Empirical analysis of non-compliance

July 2023

The Project is funded by the European Union via the Technical Support Instrument, managed by the European Commission Directorate General for Structural Reform Support
Specific Contract REFORM/SC2021/045

Building a better working world

This document was produced with the financial assistance of the European Union. Its content is the sole responsibility of the author(s). The views expressed herein can in no way be taken to reflect the official opinion of the European Union.

Acknowledgments

This report was written by a team of experts from the EY Economic Analysis Team. General supervision of the empirical analysis was done by Marek Rozkrut. The work was coordinated and conducted by Michał Kowalczuk, Magdalena Karska and Piotr Dybka. Persons that performed key analyses included also Stanisław Bartha, Anna Komisarska and Maciej Łopusiński. We are also grateful for valuable comments obtained from Andrzej Torój.

We acknowledge in-depth discussions and many relevant suggestions of the National Revenue Agency in Bulgaria as well as great support obtained from EY Bulgaria.

Contents

List of acronyms

Abbreviation	*Description*
CDA	Currency demand approach
CRM	Compliance risk management
DG REFORM	Directorate-General for Structural Reform Support
EY	Ernst & Young
EC	European Commission
EU	European Union
IMF	International Monetary Fund
MIMIC	Multiple Indicators Multiple Causes model
NRA	National Revenue Agency
OECD	Organisation for Economic Co-operation and Development

Executive summary

Tax gap is the difference between taxes that theoretically should be collected (based on the scale of economic activity and binding regulations) and the actually collected taxes. Sources of the tax gap include the shadow economy, tax frauds, tax evasion and other (e.g. legal disputes and bankruptcies). Shadow economy (non-observed economy) comprises various kinds of unreported value added (≈ GDP) of registered and unregistered businesses and is responsible often for a significant part of the total tax gap.

Our study covers three different areas of tax non-compliance in Bulgaria, analysed with different methodologies. The first focuses on the shadow economy and related part of the tax gap. The second examines the gap specifically related to personal income tax (PIT), while the third investigates the value-added tax (VAT) gap.

Shadow economy and related part of the tax gap

In this area we mainly used the currency demand approach (CDA), which is an econometric analysis of the cash in circulation. It allowed us to estimate the cash shadow economy (shadow economy generated by cash payments) in Bulgaria, related lost government revenues as well as identify they determinants. Our dataset covered about 100 countries (including Bulgaria) observed over the years 1996-2020 (panel data). All data points for this analysis were obtained from publicly available sources.

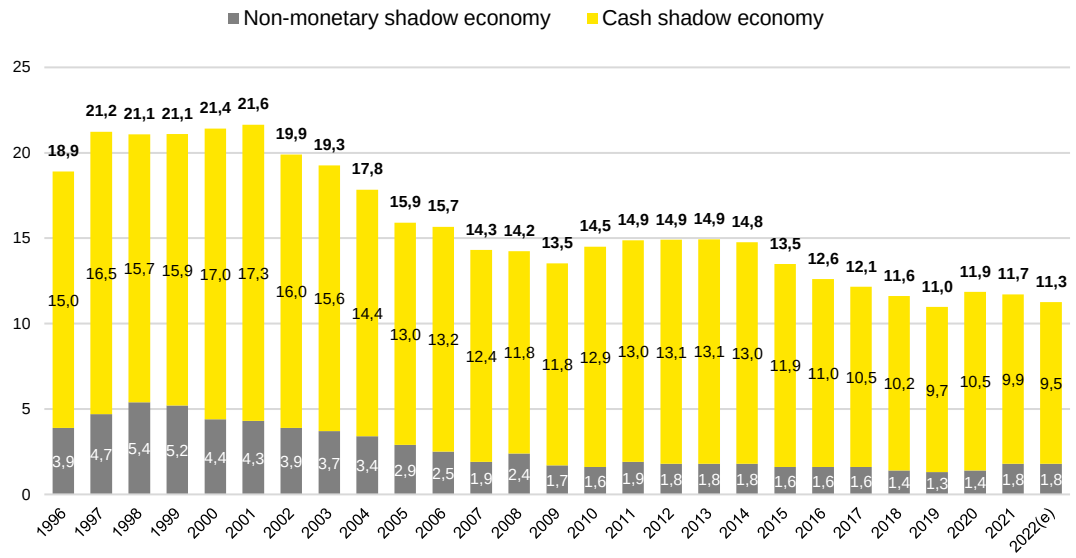
According to our estimates, the shadow economy in Bulgaria in 2022 amounted to 11.3% of GDP, out of which the cash shadow economy was equal to 9.5% of GDP and the non-monetary shadow economy (household production of goods for own use) was equal to 1.8% of GDP. Since 2001 there has been a downward trend in the shadow economy with some cyclical fluctuations (e.g. after the 2009's recession and during the pandemic in 2020) (see Chart ES.1).

