sciendo

Applying Machine Learning Techniques to Estimate the Size of the Romanian Shadow Economy

Andreea-Daniela IVAN
Bucharest University of Economic Studies, Romania
ivanandreea19@stud.ase.ro

Adriana AnaMaria DAVIDESCU
Bucharest University of Economic Studies/National Scientific Research Institute for Labour and Social Protection, Bucharest, Romania
adriana.alexandru@csie.ase.ro

Marina-Diana AGAFIȚEI*
Bucharest University of Economic Studies/National Scientific Research Institute for Labour and Social Protection, Bucharest, Romania
**Corresponding author, diana.agafiei@csie.ase.ro*

Maria Cristina GEAMBAȘU
Bucharest University of Economic Studies, Bucharest, Romania
Geambasumaria14@stud.ase.ro

Abstract. *The research investigates the underground economy in Romania, an essential phenomenon that affects tax revenue, competitive conditions in the marketplace, and economic stability. This empirical research estimates the informal economy using the most recent machine learning algorithms, with a focus on support vector regression (SVR) In our application, support vector regression (SVR) provided estimates of the coefficients needed for estimating the currency demand approach (CDA) and measuring the size of unobserved economic activities. The literature has examined different forms of measurement of the informal economy: direct approaches, indirect approaches, and statistical approaches such as the CDA and multiple-indicator multiple-cause (MIMIC) model. The classical methods have a number of issues, such as the numerous assumptions and poor data availability. Artificial intelligence (AI) or Machine learning (ML) can offer even more potentially accurate measures than what the classical methods have offered in the past. This paper integrates measures of economic indicators, such as cash in circulation, tax revenues, inflation and unemployment rates, with measures of digitalisation including numbers of cards, ATMs and POS terminals. In addition to SVR, other machine learning algorithms, such as Random Forests and Gradient Boosting, were estimated in order to evaluate which type of model produces the best predictive models of liquidity demand. Overall, the evidence establishes that SVR is a clear improvement over standard methodologies because it models complex and nonlinear relationships in the data. The research illustrates the acute fluctuations registered by Romania's underground economy from 2000 to 2023, thus illustrating how economic crises and changes in fiscal policies dictate the amount of informal economic activity. The paper suggests improved calculation methods resulting in better estimates of natural economic activity and provides ongoing information for economic forecasters and policymakers wishing to maintain the level of informal activities at acceptable levels.*

Keywords: underground economy, machine learning, Romania, artificial intelligence, currency demand, informal sector.

DOI: 10.2478/picbe-2025-0196

Introduction

The underground economy is a complex, ongoing phenomenon that continues to present challenges for policymakers and economic researchers across the globe. The underground economy includes any sort of economic activity that is being concealed from government authorities in order to evade taxes and regulations. These economies or activities significantly undermine government revenues, contribute to an inconsistency in the regulatory framework of competition, and can destabilize economic activity. Therefore, understanding the underground economy, estimating what share it accounts for, and finding ways to policy-wise alter the effects of these economies are essential.

In Romania, the underground economy is governed by social, economic, political, and institutional factors. The most common methods of estimation are household surveys as well as indirect methods such as the cash demand method, and the MIMIC (multiple-indicator multiple-cause) model. However, these traditional methods, which rely on strict assumptions, fail to fully capture the complexity and dynamics of the informal economy.

This study offers a new perspective on the estimation of the informal economy, namely through the use of machine learning approaches, specifically the Support Vector Regression technique (SVR). SVR can help identify nonlinear relationships among the determinants of key economic indicators and cash demand, and provide a more robust, data-driven estimation mechanism. Additionally, the incorporation of the digitalization indicators, as measured by the number of bank cards, and the presence of ATMs and POS terminals, illustrates the changing financial behaviour of individuals as well as the increased accuracy in terms of its predictive potential.

The research seeks to address the following key questions:

1. What are the main drivers of cash demand, and how does it affect the underground economy in Romania?
2. We know that machine learning algorithms have been used to estimate the size of the shadow economy, but how do they compare with traditional econometric models?
3. How is the informal economy reconfigured due to digitalization?

In answering these questions, this study contributes to the literature on the underground economy and provides recommendations for policymakers. The results will contribute to the design of improved fiscal and regulatory measures for dealing with informality. In the next sections, we will provide the literature review, research methods, empirical results, and economic policy implications.

Literature review

Theoretical foundations of the underground economy

The underground economy has long posed a challenge for policymakers in economies transitioning with structural changes like Romania. Prior research has found excessive tax burden, compliance burden and lack of trust in institutions as strong determinants of informality (Schneider & Enste, 2000). Building on earlier work, Tanzi (1980) and Feige (1990) presented theoretical categories, or models, in which informal economic activity was generated by tax avoidance and monetary factors. While these classical perspectives form the basis on which the indirect estimates of the informal economy rest.

Socio-political conditions also shape informal economic activity, in addition to economic drivers. The underground economy can flourish in a country that is characterized by corruption, weak compliance with labour standards, and low trust in its institutions (Asllani, Dell'Anno &

Schneider, 2024). Moreover, the informal economy has displayed countercyclical tendencies, growing during times of economic crises and declining during periods of economic stability (Dell-Annoi & Davidescu, 2019). Findings like these encourage attempts to improve estimating techniques in the name of greater conceptual coherence.

Traditional estimation methods and their limitations

The size of the underground economy has been estimated using several methods, including the currency demand approach (CDA) and the multiple-indicator multiple-cause (MIMIC) model. CDA has been the primary method employed in Romania in which excessive cash demand is thought to correlate with undeclared or partially declared economic activity (Dell'Anno & Davidescu, 2019). The MIMIC model has also been used to provide estimates of informal sector dynamism using financial indicators and broad institutional indicators (Schneider et al., 2010).

While they are all relevant, each of these approaches has limitations. In an increasingly digitalized economy, it is doubtful that the cash-to-GDP ratio would remain stable, however, CDA does assume this. Similarly, MIMIC models depend on indicator choice, and estimates differ across studies. They are also anchored to prespecified functional forms, which may not accurately capture the nuance of informal economic behaviour.

