywords: machine learning; shadow economy; government spending.

JEL classification: C63; H50; E26.

* Bucharest University of Economic Studies, Romania; e-mail: *codrut.ivascu@fin.ase.ro* (corresponding author).
** Bucharest University of Economic Studies, Romania; e-mail: *stefonisorina@gmail.com*.

Article history: Received 2 June 2022 | Accepted 6 December 2022 | Published online 14 March 2023

To cite this article: Ivașcu, C.-F., Ștefoni, S. E. (2023). Modelling the Non-Linear Dependencies between Government Expenditures and Shadow Economy Using Data-Driven Approaches. *Scientific Annals of Economics and Business*, 70(1), 97-114. https://doi.org/10.47743/saeb-2023-0001.

1. INTRODUCTION

Market transactions that are not deliberately declared to the competent authorities of the state represent a high proportion of the GDP of that state. These, along with other informal activities (e.g., criminal activities, domestic activities) constitute the phenomenon of underground economy (hereinafter referred to also as *informal economy* or *shadow economy*). Shadow economy, within a society that prioritizes the co-existence and the well-being of the society, is equivalent to non-assumption and evading responsibility.

The negative repercussions of informal economies are the major damages to government revenue size and the infringement of tax regulations. As such, policy makers conduct strategies to fight the increase of the shadow economy, using all the measures at their disposal. As governments allocate resources collected from society towards different priorities (*e.g.*, education, health, social protection), it is vital to have a better understanding of how their decisions regarding this distribution influence shadow economy.

Just as in an individual's life structuring spending in a way that brings the best results, in a society, there is a need for proper management of tax revenues so that sectors that are vulnerable and of interest can be funded. In economic terms, better financial management of a state will be reflected, among others, in a reduced dent in the underground economy. Corroborating the above two statements, we can say that an efficient distribution of tax revenues should also be reflected in a lower level of the underground economy. Regarding the efficiency of government spending, indicators have also been developed by the World Economic Forum Global Competitiveness Index, placing countries in a range between 1-7. It is in this direction that the present study aims to be developed, to see what the premises are for which we can consider the spending of a state as efficient or not, considering the level of the underground economy.

The laws of worldwide states differ according to each state approach to their public economic strategy and history. As far as EU member states are concerned, the legislative and economic elements are largely homogeneous, based on broadly the same future development prospects. Focusing on the latter category of states will allow us to formulate relevant conclusions that can be generally applicable to them. Moreover, the data on the underground economy are difficult to find from a reliable source, but as concerns the EU member states, we have identified databases that were previously analyzed in the specialized literature and that are considered relevant and reliable, as will be described in this paper.

There is a rich empirical literature that studies the relationship between shadow economy and different casual variables. Various categories of expenditures were also studied in relation to the evolution of the informal economy, in different forms and through different methods. Although well documented, most of the empirical studies use a linear model in order to explain the sensitivity of the shadow economy to different variables. However, the interactions between financial and economic variables are very complex and, in most cases, highly nonlinear.

This paper aims to model the dependencies between three main public expenditures (*i.e.*, social protection expenditure, education expenditure, health expenditure) and shadow economy size using data-driven approaches. To the best of the authors' knowledge, this is the first article to use machine learning to model the informal economy and government expenditures as determinants. Our goal is to calibrate an algorithm that can explain the variation of shadow economy better than a linear model in a real word setting. We are also trying to find out if there are some inflection points after which government expenditures become counterproductive in

decreasing the size of shadow economy, thus finding an optimal level of public spending based on the cross-sectional data of 28 EU states between 1995 and 2020.

Our input to the literature is two-fold: i) empirical, by considering a range of government expenditures describing the economic strategy of the governments, as well as an innovative method to study the link between the underground economy and government spending, namely machine learning; ii) practical, through a series of recommendations.

The sections in this paper are organized as follows: Section 2 presents part of the significant proceedings in our research; Section 3 briefly describes the machine learning algorithms used in this study; Section 4 presents the methodology and Section 5 the results of a linear analysis; Section 6 presents the results of the nonlinear analysis. Conclusions and further research are presented in Section 7.

2. LITERATURE REVIEW

Many determinants of the shadow economy have been analysed over time in the scientific literature. From the most common determinants of tax burden, social security contributions, tax complexity, and uncertainty (Alm *et al.*, 1992; Schneider & Enste, 2002; Smuga *et al.*, 2005; Schneider & Williams, 2013 and others) to perceptual determinants such as quality of public institutions, public services, level of development, regulations, moral tax (Schneider & Enste, 2000; Aruoba, 2010; Schneider *et al.*, 2010 and others) have been intensively analysed by the authors both in relaxing economic contexts and in periods when states have been forced to proceed with certain economic measures that have been controversial in the eyes of society.

If, in terms of fiscal policy, its relationship with the informal economy has often been demonstrated in the literature by authors such as Cebula (1997) or Duncan and Peter (2014), the same is applicable with regard to other government policies. For example, budgetary policy, which is closely related to fiscal policy, has not been addressed as often. In the following, we will set out some of the views that have been expressed on the latter policy.

Other authors have considered that governments can control the willingness of individuals to evade the formal area of the economy. Malaczewska (2013) studies a pattern of the shadow economy and the idea of beneficial government spending, using the sensitivity analysis of Nash equilibrium. The findings also emphasize that if the probability of detecting activities specific to the underground economy increases, then the government can be expected to increase the level of the amounts spent on control institutions (such as the *Agencies for Fiscal Administration*). The author concludes that households encouraged by useful government expenditure will give up underground activities and will migrate to the formal sector.

In the same direction, another view of the relationship between government policy and informal activities is captured by Aruoba (2010). The level of government spending is responsible for the variance in fiscal rates. The government implements strategies to finance the balanced number of expenditures for each sector using revenues collected from imposed taxes.

Government spending was also investigated by Igor and Schneider (2017). The authors show that there is a positive link between government (military) expenditures and the shadow economy in the Baltic states. In the same study, for government expenditures (health), the results indicate a negative relationship in connection with the shadow economy. So military expenditures, being a less transparent category and less visible to the general public, do not

provide (*e.g.*, contrasting with the health expenditures) so many tangible benefits in lowering the agents' inclination to divulge their incomes.