It is worth noting that the likely higher share of unregistered employment in the total employment (vs the share of the shadow economy in GDP) in Bulgaria does not imply similar share of the shadow economy in GDP. The reasons include relatively low value added generated by unregistered employees and the fact that some of this value may be finally registered, e.g. a new building. One should also note that most of GDP generated by large companies and public entities/companies is likely reported (potential presence of other sources of tax gap or corruption is a different topic), leaving only the remaining part of GDP subject to unreported activity.

In 2022, the key contributors to the shadow economy size included (relatively low) government effectiveness (3.9% of GDP) and integrity of the legal system (2.2% of GDP) as well as taxation level (2.3% of GDP). Before 2017, unemployment rate was another important factor.

We estimated that in Bulgaria in 2022 lost government revenues due to the cash shadow economy totaled 1.88% of GDP, including lost VAT (1.21% of GDP) and income taxes (0.68% of GDP).

Chart ES.1 – Total, cash and non-monetary shadow economy in Bulgaria (% of GDP)



Notes: (e) – initial estimate based on incomplete data and additional assumptions.
Source: EY.

Unregistered income and the PIT gap

We applied the traces-of-true-income (Pissarides-Weber) approach, a micro-econometric model, to estimate the level of income underreporting in Bulgaria. This method indirectly measures tax non-compliance by examining disparities in expenditure and reported income patterns through econometric modelling. Our dataset, prepared in an anonymized form by the National Statistical Institute, included individual-level Household Budget Survey data extended by the information on income from National Revenue Agency data on annual tax returns. The dataset was produced and shared especially for this project and included observations for the years 2017, 2018, 2019 and 2021.

Our analysis uncovered significant income underreporting among both self-employed individuals and private sector employees in Bulgaria (public sector workers were assumed to be fully compliant in this method). In our sample, on average for the 2017-2021 period, the model revealed an income gap of 26.0% (of reported and unreported net labor income) for private sector employee households and 50.7% for self-employed households. Yet, these shares should not be interpreted for the total income of such groups in the economy due to the significant underrepresentation of more affluent households in the analysed sample.

Using certain assumptions, we translated the obtained estimates into different macroeconomic figures which are presented in table ES.1. We estimated that identified unreported labor income accounted for approximately 6.37% of Bulgaria's GDP. Out of this total, 5.36 percentage points were attributed to the private sector, with the remainder attributed to self-employed individuals. The estimated PIT and social security contributions gaps resulting from this income underreporting were equal to 13.8% and 16.5%, respectively.

Table ES.1 – Macroeconomic results of the unregistered income analysis, 2017-2021 averages

Macro-level estimates	Average value
Unreported labour income as % of GDP	6.37%
Unreported labour income of private sector employees as % of GDP	5.36%
Unreported labour income of self-employed as % of GDP	1.01%
Lost PIT revenues as % of GDP	0.54%
PIT gap as % potential PIT revenues	13.8%
Lost revenues from social security contributions as % of GDP	1.71%
Social security contributions gap as % of potential social security contributions	16.5%

Note: Potential PIT revenues and potential social security contributions are hypothetical values in the situation of full registration of income (perfect compliance).
Source: EY, Eurostat and NRA (for social security contributions revenues), NSI (for GDP), Ministry of Finance (for PIT revenues)

We also examined income underreporting for more disaggregated socio-demographic groups. Income gap within self-employed and/or private sectors households seemed to differ significantly along the following dimensions: with/without children, settlement size, age group, unemployed in the household and industry of employment.

VAT gap

We performed estimation of the VAT gap at the sectoral level in Bulgaria. We used an econometric model with (output) VAT gap estimate as explained variable (based on the difference between potential (output) VAT estimate and declared (output) VAT). We also conducted a partial analysis with an alternative VAT gap measure, based on the VAT audits data, but it appeared to be significantly more biased and less accurate (due to non-random assignment of audits, small number of audits in some sectors, etc.). Our dataset included 84 sectors in Bulgaria observed over the years 2014-2021 (panel data). It combined various data obtained from the National Revenue Agency (including sectoral VAT revenues) and publicly available sources.