Various alternative methods utilizing nonlinear modelling and machine learning have been developed to overcome these limitations. The growing availability of economic data and the emergence of artificial intelligence create opportunities to allow access to more flexible, data-driven models in order to capture the unique aspects of nonlinear relationships found in the underground economy.

Machine learning approaches for shadow economy estimation.

Recent developments in artificial intelligence allow the application of methods of machine learning to estimate the underground economy. Therefore, Support Vector Regression (SVR), Random Forest, and Gradient Boosting represent powerful methods for detecting nonlinear patterns in economic data (Felix, Alexandre & Lima, 2024). Machine learning models use nonlinear patterns, unlike econometric models, and this can improve prediction accuracy.

For instance, Shami and Lazebnik (2024) demonstrated that machine learning outperformed regression-based models in predicting currency demand (an important variable in estimating the underground economy). Their comparison of Random Forest models to traditional linear regression models showed that machine learning methods predict cash demand much more accurately, a key proxy for unregistered economic activity.

In the same way, Felix, Alexandre, and Lima (2024) investigated the informal economy and government spending relation of 11 models (4 traditional regression methods (Lasso, Ridge, and Elastic Net) and 7 machine learning algorithms (SVR, neural networks, Random Forest, LightGBM, CatBoost, Bagging and XGBoost)) using data of 122 countries (2004 - 2014). The authors discovered that machine learning models materially outperformed the linear models. All machine learning models had R^2 values significantly over 90% (except for SVR which was 69.6%) while the best of the traditional models was only 58%.

In another study, Ivaşcu and Ştefoni (2023) examined the link between the underground economy and important government expenditures (social protection, education, healthcare) using Random Forest, XGBoost, neural networks, and SVR using EU data (1995-2020). Their finds revealed that informality declines where social protection exceeds 20% of GDP, healthcare spending surpasses 6%, and education spending is rather between 6-8%. Again, the machine

learning methods outperformed the classical regression models, which showed their better predictive ability in economic contextualization.

Research gaps and contribution of this study

Even with the growing interest in the application of machine learning in other works, there are not many examples of the use of machine learning approaches to estimate Romania's underground economy. The studies already completed used econometric models, but there is a lack of application of any AI-driven methodologies. This study seeks to fill this gap by applying SVR (and other machine learning methods) to the estimation framework, to achieve a more refined and data-driven approach.

Data and Methodology

Dataset

The dataset included a number of relevant economic variables of potential interest in estimating the informal sector in Romania for the period 2000-2023, on a quarterly basis, using sources from the National Bank of Romania and from the World Bank, Eurostat and the International Monetary Fund (which can be viewed in Table 1 of Appendix A).

The variables accounted for in the analyses are: currency in circulation, monetary aggregates, real GDP, GDP per capita, inflation rates, unemployment rates, personal remittances received, deposit interest rates, real exchange rate, enforcement strength, bank cards per capita, ATMs per capita, POS terminals per capita, GDP deflator, and population in Romania.

The currency ratio in circulation to the monetary aggregate reflects the frequent use of cash in transactions, a characteristic of underground activities, as cash transactions leave fewer traces and are more difficult for tax authorities to track and quantify.

The extent of missing observations affected our data availability, especially for the digitalization variables, foreign remittances, deposit interest rates, and enforcement strength. Linear interpolation was used to fill the dataset, providing a fast and simple way of estimating the missing data between the two points.

The formula for imputing missing variables using linear interpolation is as follows:

$$y = y_0 + (x - x_0) * \frac{y_1 - y_0}{x_1 - x_0} \quad (1)$$

To enable comparability over time and in real terms, currency in circulation, the monetary aggregate, and remittances received from abroad were normalized by the GDP deflator. Tax revenues (the sum of direct taxes + indirect taxes + social contributions) were represented in terms of a percentage of real GDP. To normalize and give an easily comparable picture of accessibility and use of financial services, the digitalization variables were normalized to the population in Romania. The lowest value observed during the analysis period for the inflation rate was -0.76, so a constant of 1.76 was added. A detailed description of the methods used for transforming and processing the data is provided in Table 1 from Appendix A.

Lastly, we performed logarithmic transformations on variables such as currency in circulation, monetary aggregates, remittances from abroad, real GDP, GDP per capita, inflation rates, tax revenues, and digitalization indicators to stabilize variability within the data.

Furthermore, Géron (2019) emphasizes that time series should be converted into stationary time series before being introduced into nonlinear models to get the most out of an algorithm. The stationarity of the time series was checked using the ADF (Augmented Dickey-Fuller), KPSS, Phillips-Perron and Zivot-Andrews tests and their seasonality. The seasonality analysis was

conducted graphically through seasonal decomposition, ACF, and PACF, and it was also based on the results of the Hegy test.

According to the results on the stationarity of the time series included in the analysis (Table 2 from Appendix B), the ADF, KPSS, Zivot-Andrews, and Phillips-Perron tests indicate that the series is non-stationary. Prior to applying the first-order difference, we interpreted the seasonality diagram (Figure 1, Appendix C) and concluded that heavily seasonal variables include: tax revenues, real GDP, GDP per capita, and the unemployment rate. This was also confirmed through the Hegy test. In this manner, STL differentiation was applied to account for seasonal impacts and short-term volatility.

However, the deseasonalized series still exhibited non-stationarity, so the first-order difference was applied to all variables in the dataset. Table 3 Appendix D contains the results regarding the order of integration of the different series. The KPSS and Phillips-Perron tests confirm the stationarity of all variables after applying the first difference I(1).

Currency Demand Approach (CDA)

This study estimates the size of Romania's underground economy using a machine learning-based approach based on qualitative and quantitative techniques. Support Vector Regression was applied to calculate the coefficients needed for the CDA model, improving the conventional methodology by using machine learning to increase estimation performance.

The currency demand approach is an important tool in econometric analysis in order to study informal monetary dynamics as well as cash demand.

This model is based on the direct relationship between cash demand and hidden economic activities. This method is based on two assumptions, namely: that the underground economy relies almost exclusively on cash transactions and that cash demand in the official monetary system can be explained by macroeconomic variables such as gross domestic product (GDP), interest rates, taxation, the level of financial regulation, and inflation.