Social protection is a variable that has been considered when it comes to the informal economy. In this sense, the European Union funded a project led by ARS Progetti S.P.A. et al. (2017), the project's purpose being to develop valuable approaches to strengthen social protection among people in the informal economy. Weaker public services negatively influence social perception regarding government policy Kelmanson *et al.* (2019). Mara (2021) argues that in the long run a possible solution to reduce the size of the shadow economy is to increase social protection expenditures, but points out that this solution needs to be accompanied by other important policies, such as reducing corruption, securing property rights, and maintaining a reasonable tax burden.

Surprisingly, with regard to the impact of education on the shadow economy, the results in the literature are contradictory. There are authors who find a positive correlation between education and the size of the shadow economy. Stulhofer (1997) and Hanousek and Palda (2004) have studied the cases of two countries in transition, respectively, Croatia and the Czech and Slovak Republic. Their findings reveal that a higher level of education has an increasing effect on tax evasion, as a component of the informal economy. Furthermore, Torgler *et al.* (2010) reveal in their paperwork a positive correlation between education expenditures and the informal economy. The authors conclude that as the amount of education increases, more opportunities are developing in the shadow economy. This connection is also found by Pang *et al.* (2021), across the provinces of China.

On the other hand, there are authors that observe a negative impact between shadow economy and education expenditures. Some of the important explanations by which such a discrepancy could be revealed are (i) the form in which *education* is quantified in these studies, (ii) the models through which they were studied, and the related factors involved in those models. We note that, as a general rule, for studies on education in terms of cognitive skills, forming values (such as tax morality), or school attendance (Hastie *et al.*, 2009; Gerxhani & van de Werfhorst, 2013), and in studies that also consider factors related to the environment of weak public services, the quality of institutions, and the perception of the people on government efficiency (Buehn & Farzanegan, 2013; Berrittella, 2015), higher levels of education have resulted in decreasing the size of the shadow economy. These mixed results may suggest that other factors, such as other government expenditures, may be important in understanding the final effect of education on the shadow economy.

Tax evasion and informal economy data are limited and hard to quantify because of the choices made by citizens according to their own values and because individuals engaging in informal economic activities often remain undetected. Using the causes and indicators of the informal economy, using the widely known MIMIC method, authors such as Slemrod (2007; 2012), Dell'Anno (2007), Schneider *et al.* (2010), Alm and Embaye (2013), and Medina and Schneider (2018) have managed to quantify the size of the shadow economy. The data collection published by these authors will be used to improve the study here, as will be made clear in the following sections.

Regarding the non-linear relationship between the data, Wu and Schneider (2019) show that there is a U-shaped relationship between the shadow economy and the level of development using a quadratic regression equation. However, they noted a limitation in their study by suggesting the possibility of an asymmetric relationship between variables. In our study we are modelling the nonlinear relationship using nonparametric models. In addition,

we analyze the simultaneous impact of three variables, respectively, social protection, education, and health expenditures.

3. MACHINE LEARNING MODELS

The interactions between financial and economic variables are very complex and, in most cases, highly non-linear. Although many scholars use a linear analysis in their studies to explain the behaviour of provided data in a real-word setting, more performant models are required. Machine learning algorithms are very handy in capturing non-linear dependencies. In this paper, 4 different models have been used, namely *Support Vector Regression*, *Random Forest*, *XGBoost,* and *Neural Networks*. All models have impressive results in financial and economic applications, as in Medeiros et al. (2021), for inflation forecasting, Aziz and Dowling (2019) for risk management, or Ivașcu (2021) for option pricing.

3.1 Support Vector Regression

The first model to which we refer is the Support Vector Regression (SVR). The Support Vector Machine algorithm, with the extension of SVR, has been developed by authors such as Boser *et al.* (1992); Guyon *et al.* (1993); Vapnik (1995); Vapnik *et al.* (1997).

The idea of SVR is based on the computation of a linear regression function in a high-dimensional feature space where the input data are mapped via a non-linear function (kernel). In contrast to OLS, the objective function of SVR is to minimize the coefficients, not the squared error. Instead, the error term is handled within the constraints, where we set the absolute error less than or equal to a specified margin, called the maximum error, ϵ. In practice, these constraints are very restrictive and often fail to account for prediction errors. Two new variables, namely ξ_i, ξ_i^*, were introduced, in order to relax the optimization conditions. Therefore, we will arrive at the formulation presented in Vapnik (1995):

$$minimize\ \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*)$$
$$subject\ to\ \begin{cases} y_i - \langle \omega, x_i \rangle - b \leq \epsilon + \xi_i \\ \langle \omega, x_i \rangle + b - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \quad (1)$$

Here, $\{(x_1, y_1), \dots, (x_l, y_l)\} \subset \mathcal{X} \times \mathbb{R}$ represent the training data and $\mathcal{X}$ represents the multidimensional space determined by the input parameters. The coefficient denoted by ω and b is the constant of linear regression. $C > 0$ represents the trade-off between the reduced slope of the function and the magnitude of the deviation above that will be tolerated. To ensure nonlinearity, a kernel function will be applied. One of the most common kernels is the radial basis function (RBF), respectively $k(x_i, x) = \exp(-\gamma\|x_i - x\|^2)$. An interested reader can check the full derivation in Smola and Schölkopf (2004).

3.2 Random Forest

The Random Forest model is a model that has been introduced by Breiman (2001), and it is one of the most efficient algorithms for both classification and regression tasks. The model is based on bagging principle (Breiman, 1996), an aggregation scheme that (i) generates multiple sets of data by bootstrapping from the original input set, (ii) makes a prediction for each set using the CART model (Breiman *et al.*, 1984) and (iii) aggregates the predictions in a single result.

The CART-split criterion is used in the construction of a single tree to find the best *cut* perpendicular to the axes. At each node in each tree, the best *cut* is selected by optimizing this informative criterion based on the Gini impurity (on classifications) or the square errors of prediction (on regressions).

3.3 Extreme Gradient Boosting (XGBoost)

The XGBoost model developed by Chen and Guestrin (2016) is an efficient and scalable implementation of the Gradient Boosting Machine. Its popularity in the Machine Learning competitions is due to numerous optimizations like (i) the addition of a regularization term that improves the generalization ability, (ii) the multithreading parallel computing, which increases the speed over 10 times according to Chen and Guestrin (2016), and (iii) the efficiency of dealing with missing data.