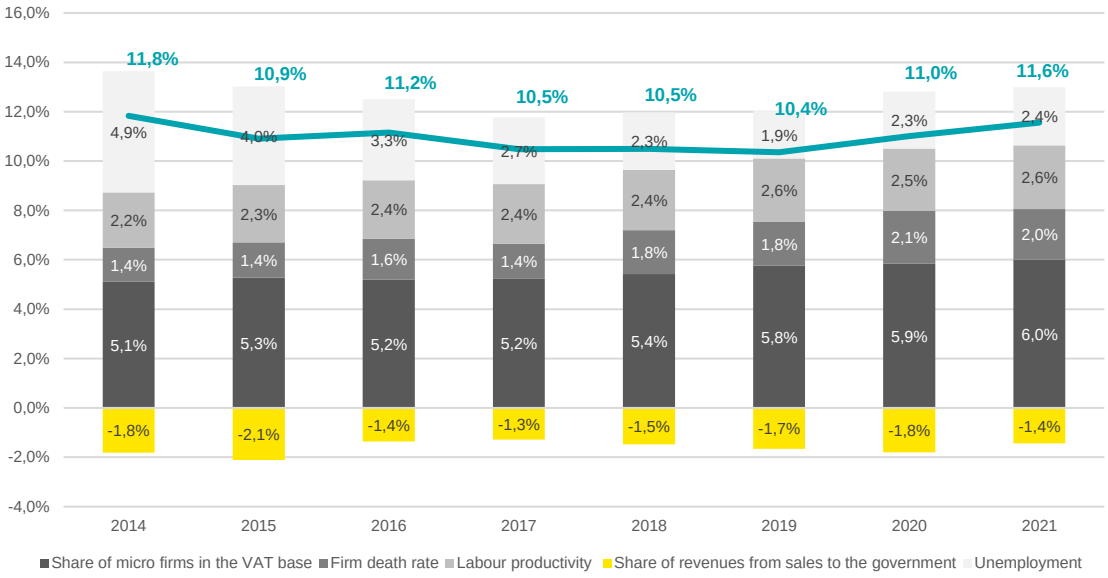
We found that the (output) VAT gap (% of potential VAT in the sector) was greater in the sectors with greater role of micro enterprises, more bankruptcies, and when unemployment rate was higher. Less intuitively, it was also larger in industries with higher labour productivity (maybe due to VAT frauds or evasion). In addition, the greater was the role of business-to-government transactions, the smaller was the sectoral VAT gap.

Given that even our (output) VAT gap dependent variable was subject to certain inaccuracies, our initial VAT gap estimates from the model should be rather used for comparisons between sectors and over time (not for determining the absolute VAT gap scale in the country). Due to this we calibrated (scaled) our results to have the same 2016-2019 average country-level VAT gap as in the previous VAT gap study of the European Commission. This allowed us to generate the final set of our VAT gap estimates for different sectors and years.

According to our results, the VAT gap in Bulgaria was in the downward trend between 2014 and 2019 but increased during the 2020-2021 pandemic years. Key contributors

to the VAT gap included the share of micro firms (in the VAT base) and unemployment rate (see Chart ES.2).

Chart ES.2 – Contributions of variables to the VAT gap estimate scaled to the European Commission's 2016-2019 average VAT gap estimate (% of potential VAT)



Notes: Blue line = net effect of positive and negative contributions. Potential VAT is a hypothetical value of declared VAT in the situation of full compliance.
Source: EY.

Sectors with the largest VAT gap (as % of potential VAT in the sector) included various professional services, other service activities and trade, while the lowest VAT gaps were found among different manufacturing sectors.

1. Introduction

The report was produced by EY within the framework of the Project REFORM/SC2021/045 “Strengthening the Compliance Management by Assessing External Context and Taxpayers Behaviour in Bulgaria”. The project was funded by the European Commission (EC) through DG REFORM. The National Revenue Agency (NRA) in Bulgaria was the main beneficiary of this work. The purpose of the report was to conduct an empirical analysis of the external context and its influence on non-compliance.

The report is divided into a few chapters.

In the second chapter, we describe in more details the definition of the tax gap, shadow economy and some other related concepts which are relevant for our analysis.

In the third, fourth and fifth chapter we empirically analyse (a) the shadow economy and related part of the tax gap, (b) unregistered income and PIT gap and (c) (sectoral) VAT gap, respectively. For each area of the research separately, we discuss our dataset and identified key variables. Further, we present the estimated scale of tax non-compliance in the country and contribution of different factors.

This is the publicly available version of the report. The full report (available to the NRA and EC) included also examples showing how the developed tools could be used to analyse future scenarios of tax non-compliance, suggestions for future development of the tools as well as some additional details. The full report covered in the main text many methodological aspects that were requested by the NRA. To simplify the reading of this document most information related to the data preparation and methodology was moved to the technical appendix.

Apart from the report, the project deliverables included three spreadsheet tools (one for each of the analysed areas) that show calculations for our key results and allow the user to analyse future scenarios of tax non-compliance in Bulgaria (not publicly available).

To our best knowledge, the conducted study includes many innovative elements that were not earlier covered by other researchers and the economic literature.