According to Tanzi (1980), the model's equation is as follows:

$$LOG\ \left(\frac{C}{M_2}\right) = \alpha + \beta_1 \log(Y) + \beta_2\ LOG\ (1 + \ T) + \beta_3 \log(R) + \beta_4 D + \varepsilon \quad (2)$$

C/M_2 represents the ratio of currency in circulation (C) to the monetary aggregate (M), Y is the level of economic activity (usually Gross Domestic Product), T represents the tax burden (tax revenues, income tax rates, etc.), R is the interest rate, and D is a set of dummy variables capturing regulations and other structural characteristics. The coefficients α and β are to be estimated, while ε is the error term.

While the model's basic equation has remained constant over time, various modifications and extensions have been introduced to adapt the method to specific economic conditions and available data. The diversity of opinions and adjustments illustrates the complexity of the underground economy and the importance of continuous adaptation to existing conditions and needs.

Dell'Anno and Davidescu (2019) utilize an adapted baseline equation of cash demand in their research, according to Tanzi (1983) and other subsequent contributions. They stated that the ratio of currency in circulation to the monetary aggregate (M1) is the sum of tax revenues, regulatory power, real GDP, inflation rate, interest rates on deposits, remittances, and nominal rate.

This paper attempts to estimate the measure of the underground economy in Romania based on the currency demand approach by applying state-of-the-art methods (Support Vector Regression). Therefore, starting from the original equation estimated by Tanzi and the adjusted

versions by Dell'Anno and Davidescu, the model presented in this paper will adapt the basic equation of cash demand by incorporating established economic determinants based on the literature, and variables related to digitalization.

The main steps for determining the size of the economy using cash demand involve estimating cash under conditions of an informal economy, estimating cash under conditions of no informality (which consists of limiting taxation), and identifying the size of the underground economy. The equation that underpins this paper is in the following form:

$$\ln\left(\frac{C}{M_1}_t\right) = \beta_0 \ln(1 + TAX_t) + \beta_1 \ln(PIBr_t) + \beta_2 \ln(REM_t) + \beta_3 \ln(1.76 + INF_t) + \beta_4 \ln(CARD_t) \quad (3)$$

Random Forest

Random Forest (Breiman, 2001) is an ensemble learning algorithm that improves prediction accuracy by aggregating multiple decision trees. Each tree is trained on random subsets of data and features, reducing overfitting. Node impurity is measured using Gini impurity or entropy:

$$Gini\ (t) = 1 - \sum_{i=0}^{C} p(i|t)^2 \quad (4)$$

$$Entropy\ (t) = -\sum_{i=1}^{C} p(i|t) \log p(i|t) \quad (5)$$

Final regression predictions are obtained by averaging three outputs:

$$\hat{Y} = \frac{1}{B}\sum_{b=1}^{B} \hat{Y}_b \quad (6)$$

Where B is the number of trees.

Gradient Boosting

Gradient Boosting (Friedman, 2001) builds models sequentially, correcting previous errors. It minimizes a loss function using gradient descent, updating predictions iteratively:

$$F_m(x) = F_{m-1}(x) + \eta f_m(x) \quad (7)$$

Where η is the learning rate. Regularization techniques prevent overfitting, enhancing predictive performance.

Support Vector Regression (SVR)

SVR extends Support Vector Machines for regression by fitting a hyperplane with a margin of tolerance ε (Smola & Schölkopf, 2004). It optimizes the loss function:

$$L(y, \hat{y}) = \max(0, |y - \hat{y}| - \varepsilon) \quad (8)$$

Kernel functions, such as linear, polynomial, or RBF, enable SVR to model nonlinear relationships effectively.

These models improve underground economy estimation by capturing complex patterns beyond traditional methods.

Results and discussions

In examining the extent of the informal economy in Romania, a variety of machine learning techniques were exploited, including Random Forest, Gradient Boosting and Support Vector Regression (SVR). Random Forest and Gradient Boosting analyses allowed the identification of the importance of the variables evaluated within the models. Real GDP and tax revenues were found to be major contributors in forecasting, whereas remittances had the smallest influence.

The comparative graph in Figure 1 illustrates that the Random Forest model generates predictions that closely match the observed values from 2000 to 2023.

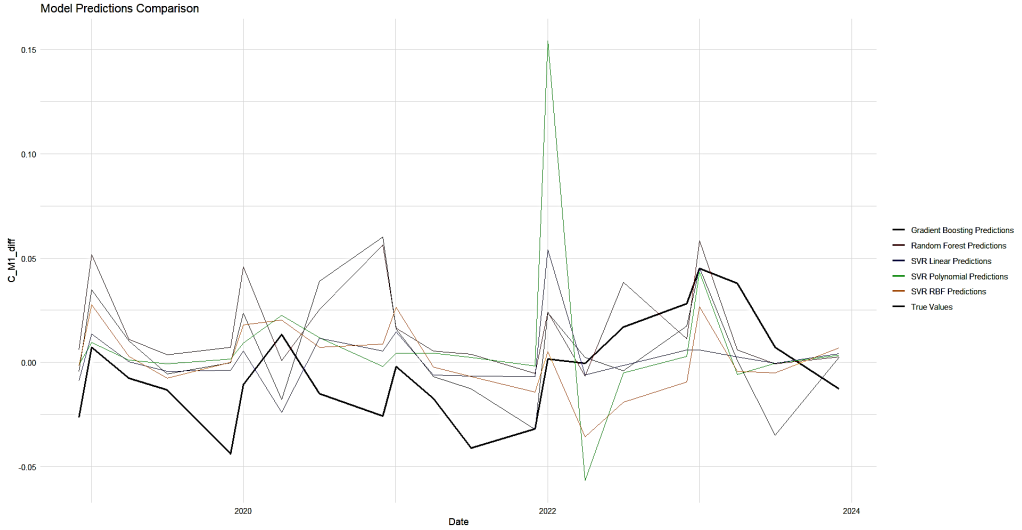

Figure 1. Currency predictions using machine learning techniques

Source: Authors' research results.