To train the model, the following optimization function must be minimized:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2}$$

where l is the cost differentiable function that measures the difference between the prediction $\hat{y}_i$ and the target value y_i and Ω, is a function of penalizing the complexity of the tree. Intuitively, the objective function will choose the model with the best prediction and the lowest complexity.

3.4 Deep Neural Network

A deep neural network is a set of interconnected processing nodes whose functionality is based on an animal's neural network. It was first introduced by McCulloch and Pitts (1943). The processing ability of the network is determined by the weights given to each node. Weights are obtained from a learning process (or adaptation) from a set of training data. According to Hornik *et al.* (1989), a neural network is a universal approximator, and any function can be modelled by using enough neurons.

Any neural network model presents three different types of layer: an input layer in which we have the explanatory variable, one or more hidden layers, and an output layer. Each layer contains many neurons. The functionality of an individual neuron is simple and direct. Each neuron sums all the signals sent to it, adds a bias term, and performs a non-linear transformation through an activation function. The activation (transfer) function is increasingly monotonic, most often a logistic, hyperbolic, or ReLu-type tangent. The signal transformed into a neuron is forwarded by a certain weight to another neuron in another layer, and the process is repeated. This step is called a follow-up step. The processing power of the network is determined by the weights given to each neuron, which are computed using the backpropagation method – for more details, see Rumelhart *et al.* (1986).

4. DATA DESCRIPTION

This research is based on a cross-section analysis of the 28 EU Member States, including the United Kingdom, for the period 1995-2020 that has been performed to determine the sensitivity of shadow economy size with respect to public expenditures. The general model specification is as follows:

$$SE = f\,(social\ protection, education, and\ health\ expenditures) \quad (3)$$

where f is a linear and nonlinear function, respectively.

Data on the size of the shadow economy are from the data set developed by Medina and Schneider (2019); that is, following the research undertaken by the authors, the largest existing data set on the size and trends of the shadow economy in 158 countries all over the world from 1995 to 2020. Medina and Schneider (2019) use the macroeconomic multiple indicators multiple causes model (MIMIC) and the currency demand approach (CDA) to estimate the size of the shadow economy. As the data provided by the mentioned authors is restricted to 2017, for years 2018, 2019 and 2020 the data were predicted using the ARIMA model, following the Box-Jenkins methodology. The indicator is expressed as a percentage of GDP, and, as also mentioned above, it represents the most comprehensive database of the underground economy, both from the perspective of the sample of states concerned and of the elements that are captured by it.

Data on social protection, education and health expenditures have been provided by the European Commission (*Eurostat database*) and variables are quantified as percentage of GDP. These variables reflect the attention given by the state to certain sectors. For the interpretation of the results, the perception of the citizens of budgetary policies is also important. The descriptive statistics of the variables are presented in Table no. 1.

Table no. 1 – Summary statistics for EU-28

	Observations	Mean	Std. Dev.	Minimum	Maximum
Shadow economy	700	18.43%	6.98%	6.40%	35.30%
Social protection	700	15.99%	4.06%	7.10%	25.50%
Education	700	5.14%	0.96%	2.80%	7.30%
Health	700	5.86%	1.43%	1.80%	8.90%

Source: own estimation

Furthermore, Figure no. 1 presented in the following represents a scatter plot matrix between the variables of interest. On the secondary diagonal, we show the distribution of each variable and above the diagonal the pairwise Pearson correlation. As it might be seen, all correlations are small or moderate (less than 0.7), with a negative direction between social protection and health, on the one hand, and the shadow economy, on the other. Education expenditures appear to be not correlated with any of the variables. The correlation implied a low level of multicollinearity. However, to confirm this, we applied a second method of multicollinearity test, VIF analysis. According to Nachane (2006), VIF values of not more than 10 are accepted as low levels of multicollinearity. The results presented in Table no. 2 confirm that this is applicable to the sample in question.

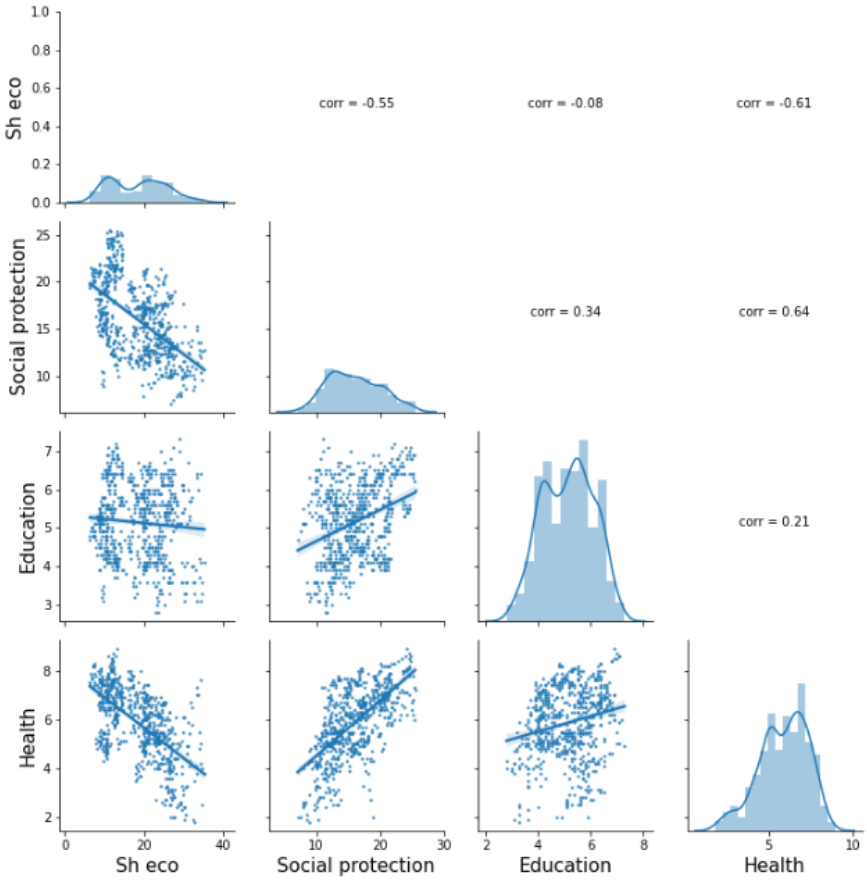

Figure no. 1 – Scatter plot matrix of the variables of interest
Source: own estimation