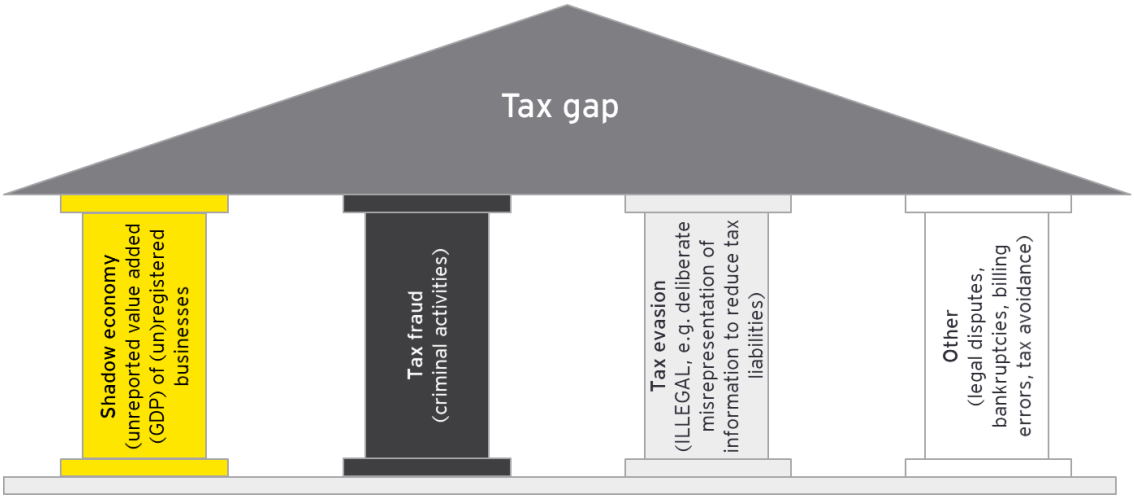
2. Definitions of relevant concepts

Before we start an empirical analysis of non-compliance it is worth introducing a few relevant concepts for such investigation, including the tax gap and the shadow economy.

2.1 Tax gap

Tax gap is the difference between taxes that theoretically should be collected (based on the scale of economic activity and binding regulations) and the actually collected taxes. Sources of the tax gap include the shadow economy, tax frauds, tax evasion and other (see Figure 1).

Figure 1 -- Tax gap and its sources



Source: EY.

Short definitions of these sources are the following:

- **Shadow economy (non-observed economy)** comprises various kinds of unreported economic activity of registered and unregistered businesses (see details further) and is responsible only for a part (often significant, though) of the total tax gap
- **Tax fraud** is a form of deliberate evasion of tax that is generally punishable under criminal law. The term includes situations in which deliberately false statements are submitted or fake documents are produced[1]
- **Tax evasion** generally involves illegal arrangements where tax liability is hidden or ignored, i.e. the taxpayer pays less tax than he/she is supposed to pay under the law (e.g. by deliberately misrepresenting information)[2]
- **Other**, often where the non-compliance is not deliberate, including legal disputes, bankruptcies, billing errors, etc.

The discussion above focuses on the **compliance tax gap**. In some research the authors also distinguish the additional component described as the **policy tax gap**,

[1] See, e.g. https://taxation-customs.ec.europa.eu/time-get-missing-part-back_en.
[2] Ibid.

which stems from the existing irregularities in the tax system (e.g. reduced rates, exemptions, specific deductions, etc.).[3] Since it results from deliberate decisions of policy makers, we do not concentrate on this aspect in our empirical analysis.

What is important for this research, the tax gap could be analysed from the perspective of various kinds of taxes, e.g. **VAT gap, PIT gap and CIT gap**, or their aggregates (e.g. tax gap on entrepreneurial income taxes – PIT and/or CIT gap depending on the business taxation form). In some cases actions of the taxpayer may simultaneously generate different types of the tax gap (e.g. unregistered revenues may increase the VAT gap and CIT/PIT gap). Yet, in some instances, it may not be the case (e.g. unregistered labour will likely not impact the VAT gap but will likely influence the PIT gap).

For some areas of the tax gap, one can further look at the specific business actions. For example, for the VAT gap, they may include unreported sales of registered businesses, inflated costs of registered businesses, failure of businesses to register, misclassification of product/business activities, other specific kinds of frauds, etc.[4]

2.2 Shadow economy

As we mentioned in the methodological report, the **shadow (non-observed) economy** is unreported value added (≈GDP) of registered and unregistered entities that includes[5]:

- **Hidden and underground activities** where the transactions themselves are not against the law, but are unreported to avoid official scrutiny (e.g., an unreported part of companies' revenues to avoid taxation).
- **Activities described as 'informal'**, typically where no records are kept (e.g., some street vendors, etc.).
- **Illegal activities** where the parties are willing partners in an economic transaction (e.g., drug selling).
- **Household production of goods for own consumption** (not sold on the market) is sometimes also treated as the **non-monetary shadow economy**.[6]

In addition to this, to clarify the scope of the shadow economy, in Frame 1 we present various activities that are not a part of the non-observed economy.

[3] European Commission, Directorate-General for Taxation and Customs Union, Poniatowski, G., Bonch-Osmolovskiy, M., Śmietanka, A., et al. (2022), VAT gap in the EU: report 2022, Publications Office of the European Union.
[4] E.g. see Keen M., Smith S. (2007), VAT Fraud and Evasion: What Do We Know, and What Can be Done?, IMF Working Paper No. 2007/031.
[5] European Commission (2013), European System of Accounts. ESA 2010.
[6] OECD (2002), Measuring the Non-Observed Economy. A Handbook.

Frame 1. What is not included in the shadow economy?