The model's performance was assessed with coefficient of determination and RMSE (Root Mean Square Error), and predictive performance was best for the Random Forest model. The least predictive performant of the three methods was the SVR method.

Although the Gradient Boosting and Random Forest models provide strong predictive performance, they both have some important drawbacks with respect to quantifying the size of the informal economy. One important downside of these models and machine learning models, in general, is the availability of coefficients to measure liquidity demand. Currently, these models are primarily suitable for prediction and classification but are less adept at measuring and interpreting the unregulated economy.

The SVR model is the most appropriate model to create a link between traditional econometrics and machine learning techniques. It enables the development of strong models where coefficients are estimated and interpreted, making SVR useful for integrating traditional and modern approaches. This convergence promotes accuracy and comprehension of the interrelationship between economic variables relevant to informality.

The SVR algorithm was applied to a refined dataset, which was segmented into training and testing subsets that included the relevant variables from the foundational equation. Following the extraction of model coefficients, the resulting equation for cash demand is presented as follows:

$$\Delta \ln\left(\frac{C}{M_1}_t\right) = 0{,}021\,\Delta^{STL}\ln(1+TAX_t) + 0{,}139\,\Delta^{STL}\ln(PIBr_t) + \Delta\,0{,}077\ln(REM_t) - 0{,}018\,\Delta\ln(1.76 + INF_t) - 0{,}175\Delta\ln(CARD_t) \quad (9)$$

In general, a positive coefficient indicates that higher tax revenues as a share of real GDP (0.021) means a more considerable demand for money accompanies increased tax rates. Economic theory indicates that increases in tax burdens will increase informal economic activity because individuals and firms will seek to escape taxes, with Schneider and Enste (2000) noting that when tax returns are higher, the degree of the informal sector increases because of the difficulties in taxing or monitoring cash.

The positive coefficient for real GDP means that increased economic activity is associated with increased demand for cash for day-to-day transactions. Cagan (1958) stated that as economic activity grows, the demand for cash increases proportionally due to the demand for cash for illegal transactions, as individuals and firms engage in a range of exchanges regardless of the legal implications.

The positive coefficient on the remittances means that more cash is demanded as remittances are received abroad. Remittances are money sent home by workers abroad to their families. Remittances have an important role in the broader economic context. Freund and Spatafora (2008) suggest that these funds crowd into the informal economy by increasing money demand as final recipients of the funds prefer not to use them for formal transactions or simply want to hold as much currency as possible in order to conduct informal (non-reported) transactions.

Conversely, the inflation rate and the number of cards used are the only variables displaying negative coefficients, per classical economic theories. The inflation rate coefficient tells us that an increase in inflation causes cash demand to fall. When inflation occurs, the real value of cash decreases, thus individuals and firms hold less cash and make efforts to maintain and grow wealth by instead investing in real estate or holding interest-earning assets. Friedman (1989), in examining inflation, suggested that in the context of increasing inflation, the public will seek to protect itself against an erosion of purchasing power by holding less cash.

Conversely, the negative coefficient referring to the digitalization variable, identified by the number of cards in circulation, clearly and logically illustrates that as the usage rate of payment cards goes up, the cash demand goes down. The shift towards digitalization and technological advancement involves a growing reliance on cards or other forms of electronic payments, thereby reducing the significance of cash in transactions.

Estimating cash in the context of informality involves determining the estimated level of cash by substituting values from the dataset into equation (10). Given the differentiated series, the first value will be that from the original dataset, as presented in equation (11), and the series for the variable will be reconstructed using the following formulas:

$$\Delta\, ln\left(\hat{C}\right)_{i,0} = ln(C)_{i,0} \tag{10}$$

$$\Delta\, ln(C\hat{}\,)_{i,t} = ln\left(\hat{C}\right)_{i,t-1} + \Delta\, ln\left(\hat{C}\right)_{i,t/t-1} \tag{11}$$

$ln(C)_{i,0}$ represents the missing lag's actual value from the dataset, and $\Delta\, ln\left(\hat{C}\right)_{i,t}$ represents the estimated values.

It is important to note that these values are expressed in logarithmic terms, and the exponential function will be applied to convert them into real terms.

After estimating the total cash demand using the actual values of economic variables, the level of legal cash (excluding informal forms) is similarly determined.

The scenario applied in this study has a minimum constraint. The minimum tax revenue percentage on the dataset, as indicated (22.5%), occurred in Q4 2008. Therefore, this variable will be limited to the minimum constant value in the equation obtained to determine the cash level in this context. This scenario assumes the absence of incentives for informal economic activities. Koloane and Bodhlyera (2022) argue that a minimum tax rate is the optimal approach to prevent tax evasion.

Nonetheless, Tanzi (1983) subsequently argued after evaluating both the minimum and zero-tax options, that a limited tax value of 0 best represents the extent of the informal issue. However, this scenario is unrealistic since taxes are the main source of revenue for governments

and would disrupt a country or region's ability to effectively provide programs and services needed or maintain a stable economy.

The difference between total cash demand and legal cash demand is called the extra currency (EC) level, which is, in effect, the level of cash for the underground economy. The difference between money aggregate (M1) and extra currency can be seen as that portion of the M1 in circulation for official and legitimate economic activities, and the difference is then applied in the velocity equation as (Tanzi, 1983):

$$v = \frac{PIB}{M_1 - EC} \quad (12)$$

Velocity is the speed of money exchanges and refers to how many times a unit of currency is used to purchase goods and/or services in a specific period. It measures how quickly money flows through the economy. Finally, the size of the underground economy is estimated by multiplying the velocity of money by its cash volume demand for the informal economy.

After applying the calculations outlined above, the figure below shows the evolution over time of the size of the unregulated economy in Romania.