Table no. 2 – Results of the VIF analysis

Variable	Coefficient Variance	Uncentered VIF	Centred VIF
Social Protection	0.0026	1.9265	1.9322
Health	0.0148	1.7728	1.2151
Education	0.0397	2.2724	1.1248
C	2.1699	2.3709	

Source: own estimation

In order to check the causality relationship among variables, we conducted a panel Granger causality test. The results of Table no. 3 validate the causality between the shadow economy and the factors analyzed. It is worth noting the existence of the bidirectional relationship between shadow economy size and social protection and health expenditure. On this note, Mara (2021) obtained in a recent paper the same bidirectional relationship between shadow economy and public expenditure. Schneider (2006) argues that an increase of the

shadow economy can lead to reduced state revenues, which in turn reduce the quality and quantity of publicly provided goods and services, with the consequence of even stronger incentives to participate in the shadow economy.

Table no. 3 – Results of panel Granger causality test

Null Hypothesis:	F-Statistic	Result	Conclusion
Social Protection does not Granger Cause Shadow Economy	5.3837***	Yes	**SP → SE**
Shadow Economy does not Granger Cause Social Protection	11.7129***	Yes	SE → SP
Education does not Granger Cause Shadow Economy	13.8180***	Yes	**E → SE**
Shadow Economy does not Granger Cause Education	0.2496	No	
Health does not Granger Cause Shadow Economy	12.4888***	Yes	**H → SE**
Shadow Economy does not Granger Cause Health	7.0464***	Yes	SE → H
Education does not Granger Cause Social Protection	2.8823*	Yes	E → SP
Social Protection does not Granger Cause Education	1.4572	No	
Health does not Granger Cause Social Protection	1.8958	No	
Social Protection does not Granger Cause Health	1.8754	No	
Health does not Granger Cause Education	6.9022***	Yes	H → E
Education does not Granger Cause Health	2.1007	No	

Note: ***, ** and * indicate the significance at 1%, 5%, and 10% levels, respectively.
Source: own estimation

5. LINEAR ANALYSIS

As the linear analysis is widely used in shadow economy studies, we also conducted such analyses, in order to compare the results obtained herein with the ones to be obtained using a non-linear approach. In this regard, to analyze the effect of public expenditures on shadow economy, we apply the following methodology. Based on the direction in which our study is heading, namely, quantifying a possible link between certain types of government expenditure and the shadow economy, we outlined the following form of linear regression:

$$Shadow\ Economy_t = \beta_0 + \beta_1 * Social\ Protection_t + \beta_2 * Education_t + \beta_3 * Health_t + \epsilon_t \quad (4)$$

Given the panel set-up, a modelling concern is related to the heterogeneity across time and countries, that is the choice between a fixed-effects and random-effects specification, compared to a simple Pooled OLS specification. First, we start by conducting the Breusch-Pagan LM test to check if the variance components (period and cross-section) have significant effects. According to this test, only cross-sectional effects are statistically significant, suggesting that a random-effects specification is preferred to a pooled OLS model. Hence, we proceed by executing the Hausman test for selecting between fixed- and random-effects models. When testing the null hypothesis, the chi-square test is consistently zero, implying an estimation of negative variance. According to Psychoyios *et al.* (2021) this is not uncommon, provided that the Hausman statistic can be negative even asymptotically. The testing results support the use of the random-effects specification over fixed-effects, as we fail to reject the null hypothesis.

Table no. 4 – Random effects estimates
(dependent variable: size of the shadow economy, % of GDP)

Intercept	Social Protection Expenditure	Education Expenditure	Health Expenditure
15.1885***	-0.1249***	2.8899***	-1.6406 ***
(1.47)	(0.05)	(0.19)	(0.12)

Note: Std. Errors are displayed in parentheses under the coefficient estimates.
*** Statistically significant at the 1% level.
Source: own estimation

Table no. 4 reports the results of the Random Effects specification of the multiple regression model of Equation (3). All variables are statistically significant at the 1% level. At first glance, it is clear that social protection and health expenditures have a negative relationship with shadow economy size. An increase in social protection expenditure by 1 percentage point will decrease the size of the shadow economy by almost 0.12 percentage points, and an increase in health expenditure will decrease the shadow economy by more than 1.64 percentage points. These results agree with studies conducted by other authors who have confirmed the negative relationship between health expenditure and the shadow economy (Igor & Schneider, 2017; Kelmanson *et al.*, 2019 and others). Health and social protection services are directly related to the informal economy through the impact they have on citizens. According to our results, a state with a functioning social system that is also investing in health means a state with a declining level of the shadow economy.

On the other hand, there is a direct relationship between public expenditures on education and the informal economy: when public expenditures on education increase by 1 percentage point, the informal economy will also increase by almost 2.89 percentage points. The result is expected, given that factors that reflect the public perception on public institutions or on government efficiency, as well as socio-cultural factors, have been disregarded. Also, this positive correlation confirms the results of other authors such as Stulhofer (1997); Hanousek and Palda (2004); Torgler *et al.* (2010); Pang *et al.* (2021). One explanation for this result may be that in the case of countries that are either developed or developing, the level of education is already high, the citizens having internal values that come from both school and home (including self-education). If education expenditure increases (and thus it would have positively influenced the educational level of citizens), citizens will benefit from the education received in a satisfactory way. In the absence of perceptual indicators, they will not shy away from underground activities. In addition, a better education often comes with finding barely legal and resourceful approaches to tackle an activity that is part of the underground economy.

Figure no. 2 shows the scatter plot between the estimated values of the size of the shadow economy and the observed values. As might be seen, the errors of prediction are quite high, the differences between estimated points and estimated regression line being over 15 percentual points. Public spending on social protection, education, and health does explain only 42.35% of shadow economy variance. Naturally, this aspect is to be expected, as several other factors influence the shadow economy, especially perceptual factors, which have been studied by many authors in the literature (tax morality, fiscal policy, human development, the quality of public institutions, etc.). However, this could also suggest that there is not a linear relationship between public expenditures and shadow economy size, suggesting that previous results may be questioned. To check if this is the case, we applied the *Ramsey Regression Equation*

Specification Error Test (RESET) test to see whether non-linear combinations of the fitted values help explain the response variable. We obtain an F-statistic of 6.55 (p-value=0.0109) which indicates the possibility of a nonlinear relation. Therefore, a more advanced analysis is required.