The shadow economy and the total economy, according to the national accounts guidelines used by statistical offices (e.g. to estimate the size of GDP), exclude activities that are not related to ”production” or that are hard to valuate. For this reason, the shadow economy and the total economy exclude:

- **Illegal activities where at least one of the parties is not a willing participant** (e.g. theft) and/or that **do not lead to the creation of goods or services** (e.g. tax fraud, corruption, etc.);
- **Value of traded second-hand goods**, since such trade leads mostly to a change in ownership of the already existing goods (not to creation of new goods)[7];
- **Household ”production” of services for own consumption** (e.g. cooking for the family), since it is difficult to assign a specific monetary value to them (they are generally excluded from the national account system, e.g. from GDP calculations; imputed rents of owners-occupiers are an exception to this rule[8]).

Cash shadow economy is unregistered economic activity generated by cash payments. It was analysed in the series of EY research by EY (see e.g., EY (2019)[9]). Cash allows the seller not to report the transaction. With only a few exceptions, if an electronic payment was used instead of cash, it would be difficult to hide a transaction.

Cash shadow economy can be broken down into the ‘**passive shadow economy**’ and the ‘**committed shadow economy**’. In the ‘passive shadow economy’ consumer pays with cash (e.g., due to personal preference or lack of other payment infrastructure) and seller uses this opportunity to benefit from not reporting the transaction (consumer is often unaware of it). In such case cash is the cause of the shadow economy and policies that limit cash payments or increase their registration may help. In the ‘committed shadow economy’ the seller offers the consumer a lower price (without tax) or an opportunity to buy an illegal product/service if payment is made in cash. In such case cash is just the consequence of their joint willingness to act in the shadow economy and to tackle this problem controls, incentives, education and tax morale improvements may be needed (they are also important for the passive shadow economy).

2.3 Shadow employment

Shadow employment (also unregistered or informal employment) may be defined as an employment relationship without a formal job contract. It may happen both within unregistered and registered enterprises.

In the context of the shadow economy (which is an important source of the total tax gap), it is worth noting that the share of the shadow economy in the total economy (GDP) is something different than the share of unregistered employment in the total employment.

[7] In contrast to the value of second-hand goods, margins related to their trade are treated as “production” of services and constitute a part of either the registered or shadow economy.
[8] Imputed rents are related to housing services that homeowners implicitly provide for themselves. They are estimated to be equal to the rents that homeowners would have paid to live in dwellings of the same type, in the same district and with the same service facilities. They are included in GDP. If they were not, the GDP would be affected by changes in the share of people living in their own dwellings. It is assumed that, for example, a situation in which two homeowners living in their own dwellings start letting their dwellings to each other and paying regular rents should not affect the level of GDP. Indeed, such a change does not impact the level of GDP, since these ”new” rents have already been included in GDP as imputed rents.
[9] EY (2019), Reducing the Shadow Economy in Albania Through Electronic Payments.

Most often the first share is lower than the second one. The are several reasons for that. The value of goods and services generated by unregistered workers may be lower (compared to registered) due to factors such as lower education and skills, limited access to capital, or poor organization of work and production processes. In addition, at least part of the value of products and services generated with the use of undeclared work may be included in the registered economy (e.g. a house built with the use of some unregistered workers that is later legally sold) (see Frame 2 for an illustrative example). Moreover, undeclared work may involve fewer working hours, e.g. seasonal employment in agriculture.

While the factors outlined above explain why the share of the shadow in GDP may be lower than the share of informal workers in total employment, the opposite situation may also occur, e.g. when many businesses, despite having registered employees, do not report a significant share of their revenues to avoid paying taxes. Yet, in the case of Bulgaria, this effect rather does not outweigh the factors described above.

Frame 2. Why at least part of the value of products and services generated by unregistered employment may be included in registered (non-shadow) economy: illustrative example

- Suppose that there is a company that sold its products worth BGN 10000 to consumers but registered a revenue of only BGN 8000.
- Suppose that the only cost to this company is wages equal to BGN 6000. However, assume that only half of the wage value is officially registered, while the rest is paid in cash ”in an envelope” directly to the employees (to avoid taxes, formal actions required to register their employment, etc.).
- From the perspective of the registered GDP calculation, the labour share in the registered value added is equal to BGN 3000 (as only half of employees’ actual compensation is officially registered), while the rest of the registered value added (8000 – 3000 =BGN 5000) is reported as a return to the company’s capital.
- Even though BGN 3000 is paid in the form of unreported wages, it is reflected in the registered level of the value added. It is ”captured” in the form of the inflated company income. In other words, in this example, the shadow labour market activity results in understating the actual labour share in value added and overstating a return to the company’s capital. While it affects the structure of the generated value added, it does not influence the level of the registered economic activity.
- By contrast, the fact of not reporting some of the company’s sales to consumers does result in unreported value added of BGN 2000, i.e. it leads to an increase in the level of the shadow economy.
- As a result, from the GDP calculation standpoint, the crucial element is the registration of the final sales. In the case of wages, if somebody has unregistered workers but declared all his revenue – the workers will pay less PIT (their registered income is lower) but the employer would pay more CIT or PIT on entrepreneurial income (because there is no option to declare unregistered wages as cost). We expect that in the case of unregistered wages, part of the final sales should also be not reported (otherwise, the benefits of not registering wages are limited) and therefore our approach focuses on how much value added is not reported.