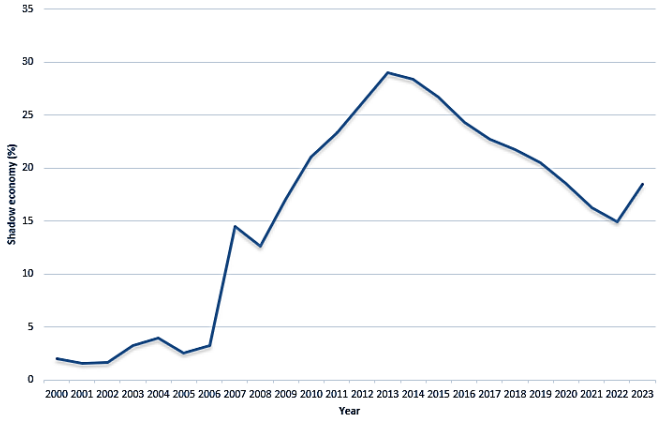

Figure 2. The progression of the shadow economies in Romania from 2000 to 2023

Source: Authors' research results.

The period from 2000 to 2024 reflects moderate growth which can be explained by a post-communist economic transition. During these years institutional and financial reform was developing. During this time the informal economy was influenced by instability in legislation and administration related to tax collection.

In 2007, Romania's accession to the European Union brought significant changes to the institutional and public sectors to align with European standards, resulting in compliance with the strict rules of the European Union.

A further yet major event that effectively destabilized global markets was the 2008 economic crisis. The recession resulted in an increase in the unemployment rate due to massive job losses and business closures. The outcome of the economic decline was demonstrated through the immediate growth of the informal economy as the population struggled to survive the economic shock through unofficial means.

The underground economy peaked in 2013 when the extent of this phenomenon reached nearly 30% of the Gross Domestic Product. Austerity measures and the government's failure to impact tax evasion made substantial contributions to this outcome. In 2013, reports from the National Bank of Romania stated inflation was at 1.6%, its lowest level since the 1989 revolution.

Based on economic theory in this context, the tremendous increase is justified by the inverse relationship between inflation and cash demand.

Moraru and Popovici (2014) argue that special construction taxes targeting non-residential infrastructures (industrial buildings and warehouses) were introduced in this reference year. This pressure on businesses prompted a significant migration toward unofficial activities to avoid taxation.

The COVID-19 pandemic, which began in early 2020, had devastating effects on populations and official economies worldwide. The economic activity from the first quarter of 2020, when the alert was declared, was limited by restrictions intended to limit the virus's speed of transmission and progression, which led to chaos due to severe limitations on commercial practices. These measures led to a sharp decline in consumption, increased unemployment due to business closures resulting from lockdowns, and financial difficulties for the population.

According to the 2021 report on the economic impact of COVID-19 provided by the International Monetary Fund, the Romanian government intervened with a series of support measures, including subsidies and technical unemployment schemes, deferral of tax payments, and state-guaranteed loans.

The year 2022 saw a significant decrease in the underground economy. This development was influenced by a series of economic and social events that impacted the national economy: growth in GDP, high inflation, and the war between Ukraine and Russia.

The inflation rate played a crucial role in this decline, averaging 13.7% annually. This was driven by the immense increase in energy and food prices caused by the unfavourable context caused by the war between Russia and Ukraine. The Russian invasion of Ukraine caused a significant crisis in the energy sector at a European level with high prices translating into a sizeable rise in cost of living pressure and a lot of uncertainty as to how or when economies would recover. The increases in these economic factors directly impacted the reduction of the informal sector.

Conclusion

The objective of this research was to analyze the dimension of the Romanian underground economy by using the model based on liquidity demand. The coefficients of the necessary variables inputted in the regression were calculated using machine learning methods like SVR.

The SVR model accurately and confidently estimated the unregulated economic phenomenon. The study's results concluded that rising fiscal revenues lead to increased demand for liquidity, thus forcing economic agents and persons to enter the unregulated sector to avoid being taxed.

There was also a strong positive correlation between GDP and cash, indicating that a lot of economic activities happen off the books. Apart from recent government policies, a growth in demand for cash was not the only cause, another key element was the influx of remittances from abroad that were fueling the shadow economy.

On the other hand, the correlation of inflation with the dependent variable was negative, which was expected based on economic knowledge since inflation devalues liquid assets. The variable of digitalization served as a novel element, where increased card usage leads to reduced cash demand, reflecting a broader formalization of the economy and a decrease in the underground economic sector.

The graphical representation of the evolution of the underground economy in Romania illustrated significant variations influenced by a series of economic, fiscal, and legal events during

the studied period. Economic growth periods were often accompanied by increases in the underground economy, suggesting that a portion of economic activity remains unregulated.

Moreover, the informal sector fluctuated through periods of economic crisis or extreme fiscal regimes, demonstrating the impact of these events on off-the-books economic activity. For example, the underground economy grew even more during 2008 to 2013, after developing a strong upward trend that began in 2008, due in large part to the global financial crisis, as well as government initiatives.

A significant drop occurred in 2022 when the war between Russia and Ukraine began, which resulted in significant price rises in energy and food. Prices began to increase significantly, and the (GDP) inflation rate rose rapidly, reducing the total size of the underground economy.

Limitations of this research were the availability of data, restrictions posed by machine learning methods and a limited amount of literature from specialists. The quality and availability of economic data pose a significant challenge and in some cases, imputation was performed on several variables which had missing values, with linear interpolation being used. This made it possible to complete the dataset and reduce the amount of bias identified in other studies, however, it may have limited some of the accuracy of the analysis, albeit minor, while also changing some of the fundamental properties of the data.

Although the SVR model effectively extracted the coefficients needed for the liquidity demand method, it proved to be the only suitable method for determining the size of the underground economy at the national level in this context. Although Random Forest and Gradient Boosting demonstrated better predictive capabilities and increased accuracy, they were inadequate for this study, remaining useful only for prediction and classification problems. This limitation restricted the potential use of more advanced and precise techniques.

Another significant limitation emerged during the literature review process. With state-of-the-art machine learning techniques combined with more traditional econometric methods, research on the underground economy is still very much in its infancy and underexploited. This represented an obstacle as there is no comprehensive and well-validated body of literature to offer that could make available existing theoretical and empirical models that would have better supported and commented on the findings.