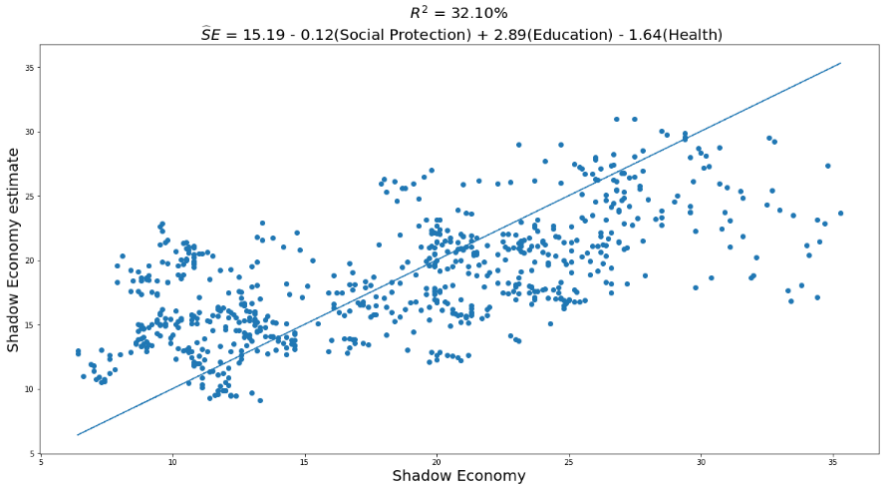

Figure no. 2 – Scatter plot of shadow economy estimates vs observed values
Source: own estimation

6. NONLINEAR ANALYSIS

To select the best heuristic method, the following methodology has been employed. The data set has been randomly divided into 2 different sets, respectively train set (in sample data) and test set (out of sample data), with 80% of the data in the train set and 20% in the test set. The models used in this analysis are presented in Section 3. Linear regression has been considered as a reference point. The neural network (NN) model has 3 hidden layers with 100, 70 and 30 neurons and the activation function is represented by ReLu. XGBoost (XGB) and Random Forest (RF) have a maximum depth in tree construction of 10 and Support Vector Regression (SVR) a tolerance of 10^{-6} and kernel represented by radial bias function.

Table no. 5 shows the root mean squared error (RMSE) and R-squared values for both the in-sample and the out-of-sample data for each model. All machine learning algorithms outperformed the dummy linear regression for both in-sample and out-of-sample data. Support Vector Regression has comparable results with a linear regression suggesting that the data could not be modelled in a linear way even in higher dimensions. Neural Networks model does not significantly reduce the error of prediction. However, a better architecture could increase the R-squared value. Decision tree algorithms, respectively, XGBoost and Random Forest clearly outperform the other methods, XGBoost being the best. Due to their mathematical formulation, the models fit almost perfectly the in-sample data. The out-of-sample RMSE is more than 2 times lower than the one from linear regression, suggesting that a heuristic approach could model better the dynamics between public spending and shadow economy size.

To determine the optimal public spending that can decrease the size of parallel economy, the most performant model, respectively, the XGBoost model has been selected. In order to

increase the accuracy of the algorithm, another training has been performed using all cross-sectional data sets. Figure no. 3 shows the most important variables (features) in tree construction (please see Hastie *et al.*, 2009, p. 367 for technical details). The model assigns the most relevance to social protection expenses, almost the same as combined education and health expenses. Therefore, a fiscal policy focused on social protection could further reduce the shadow economy. This can be explained by the fact that the social sector is, in essence, perhaps the closest to the people. A well-organized social sector, in which people are offered employment and vocational training opportunities, as well as help with the problems they face (unemployment, disability, old age), in which they are assured that they will not be discriminated and that they will receive equal treatment, will lead to a situation in which citizens will have neither time nor need to seek informal economy activities.

Table no. 5 – Root mean squared error and R-squared results for both train sets (in sample) and test sets (out of sample) sets for machine learning algorithms

	RMSE (%)		R-squared	
	In sample	**Out sample**	**In sample**	**Out sample**
Linear	5.28	5.98	32.17%	31.69%
NN	4.94	4.65	56.13%	55.34%
XGB	0.02	2.12	99.99%	83.12%
RF	1.16	2.93	92.62%	79.42%
SVR	5.18	4.69	43.40%	42.46%

Source: own estimation

In order to investigate the non-linear interactions between public expenses and shadow economy size, we simulate over 30,000 combinations of expenses rate (% of GDP) for the analysed variables: social protection, education, and health. Using XGBoost model trained on a full cross-sectional dataset, we have estimated the shadow economy size for each of the simulated combinations. On the basis of this heuristic, we can identify which combination of variables minimizes the size of the informal economy. Please note that estimated values are based on historical data of 28 EU countries for 25 years; therefore, some biases could occur due to sociocultural and economic differences between countries.

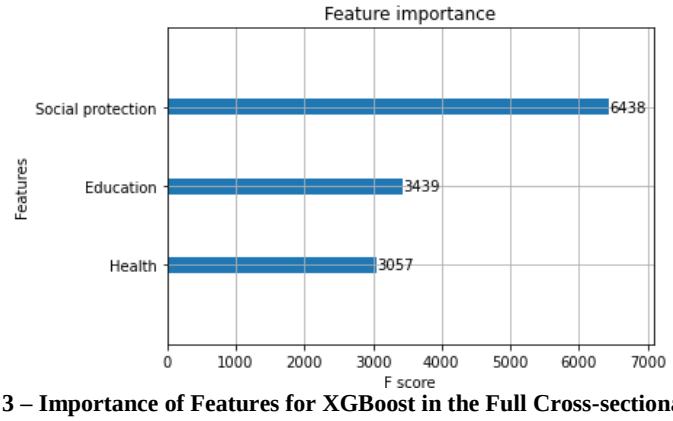

Figure no. 3 – Importance of Features for XGBoost in the Full Cross-sectional Data Set
Source: own estimation

Figure no. 4 shows the 3D scatter plot of the combinations of simulated expenses. The points have been colored on the basis of estimated values of shadow economy size. It is worth noting that health expenses less than 4% of GDP determine higher rates of informal economy and, in combination with a policy that does not favor social protection, a bigger parallel economy may develop. This was the case of Romania, Bulgaria and Cyprus in the 1990s and early 2000s, when shadow economy size represented over 30% of GDP. It seems that a higher level of social protection (over 15% of GDP) minimizes the informal economy, but this is not always true. For example, Ireland, Slovakia, or the Czech Republic report a small level of shadow economy even if public spending on social protection is relatively low. However, more than 5% of GDP must be allocated to health expenses.