3. Shadow economy and related part of the tax gap

In this chapter we discuss our analysis of the shadow economy and related part of the tax gap. We describe the main idea and background of the applied currency demand approach (CDA), our dataset, selection of variables explaining the cash demand as well as obtained shadow economy estimates and related figures.

Section A1 of the technical appendix explains our analytical steps, data preparation process, method for estimation of econometric model and initial selection of variables with Bayesian model averaging (BMA) techniques.

3.1 Main idea and background of the method

For our analysis of the shadow economy, we use the currency demand approach (CDA). The key assumption in the currency CDA framework is that most of the unregistered transactions are settled with cash (there are some exceptions, e.g. illegal transactions with cryptocurrencies). The CDA approach aims to econometrically decompose the demand for cash into the two components: (1) cash used to facilitate the unregistered transactions (shadow cash), explained with variables described as "shadow economy determinants" and (2) cash used in the formal economy, explained with "control variables".

This idea started with early contributions of Cagan (1958)[10], followed by Gutmann (1977)[11] and Feige (1979)[12] and with important developments provided by Tanzi (1980,1983)[13]. Later, the relevant contributions were provided by Giles and Tedds (2002)[14], Embaye (2007)[15], Ahumada et al. (2008)[16], Thießen (2010)[17] and Ardizzi et al. (2014)[18], to name the few. The CDA framework was further developed by coauthors of this report, including addressing many issues encountered in the previous literature (see Dybka et al., 2019[19] and EY (2019)[20] for a detailed discussion of the issues and improvements) and analysis of uncertainty of the CDA-based shadow economy estimates (Dybka et al. 2022[21]).

[10] Cagan, P. (1958), The demand for currency relative to the total money supply, Journal of Political Economy, 66(4), 303–328.
[11] Gutmann, P. M. (1977), The subterranean economy, Financial Analysts Journal, 33(6), 26–27.
[12] Feige, E. L. (1979), How big is the irregular economy?, Challenge, 22(5), 5–13.
[13] Tanzi, V. (1980), Underground economy built on illicit pursuits is growing concern of economic policymakers, Survey no. 4–2
Tanzi, V. (1983), The underground economy in the United States: Annual estimates, 1930–80, Staff Papers (International Monetary Fund), 30(2), 283–305.
[14] Giles, D. E., Tedds, L. (2002), Taxes and the Canadian underground economy. Toronto: Canadian Tax Foundation.
[15] Embaye, A. (2007), Underground economy estimates for non-OECD countries using currency demand method, 1984–2005, MPRA Paper 20308. Germany: University Library of Munich.
[16] Ahumada, H., Alvaredo, F., & Canavese, A. (2008), The monetary method to measure the shadow economy: The forgotten problem of the initial conditions, Economics Letters, 101(2), 97–99.
[17] Thiessen, U. (2010), The shadow economy in international comparison: Options for economic policy derived from an OECD panel analysis, International Economic Journal, 24, 481–509.
[18] Ardizzi, G., Petraglia, C., Piacenza, M., & Turati, G. (2014), Measuring the underground economy with the currency demand approach: A reinterpretation of the methodology, with an application to Italy, Review of Income and Wealth, 60(4), 747–772.
[19] Dybka, P., Kowalczuk, M., Olesiński, B., Rozkrut, M., Torój A. (2019), Currency demand and MIMIC models: towards a structured hybrid method of measuring the shadow economy", International Tax and Public Finance, vol. 26(1), pages 4-40
[20] EY (2019), Reducing the Shadow Economy Through Electronic Payments. Technical appendices, https://assets.ey.com/content/dam/ey-sites/ey-com/en_pl/topics/eat/pdf/03/ey-shadow-economy-study-technical-appendices.pdf
[21] Dybka, P., Olesiński, B., Rozkrut, M., Torój, A. (2022), Measuring the model uncertainty of shadow economy estimates, International Tax and Public Finance.

3.2 Dataset and considered factors

First, we provide key information on the prepared dataset and factors that we have considered.

- **Type of data:** The data consists of various countries (including Bulgaria) observed over different years (panel dataset). Due to the availability of data, we decided to focus on the 1996-2020 period[22]. We analyzed data for 187 countries (on account of data gaps, the number of countries in the final model is equal to 101). Due to the required data structure, our analysis does not allow to consider factors accessible only for a few years, relatively low number of countries or only at the individual/sectoral level.
- **Data sources:** We used only publicly available data.[23] The main sources of the information included: International Monetary Fund, World Bank, Fraser Institute, International Labour Organization and Tax Foundation. Research projects and institutions from which we at least initially collected or considered some variables covered also: World Values Survey, European Central Bank, International Bank for Settlements, Global Findex Database, Oxford Economics, Transparency International, World Health Organization and national central banks.
- **Variable categories:** First category is the explained variable (share of the currency in circulation in the M1 monetary aggregate) that under certain conditions approximates the level and changes in the cash shadow economy. As we mentioned in section A1.1 of the technical appendix, the explanatory variables can be divided into: shadow economy determinants and control variables. While testing if the given explanatory variable is statistically significant in the econometric model is an empirical question, the assignment of the variable to the shadow or control variable category is the decision of the researcher based on theory and other studies. For example, besides from the effects related to the shadow economy, there is no simple way to explain the impact of tax or public governance quality on the currency demand. On the other hand, control variables do not impact the shadow economy but may have other influence on the currency in circulation, e.g. through the economic/technological development or changes in monetary conditions.