In light of these factors, future versions of this study should include longer time frames and more sophisticated missing value imputation techniques to mitigate their impact on understanding and modelling the underground economy. On the other hand, training on more determinants in the regression equation would yield a better understanding of their significance and behaviour. It might also be useful to invest in a careful assessment of how these novel techniques can be combined with classical methods (for example, using alternative estimation methods like MIMIC (Multiple Indicators Multiple Causes)).

The results of this research support the argument that policymakers should pursue not only fiscal policies but also digitalization policies to alleviate the size of the underground economy. The negative relationship between digitalization (as measured via card transactions) and cash demand suggests that policymakers should increase the implementation of electronic payment methods to deter informal transactions. It has been shown that a 10% increase in electronic payments for four consecutive years could reduce the size of the shadow economy by as much as 5% in size (Schneider, 2013). The shift towards digital transactions promotes transparency, decreases anonymity around cash exchanges, and ultimately creates better tax compliance and minimizes informal economic activities.

To maximize the potential of digitization, the policy actions above, that promote financial inclusion, will contribute to the increase of the uptake of digital payment solutions. Increasing acceptance of cards and digital payments is an essential first step to providing all businesses - particularly small and medium enterprises - with access to effective and low-cost digital payment infrastructure. Other tools to encourage a preference for digital payments over cash are also relevant, including tax rebates for electronic transactions, or lowering transaction fees. Overall, these stimulate alternative cashless transaction growth. Limits to cash transactions, in the form of regulating cash and reducing tax evasion and underreporting what they earn, as is being undertaken in some European countries, have also been contemplated.

Additionally, enhancing tax monitoring through artificial intelligence and big data analytics can improve tax enforcement by detecting irregularities in transaction records. Strengthening institutional trust by improving transparency and efficiency in fiscal policies can also foster voluntary tax compliance and reduce reliance on informal transactions. By integrating these digitalization strategies with traditional economic instruments, Romania can move toward a more transparent and structured economy while effectively curbing the negative impact of the underground sector.

Acknowledgement

This paper was co-financed by the Bucharest University of Economic Studies during the PhD program. The research study has been elaborated within the Data Science Research Lab for Business and Economics of the Bucharest University of Economic Studies. This research was supported by Project 101182756 — INSEAI 2023: International Network for Knowledge and Comparative Socioeconomic Analysis of Informality and the Policies to be Implemented for their Formalization in the European Union and Latin America, funded by the Marie Sklodowska-Curie Actions (MSCA) Staff Exchanges program.

References

Asllani, A., Dell'Anno, R., & Schneider, F. (2024). Mapping the informal economy worldwide with an enhanced MIMIC approach: New estimates for 110 countries from 1997-2022 (No. 11416). CESifo Working Paper.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Cagan, P. (1958). The demand for currency is relative to the total money supply. *Journal of Political Economy*, 66(4), pp. 303-328.

Dell'Anno, R., Davidescu, A. (2019). Estimating shadow economy and tax evasion in Romania. A comparison by different estimation approaches. *Economic Analysis and Policy*, 63.

Feige, E. L. (1990). Defining and estimating underground and informal economies: The new institutional economics approach. *World Development*, 18(7), pp. 989-1002.

Felix, J., Alexandre, M., & Lima, G. T. (2024). Applying machine learning algorithms to predict the size of the informal economy. Computational Economics, 1-21.

Freund, C., Spatafora, N. (2008). Remittances, transaction costs, and informality. *Journal of Development Economics*, 86(2), pp. 356-366.

Friedman, M. (1989). 'Quantity theory of money'. In: Money. London: Palgrave Macmillan UK, pp. 1-40.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

Géron, A. (2019). Hands-on machine learning with scikit-learn, keras, and tensorflow: concepts. *Google Kitaplar*.
Ivașcu, C., Ștefoni, S. E. (2023). Modelling the Nonlinear Dependencies between Government Expenditures and Shadow Economy Using Data-Driven Approaches. *Scientific Annals of Economics and Business*, 70(1), pp. 97–114.
Koloane, C.T., Bodhlyera, O. (2022). A statistical approach to modelling the underground economy in South Africa. *Journal of Economics and Management*, 44, pp. 64-95.
Moraru, C., Popovici, N. (2014). Analysis of Fiscal Policy Measures during 2005 - 2013. The Case of Romania. Ovidius University Annals, *Economic Sciences Series*, 0(2), pp. 224-228.
Schneider, F., Enste, D. H. (2000). Shadow economies: Size, causes, and consequences. *Journal of Economic Literature*, 38(1), pp. 77–114.
Schneider, F., Buehn, A., & Montenegro, C. E. (2010). New estimates for the shadow economies all over the world. International Economic Journal, 24(4), 443-461.
Schneider, F. (2013). The shadow economy in Europe, 2013. A.T. Kearney & Visa Europe.
Shami, L., & Lazebnik, T. (2024). Implementing machine learning methods in estimating the size of the non-observed economy. Computational Economics, 63(4), 1459-1476.
Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing, 14(3), 199-222.
Tanzi, V. (1980). The Underground Economy in the United States: Estimates and Implications. *Banca Nazionale del Lavoro*, 135(4), pp. 427-453.
Tanzi, V. (1983). The Underground Economy in the United States: Annual Estimates, 1930-1980. *IMF-Staff Papers*, 30(2), pp. 283-305.