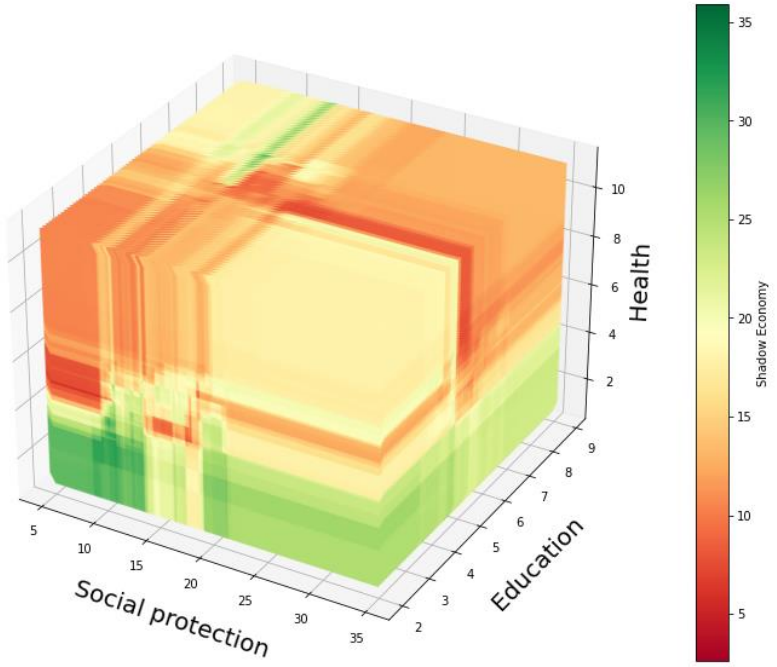

Figure no. 4 – Shadow economy size for over 30,000 simulations of combinations of expenses (% of GDP)
Source: own estimation

Regarding education, the results are debatable, depending on health and social protection policies. For example, high public spending on education reduces shadow economy only if social protection and health expenses are also high (for example, the cases of Sweden, Denmark, Finland, etc.). However, if social protection expenses are low, the shadow economy will increase. This is the case of Estonia in the 1990s, when an informal economy of more than 27% of GDP was reported, even if the government had allocated more than 7% of GDP to the education sector. On the other hand, an allocation smaller than 6% will negatively impact shadow economy size even if social protection is high, as in Italy or Greece.

The model suggests two scenarios in which the shadow economy registers the lowest values and is represented by red in our cube. The first minimum point of the underground economy is reached when education expenditures represent around 6-8% of GDP with high health and social protection expenditures. The second area in which the shadow economy reaches a low level is when education expenditures represent around 3-4% of GDP, with a low level of social spending between 10-15% and a level of health spending between 4-6%. In contrast, the highest levels of the underground economy occur at the lowest spending allocations in the three areas: social protection, education, and health.

This different approach on the relationship between shadow economy and government expenditures is welcomed in the literature because, especially in terms of education expenditures and shadow economy, the results show a contentious linear relationship. It is important to note that even in our research, the results regarding the influence of education expenditures are slightly dispersed, but useful directions can be outlined, most notably considering the other two categories of government expenditures, namely health and social protection expenditures.

To check the robustness of the methodology, the share of government expenditures quantified as percentage of GDP has been replaced by government expenditures quantified as euros per capita. Table no. 6 confirms the higher predictability performance of the ML algorithms compared to a linear regression in both the RMSE and the R-squared value. Again, decision tree-based models are the most performant with XGBoost having the lowest error and highest R-squared, which is more suited to model this nonlinearity. The same simulation methodology has been applied to see the relationship between variables.

Table no. 6 – Root mean squared error and R-squared results for both train sets (in sample) and test sets (out of sample) sets for machine learning algorithms (expenditures per capita)

	RMSE (%)		R-squared	
	In sample	**Out sample**	**In sample**	**Out sample**
Linear	4.22	4.36	56.35%	51.71%
NN	3.50	3.61	78.97%	71.33%
XGB	0.02	3.19	99.99%	82.29%
RF	1.54	3.21	94.88%	81.76%
SVR	4.02	4.10	70.27%	65.32%

Source: own estimation

Upon comparing the two scenarios, Figure no. 5 shows similar conclusions to Figure no. 4. In this second scenario, the lowest values of shadow economy are recorded when education expenditures are around 3000 € per capita (equivalent of the 6-8%% of GDP value in the first model), health expenditure above 2000 € (equivalent of the 6% of GDP value in the first model), and social protection over 10,000 € per capita (equivalent of the 20% of GDP value in the first model). Again, social protection expenses need to be understood in relation with the other variables, shadow economy levels being high regarding social protection costs whenever health expenses are less than 1000 € per capita.

All things considered, in terms of government policy, the way states manage budget expenditure influences taxpayers in their decision whether or not to comply with legal rules. From this perspective, the attention paid to them by governments, quantified by the level of budgetary expenditure distributed, will influence taxpayers' behavior. The main reason for this is that taxpayers are directly interested in how much of the taxes they pay go back to them, either through social benefits, better quality services, or a more innovative educational system.