  For the convenience of initial listing of variables, we introduced some additional subcategories for shadow economy determinants – like labour market/business cycle, institutional/regulatory and taxation. We also tried to match the shadow economy determinants with the groups of factors for tax non-compliance identified in the literature review in the methodological report. Yet, due to the characteristics of our dataset (country-level data, some variables covering quite broad socioeconomic aspects), they often matched with more than one group. Obviously, since control variables have only an auxiliary role in the model, they are not related to factors from the literature review.
- **Alternative variables:** For different areas often more than one variable (source) was considered. The final selection was based on the number of observations and empirical analysis.

[22] For some variables less historical time periods and/or also the year 2021 were available. To estimate the shadow economy value for Bulgaria (see further) we additionally collected all available data for this country till 2022 and make some assumptions when they were missing.
[23] With the international panel data approach, even if the NRA shared with us a variable based on not publicly available data sources for Bulgaria, we would not be able to collect similar data series for other countries and, as the result, would need to exclude such variable from the analysis. Not publicly available data sources for Bulgaria were used in our VAT gap and PIT gap analyses that do not have such requirements (see further in the report).

- **Initial exclusions from the analysis**: Most often we excluded variables due to data gaps or the fact that some underlying research – like the Doing Business report – was discontinued. The same problem applies to the number of payment cards and other factors from the electronic payment system group of variables. Although such factors could bring some value to our analysis, we finally chose not to include them due to the low data availability (only after the year 2004 and for the limited number of countries).[24]
- **Consultations**: At the request of the NRA, after they saw the first proposition of our dataset, we considered several additional variables. They were mostly sociodemographic variables, some of them with a less direct theoretical link with the shadow economy. Regrettably, part of them was also quickly rejected as a result of the insufficient data coverage for various countries and time periods.

More information about our dataset could be find in section A1.2 of the technical appendix. It contains information about to which group a given variable belongs and its closest group from the literature review. It consists also of variable description and data source. You can also find there an explained decision about excluding some variables already at the initial phase of the analysis, numbers of observations, countries, and years available. We also included additional comments, among other to address the NRA's request to link some of our macroeconomic variables with publicly available forecasts (e.g. from the IMF).

In addition, our data preparation process is described in section A1.3 of the technical appendix.

3.3 Selection of variables and results of econometric model

Currency demand analysis is based on the econometric model for which two crucial components are estimation method and selection of variables.

First, even for a given set of variables, there are different econometric methods of estimation (so called estimators) of unknown parameters that describe the relationship between the explanatory and explained variables (coefficients) as well as measure their uncertainty or variability (standard errors). Our final choice of the Panel-Corrected Standard Errors estimator (further described as PCSE) is explained in section A1.4 of the technical appendix.

Second, our innovation and significant improvement in comparison with standard currency demand models includes very long list of considered factors and our approach to initial selection of variables from such list. We applied a Bayesian model averaging (BMA) procedure in which a wide array of variants (hundreds of thousands) of CDA model was estimated, with different combinations of potential variables. The goal of this was to obtain the ranking indicator showing likelihood of variables inclusion in the "true" model. This part of the analysis is described in section A1.5 of the technical appendix.

The choice of the final specification (set of variables) was made according to the method from general-to-specific.[25] Based on the results from BMA, we specified the general model containing all the variables worth further consideration. Afterwards, we tested different specifications - among others, we swapped variables within one group

[24] Another potential issue with such variables in the currency demand framework is simultaneity, i.e. the fact that they may not only influence the dependent variable but also, to some extent, be impacted by changes in this variable that have some other sources.

[25] General-to-specific is a modelling strategy in econometrics that involves starting with a general model that includes a large number of potential explanatory variables and then using a stepwise approach to systematically narrow down the set of variables to the most significant ones that best explain the variation in the dependent variable.

from BMA if their choice was ambiguous. We also checked if obtained signs of coefficients were in line with theory and other research. Next, we removed the variables that were statistically insignificant and/or had wrong signs in the stepwise manner. List of variables included in the final version of the model along with the principal information is presented in Table 1. Afterwards, one can find Table 2 with coefficients of the selected econometric model, related standard errors, and some additional statistics.