Appendix A. Data sources

Table 1. Data sources and their descriptions

Alias	Variable name	Description	Unit of measure	Data source
C	Cash in circulation	Cash in circulation, normalized using the GDP deflator (2015 = 100)	RON	National Bank of Romania
M1	Monetary aggregate	Monetary aggregate normalized using the GDP deflator (2015=100). For the period 2000-2006, the series was recalculated because, starting from 2007, there were changes regarding the method of calculating the money supply M1.	RON	National Bank of Romania
PIBc	GDP per capita	Real Gross Domestic Product per capita	RON	Eurostat
PIBr	GDP real	Real Gross Domestic Product	RON	Eurostat
TAX	Tax revenues (% real GDP)	Tax revenues calculated as the sum of direct taxes, indirect taxes and social contributions, expressed as a percentage of real Gross Domestic Product.	%	Eurostat
ENF	Enforcement strength	The enforcement strength is determined as an average index between the 3 institutional indicators: Rule of Law, Regulatory Quality and Government Effectiveness. Scores range from -2.5 to 2.5, where high scores signify high application power.		Worldwide Governance Indicators database, World Bank
R	Interest rate	Nominal interest rate on deposits	%	International Monetary Fund
INF	Inflation rate	The quarterly rate of inflation is determined as an arithmetic mean of the harmonized monthly indices of consumer prices, from which the comparison base equal to 100 is subtracted.	%	Eurostat
REM	Personal remittances received	Personal remittances received, calculated as the sum of personal transfers and employee compensation, normalized using the GDP deflator (2015=100)	RON	International Monetary Fund - balance of payments database
REER	Real exchange rate	The real effective exchange rate	%	Eurostat
SOM	Unemployment rate	Quarterly unemployment rate	%	Eurostat
CARD	Number of cards in circulation per capita	The total number of cards in circulation, divided by the population of Romania.	unit	National Bank of Romania
ATM	Number of ATMs per capita	The total number of existing ATMs, divided by the population of Romania.	unit	National Bank of Romania
POS	Number of POS per capita	The total number of existing POS, divided by the population of Romania.	unit	National Bank of Romania
Deflator GDP	GDP Deflator	GDP deflator (2015=100)	RON	Eurostat
POP	The population of Romania	Quarterly total population of Romania	mil. People	Eurostat

Source: Authors' own research.

Appendix B. Results of stationarity tests

Table 2. Results of stationarity tests of time series in level

		Level			
Variable	**Method**	**ADF_P**	**Zivot_P**	**KPSS**	**PP**
ln(C)	None	0,8029	-3,5910	1,1194	-2,4343
	C	-2,2068	-3,0024	0,2006	-1,2688
	T&C	-1,7374	-3,5700		
ln(M1)	None	1,2728	-3,4424	0,9693	-0,9018
	C	-0,7161	-2,6252	0,1953 ***	-1,7621
	T&C	-1,7398	-3,5224		
ln(PIBc)	None	2,0968 **	-3,3922	1,1252	-2,3680
	C	-1,5338	-3,2087	0,1714 ***	-3,4465
	T&C	-2,4565	-4,0965		
ln(PIBr)	None	2,0542 **	-3,4306	1,1164	-2,1758
	C	-1,3959	-3,1517	0,1505	-3,7937
	T&C	-2,5331	-4,4869		
ln(1+TAX)	None	-0,6183	-4,8483 **	0,7493	-7,1526 ***
	C	-2,1649	-3,4904	0,1512 ***	-8,9715 ***
	T&C	-2,5870	-4,8441 *		
ln(1.76+INF)	None	1,9466 *	-2,6953	1,1126	-2,7574
	C	1,2050	-2,4961	0,2340	-3,4501
	T&C	-1,2471	-2,3426		
ENF	None	-0,3417	-3,1559	0,9157	-1,5839
	C	-1,6766	-3,9408	0,2548	-1,4073
	T&C	-1,2092	-4,6991		
ln(REM)	None	-0,7953	-4,4060	0,4997 **	-2,0776
	C	-2,0952	-3,3314	0,2209	-2,1595
	T&C	-1,9699	-4,4099		
REER	None	0,6656	-4,8581 **	0,5241	-1,5504
	C	-1,8523	-4,7178 **	0,1817 ***	-1,6446
	T&C	-2,0757	-4,8416 *		
ln(CARD)	None	-2,3040 **	-3,0436	1,1471	-2,3702
	C	-1,6506	-3,9161	0,2019 ***	-1,9480
	T&C	-1,9862	-4,1395		
ln(ATM)	None	-1,4488	-3,3661	0,8876	-2,2895
	C	-1,1108	-2,9187	0,2375	-1,4181
	T&C	-2,0771	-3,4701		
ln(POS)	None	-2,5330 **	-34,3666 ***	1,0389	-1,2550
	C	-3,5382 ***	-7,9251 ***	0,1381***	-2,7104
	T&C	-2,8110	-34,2027 ***		
SOM	None	-2,5330	-3,5506	0,7820	-2,3965
	C	-3,5382	-2,6241	0.782	-3,4392
	T&C	-2,8110	-3,7944		
R	None	-3.3656***	-3.9744	0.8556	-3.232 **
	C	-3.9517***	-3.8842	0.2496	-1.0944
	T&C	-2.7528	-3.9159		

Note: *, **, *** indicate significance level of 10%, 5%, 1% (for which series are stationary). None is the model without constant and trend, C represents model with constant and T&C is the model with trend and constant.

Source: Authors' own research.