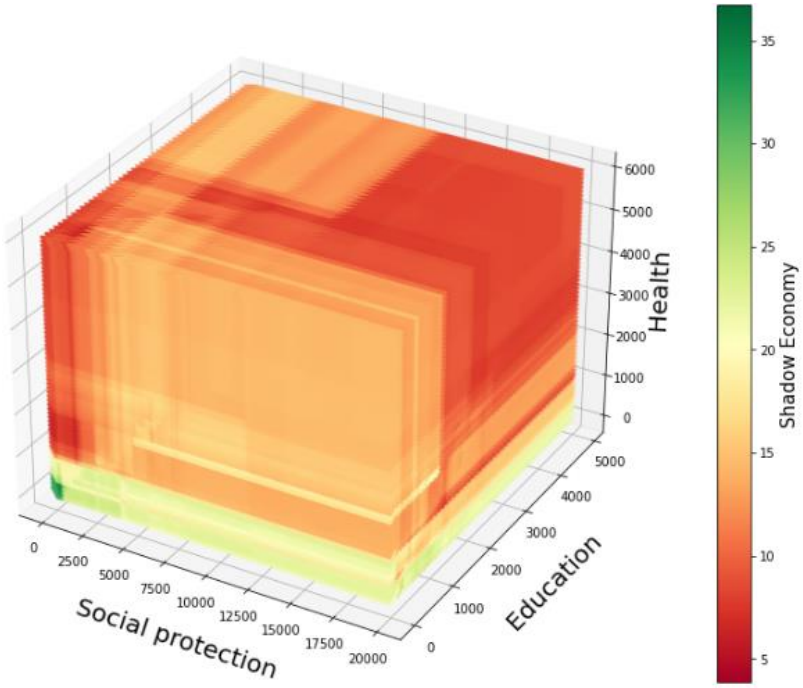

Figure no. 5 – Shadow economy size for over 30,000 simulations of combinations of expenses (€ per capita)
Source: own estimation

Thus, to reduce the level of the underground economy, states must consider an optimal distribution of budget expenditures. Our recommendation, following the analysis carried out in this paper, is that the government should not neglect the 3 important categories of expenditure, namely: education, health, and social protection. A government policy that focuses on social protection (around 20% of GDP), as this is the direction that is closest to the taxpayer and most easily perceived, but also maintains a moderate focus on education (6%-8% of GDP) and health (around 6% of GDP), will lead to a decrease in the informal economy. However, depending on certain unforeseen events that may occur, these efficiency points may fluctuate (such as the emergence of the coronavirus, which required increased attention and resources for the health sector). Moreover, for greater accuracy, the methodology presented by us in an improved and individualized version could be used by governments to generate their own analysis of the level of expenditure they should consider for a reduced level of the underground economy.

7. CONCLUSIONS

In summary, the general conclusion of our study is that government expenditures could influence the size of the shadow economy. We tried to model this influence using both linear models and data-driven approaches, and we showed that Machine Learning algorithms are suitable for this task and that they can explain much better the variation of shadow economy size than a linear model. Furthermore, we presented a methodology that can heuristically detect

optimal allocation of the public budget. Our findings suggest that social protection expenses greater than 20% of GDP, health expenses greater than 6%, and education expenses between 6% and 8% of GDP determine the lowest size of the shadow economy (for the EU member states).

The direction of our study is to confirm that the distribution of budget expenditures by states influences the level of the underground economy, on the one hand, and on the other hand, this paper intends to propose levels that can make the decrease of the underground economy more efficient. The key point of the work lies in the innovative method that we applied in studying the relationship between budget expenditures (social protection, health, and education) and the underground economy, a method that can be processed by states and could be used by them for analysis budget strategies to be implemented.

Ultimately, we also mention the directions in which our study can be further developed. An analysis with more countries considered or on a longer time frame should be enhanced so as to confirm or contradict our conclusions. Multiple variables can be added as control to increase the performance of machine learning models, such as sociocultural or economic indicators, or dummy variables that can isolate big events such as crises or EU policies. Also, different algorithms could be used in modelling, such as recurrent neural networks or memory models, to capture the dynamic of shadow economy.

Acknowledgements

We want to thank Brasoveanu Iulian Viorel, Dragota Victor, and Negrea Bogdan Cristian, Ph.D. professors at Bucharest University of Economic Studies, for useful discussions and support.

References

Alm, J., & Embaye, A. (2013). Using Dynamic Panel Methods to Estimate Shadow Economies Around the World, 1984–2006. *Public Finance Review, 41*(5), 510-543. http://dx.doi.org/10.1177/1091142113482353

Alm, J., Jackson, B., & McKee, M. (1992). Institutional uncertainty and taxpayer compliance. *The American Economic Review, 82*(4), 1018-1026.

ARS Progetti S.P.A., LATTANZIO Advisory, & AGRER. (2017). *Extending coverage: social protection and the informal economy*. Research, Network and Support Facility. Brussels.

Aruoba, S. B. (2010). *Informal sector, government policy and institutions*. Paper presented at the 2010 Meeting Papers from Society for Economic Dynamics.

Aziz, S., & Dowling, M. (2019). Machine learning and AI for risk management. In T. Lynn, Mooney, J., Rosati, P., Cummins, M. (Ed.), *Disrupting finance* (pp. 33-50): Palgrave Pivot, Cham. http://dx.doi.org/10.1007/978-3-030-02330-0_3

Berrittella, M. (2015). The effect of public education expenditure on shadow economy: A cross-country analysis. *International Economic Journal, 29*(4), 527-546. http://dx.doi.org/10.1080/10168737.2015.1081259

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). *A training algorithm for optimal margin classifiers*. Paper presented at the Proceedings of the fifth annual workshop on Computational learning theory.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140. http://dx.doi.org/10.1007/BF00058655

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. http://dx.doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees. Brooks* (Vol. 37). New York: Routledge. http://dx.doi.org/10.1201/9781315139470

Buehn, A., & Farzanegan, M. R. (2013). Impact of education on the shadow economy: Institutions matter. *Economic Bulletin, 33*, 2052-2063.

Cebula, R. J. (1997). An empirical analysis of the impact of government tax and auditing policies on the size of the underground economy: the case of the United States, 1973–1994. *American Journal of Economics and Sociology, 56*(2), 173-185. http://dx.doi.org/10.1111/j.1536-7150.1997.tb03459.x

Chen, T., & Guestrin, C. (2016). *Xgboost: A scalable tree-boosting system.* Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Dell'Anno, R. (2007). The Shadow Economy in Portugal: An Analysis with the MIMIC Approach. *Approach. Journal of Applied Econometrics, 10*(2), 253-277. http://dx.doi.org/10.1080/15140326.2007.12040490

Duncan, D., & Peter, S. K. (2014). Switching on the lights: do higher income taxes push economic activity into the shade? *National Tax Journal, 67*(2), 321-349. http://dx.doi.org/10.17310/ntj.2014.2.02

Gerxhani, K., & van de Werfhorst, H. (2013). The Effect of Education on Informal Sector Participation in a Post-Communist Country. *European Sociological Review, 29*, 446-476. http://dx.doi.org/10.1093/esr/jcr087

Guyon, I., Boser, B., & Vapnik, V. (1993). Automatic capacity tuning of very large VC-dimension classifiers. *Advances in Neural Information Processing Systems*, 147-155.

Hanousek, J., & Palda, F. (2004). Quality of government services and the civic duty to pay taxes in the Czech and Slovak Republics and other transition countries. *Kyklos, 57*(May), 237-252. http://dx.doi.org/10.1111/j.0023-5962.2004.00252.x

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*: Springer Science & Business Media http://dx.doi.org/10.1007/978-0-387-84858-7

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359-366. http://dx.doi.org/10.1016/0893-6080(89)90020-8

Igor, F., & Schneider, F. (2017). Military expenditures and shadow economy in the Baltic States: Is there a link? *Munich Personal RePEc Archive,* (76194). Retrieved from https://mpra.ub.uni-muenchen.de/76194/

Ivaşcu, C. F. (2021). Option pricing using Machine Learning. *Expert Systems with Applications, 163*(January), 113799. http://dx.doi.org/10.1016/j.eswa.2020.113799

Kelmanson, B., Kirabaeva, K., Medina, L., Mircheva, B., & Weiss, J. (2019). *Explaining the shadow economy in Europe: size, causes and policy options*: International Monetary Fund Working Paper.

Malaczewska, P. (2013). Useful government expenditure influence on the shadow economy. *Quantitative Methods in Economics, XIV*(2), 61-69.

Mara, E. R. (2021). Drivers of the shadow economy in European Union welfare states: A panel data analysis. *Economic Analysis and Policy, 72*(December), 309-325. http://dx.doi.org/10.1016/j.eap.2021.09.004

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics, 5*(4), 115-133. http://dx.doi.org/10.1007/BF02478259

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics, 39*(1), 98-119. http://dx.doi.org/10.1080/07350015.2019.1637745

Medina, L., & Schneider, F. (2018). *Shadow economies around the world: what did we learn over the last 20 years?* : International Monetary Fund. http://dx.doi.org/10.5089/9781484338636.001

Medina, L., & Schneider, F. (2019). Shedding light on the shadow economy: A global database and the interaction with the official one. *Munich Society for the Promotion of Economic Research*. http://dx.doi.org/10.2139/ssrn.3502028

Nachane, D. M. (2006). *Econometrics: Theoretical foundations and empirical perspectives*. New Delhi: Oxford University Press.

Pang, J., Li, N., Mu, H., & Zhang, M. (2021). Empirical analysis of the interplay between shadow economy and pollution: With panel data across the provinces of China. *Journal of Cleaner Production, 285*, 124864. http://dx.doi.org/10.1016/j.jclepro.2020.124864

Psychoyios, D., Missiou, O., & Dergiades, T. (2021). Energy based estimation of the shadow economy: The role of governance quality. *The Quarterly Review of Economics and Finance, 80*(May), 797-808. http://dx.doi.org/10.1016/j.qref.2019.07.001

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagation errors. *nature, 323*(6088), 533-536. http://dx.doi.org/10.1038/323533a0

Schneider, F. (2006). Shadow Economies and Corruption All over the World: What Do We Really Know? IZA Discussion Papers 2315. Institute of Labor Economics. Retrieved from IZA Discussion Papers website: https://www.iza.org/publications/dp/2315/

Schneider, F., Buchn, A., & Montenegro, C. E. (2010). New estimates for the shadow economies all over the world. *International Economic Journal, 24*(4), 443-461. http://dx.doi.org/10.1080/10168737.2010.525974

Schneider, F., & Enste, D. H. (2000). Shadow Economies: Size, Causes, and Consequences. *Journal of Economic Literature, 38*(1), 77-114. http://dx.doi.org/10.1257/jel.38.1.77

Schneider, F., & Enste, D. H. (2002). *The shadow economy: Theoretical approaches, studies, and political implications*. Cambridge: Cambridge University Press.

Schneider, F., & Williams, C. (2013). The shadow economy. Retrieved from http://dx.doi.org/10.13140/2.1.1324.1286

Slemrod, J. (2007). Cheating Ourselves: The Economics of Tax Evasion. *The Journal of Economic Perspectives, 21*(1), 25-48. http://dx.doi.org/10.1257/jep.21.1.25

Slemrod, J., & Weber, C. (2012). Evidence of the Invisible: Toward a Credibility Revolution in the Empirical Analysis of Tax Evasion and the Informal Economy. *International Tax and Public Finance, 19*(1), 25-53. http://dx.doi.org/10.1007/s10797-011-9181-0

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing, 14*(3), 199-222. http://dx.doi.org/10.1023/B:STCO.0000035301.49549.88

Smuga, T., Burzyński, W., Karpińska-Mizielińska, W., Marzec, A., Niemczyk, J., & Ważniewski, P. (2005). *Metodologia badań szarej strefy na rynku usług turystycznych*. Warszawa: Instytut Koniunktur i Cen Handlu Zagranicznego.

Stulhofer, A. (1997). Sociocultural aspects of the informal economy - between opportunism and mistrust. *Financial Practice, 21*, 125-140.

Torgler, B., Schneider, F., & Schaltegger, C. (2010). Local autonomy, tax morale, and the shadow economy. *Public Choice, 144*, 293–321. http://dx.doi.org/10.1007/s11127-009-9520-1

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer. http://dx.doi.org/10.1007/978-1-4757-2440-0

Vapnik, V., Golowich, S. E., & Smola, A. J. (1997). *Support vector method for function approximation, regression estimation, and signal processing*. Cambridge: MIT Press, Cambridge.

Wu, D. F., & Schneider, F. (2019). *Nonlinearity between the shadow economy and level of development*: Institute of Labor Economics (IZA).