Appendix C. Seasonality of data

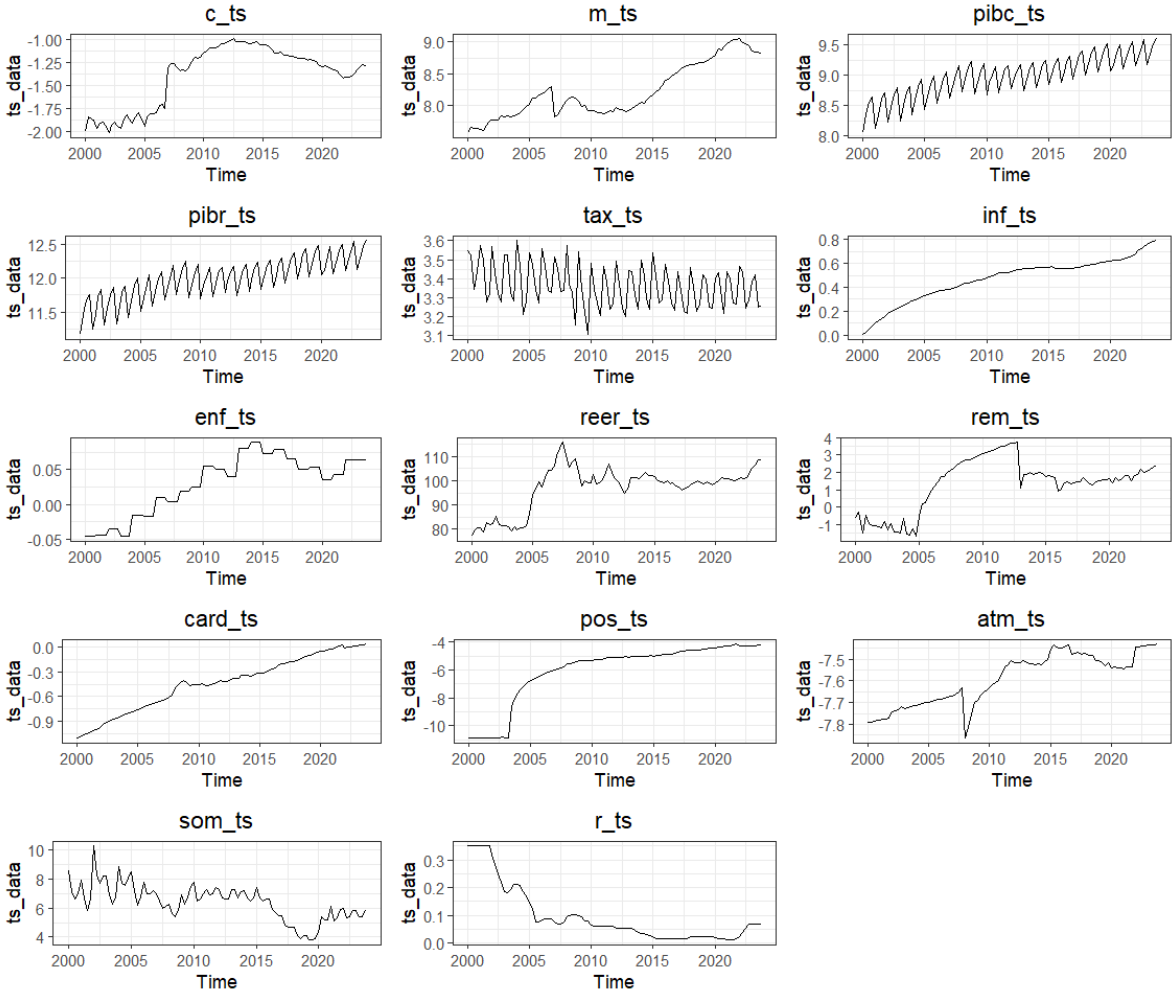

Figure 1. The evolution over time of the analyzed economic variables

Source: Authors' own research results.

Appendix D. Results of stationarity tests of first order

Table 3. Results of stationarity tests of first order and seasonally differentiated time series by STL

		The first difference			
Variable	**Method**	**ADF**	**Zivot_P**	**KPSS**	**PP**
ln(C)	None	-1,5183	-3,3967	0,4198 **	-11,1672 ***
	C	-2,1502	-3,1492	0,0598 ***	-11,3296 ***
	T&C	-2,8750	-3,6405		
ln(M1)	None	-2,3116 **	-3,6511	0,0943 ***	-8,8076 ***
	C	-2,5882 *	-3,1182	0,0993 ***	-8,7569 ***
	T&C	-2,5246	-3,4924		
ln(PIBc)	None	-2,1433 **	-5,5807 ***	0,3312 ***	-12,8058 ***
	C	-3,2924 **	-3,7791	0,0928 ***	-13,2255 ***
	T&C	-3,4355 *	-5,6401 ***		
ln(PIBr)	None	-2,2109 **	-5,3451 ***	0,2744 ***	-13,0707 ***
	C	-3,2720 **	-3,6869	0,0915 ***	-13,3068 ***
	T&C	-3,3567 *	-5,4028 ***		
ln(1+TAX)	None	-3,9511 ***	-4,3635	0,0668 ***	-30,2939 ***
	C	-3,9590 ***	-3,9735	0,0476 ***	-30,9103 ***
	T&C	-3,9342 *	-4,6059		
ln(1.76+INF)	None	1,9466 *	-3,1309	0,4300 **	-4,4783 ***
	C	-2,2316	-3,5207	0,2253	-4,8021 ***
	T&C	-1,0626	-3,7656		
ENF	None	-2,8265 ***	-4,6456 *	0,2185 ***	-9,9144 ***
	C	-3,1693 **	-3,8367	0,0794 ***	-10,0556 ***
	T&C	-3,4275 *	-4,7666 *		
ln(REM)	None	-2,4331 **	-3,6531	0,1242 ***	-7,8726 ***
	C	-2,5514	-3,0241	0,0812 ***	-7,8314 ***
	T&C	-2,6363	-4,0216 ***		
REER	None	-2,7285 ***	-4,7867 **	0,1355 ***	-12,7214 ***
	C	-2,8891 *	-3,2733	0,0990 ***	-12,6912 ***
	T&C	-2,9280	-6,3578 ***		
ln(CARD)	None	-1,7551 *	-4,8460 **	0,3431 ***	-6,8077 ***
	C	-3,1063 **	-3,6934	0,0625 ***	-6,9435 ***
	T&C	-3,5096 **	-4,9771 *		
ln(POS)	None	-2,9600 ***	-3,7965	0,3274 ***	-7,3118 ***
	C	-3,2888 **	-3,4627	0,0637 ***	-7,3942 ***
	T&C	-3,2728 **	-3,9944		
ln(ATM)	None	-2,1919 **	-4,6661 *	0,0685 ***	-10,7201 ***
	C	-2,6395 *	-4,3560 *	0,0558 ***	-10,6701 ***
	T&C	-3,5672 **	-8,2320 ***		
SOM	None	-2,8281 ***	-3,8583	0,0781 ***	-11,4863 ***
	C	-2,9085 **	-3,1353	0,0634 ***	-11,4015 ***
	T&C	-2,9174	-4,1656		
R	None	-2.1827 **	-4.8619 **	0.5729 ***	-4.3057 ***
	C	-2.0226	-4.372 *	0.0572 ***	- 4.1688 ***
	T&C	-3.1577 *	-4.9112 *		

Note: *, **, *** indicate significance level of 10%, 5%, 1% (for which series are stationary). None is the model without constant and trend, C represents model with constant and T&C is the model with trend and constant.

Source: Authors' own research.