Table 1 – Variables included in the final econometric model

group of variables for our analysis	closest group(s) of factors from the literature review in the report with methodology	name of the variable	description and source
Dependent (explained) variable	Not applicable	CASH_M1	Share of the currency in circulation in the M1 monetary aggregate (currency in circulation + transferable deposits), %. Numerator's data series: Currency Outside Depository Corporations (from Depository Corporations data table) or - in case of missing data - Currency Outside Banking Institutions (from Non-Standardized Presentation data table). Denominator's data series: M1 or sum of data on Transferable Deposits and Currency from the same tables. Source: International Monetary Fund
Shadow economy determinant: institutional / regulatory	Public governance (service) quality / perceived tax service quality/ trust in government	GOV_ EFFECTIVENESS	The value of the indicator measuring the government effectiveness from the Worldwide Governance Indicators. It ranges from approximately -2.5 (low government effectiveness) to 2.5 (high government effectiveness). It reflects perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies. Source: World Bank - Worldwide Governance Indicators
Shadow economy determinant: institutional / regulatory	Public governance (service) quality / perceived tax service quality/ trust in government	INTEGRITY	Integrity of the legal system, index with values from 0 (worst) to 10 (best) The first source is the International Country Risk Guide Political Risk Component I for Law and Order: “Two measures comprising one risk component. Each sub-component equals half of the total. The ‘law’ sub-component assesses the strength and impartiality of the legal system, and the ‘order’ subcomponent assesses popular observance of the law”. The second source is Judicial Accountability, Compliance with the High Court, Judicial Review, Transparent Laws with Predictable Enforcement, and Access to Justice for Men from the V-Dem dataset. (An adjustment for the area as a whole is made later to account uniformly for gender disparities.) Each of the V-Dem variables is individually rated using the formula (Vi − Vmin) / (Vmax − Vmin)

			multiplied by 10. Vi is the country's V-Dem score according to V-Dem, and Vmax and Vmin were set at 4.0 and 0, respectively. The five measures from V-Dem are then averaged. The final number is the average of whichever of the two sources are available. Source: Economic Freedom of the World, Fraser Institute
Shadow economy determinant: labour market / role of small and specific entities	Business form/ (financial) condition of taxpayers (level) / level of economic development	FAMILY_ WORK	The ratio of the total number of contributing family workers to the population aged 15-64, %. Contributing family workers are own-account workers in the market-oriented business that is conducted by a related person who lives in the same household. Source: International Labour Organization, own calculations
Shadow economy determinant: business cycle	(Financial) condition of taxpayers (level) / shock to financial condition	UNEMP	Unemployment rate, % of total labor force (economically active population) The same definition as "unemployment rate" (percent of total labor force) in the IMF, for which there are publicly available forecasts. Source: World Bank - modeled ILO estimate
Shadow economy determinant: taxation	Tax rate / tax at risk	AVG_MAIN_ TAXES_RATE	Average rate of VAT, CIT and PIT, % Source: VAT - International Monetary Fund, CIT - Tax Foundation, PIT - Economic Freedom of the World, Fraser Institute
Control variable and for interactions with selected shadow economy determinants (to show differences in the determinants' impact depending on the country's development level)	Not applicable / level of economic development	GDP_PER_ CAPITA	GDP per capita based on purchasing power parity (PPP), thousands of constant 2017 international dollars The same definition as "Gross domestic product per capita, constant prices" (purchasing power parity; 2017 international dollar) in the IMF, for which there are publicly available forecasts. Source: World Bank - International Comparison Program, World Development Indicators database, Eurostat-OECD PPP Programme.
Control variable	Not applicable	URBAN_ POPULATION	The share of urban population in the total population, % Source: World Bank - United Nations Population Division. World Urbanization Prospects: 2018 Revision
Control variable	Not applicable	CREDIT_GDP	Domestic credit to private sector, % of GDP Source: World Bank - International Monetary Fund, International Financial Statistics and data files, and World Bank and OECD GDP estimates.
Control variable	Not applicable	INTERNET_ ACCESS	The share of population with Internet access, % Source: World Bank - International Telecommunication Union

Control variable	Not applicable	DUMMY_IND	Binary variable controlling for the effect of demonetization in India in 2016, 2016=1 Source: Own elaboration
Control variable	Not applicable	DUMMY_ROU	Binary variable controlling for the credit boom in Romania starting in 2007, 2007-2010=1 Source: Own elaboration

Source: EY.

Table 2 – Coefficients in the final econometric model of the currency demand

	Dependent variable: CASH_M1	
		PCSE_psar1
Shadow economy determinants	GOV_EFFECTIVENESS	-4.0273***
		(1.237)
	GOV_EFFECTIVENESS_interacted	0.0934***
		(0.031)
	INTEGRITY	-0.5642*
		(0.336)
	FAMILY_WORK	0.7788***
		(0.136)
	UNEMP	0.4286***
		(0.117)
	UNEMP_interacted	-0.0090***
		(0.003)
	AVG_MAIN_TAXES_RATE	0.3524***
		(0.123)
Control variables	GDP_PER_CAPITA	-0.2345***
		(0.073)
	URBAN_POPULATION	-0.2722***
		(0.071)
	CREDIT_GDP	0.0183**
		(0.008)
	INTERNET_ACCESS	-0.1087***
		(0.020)
	DUMMY_IND	-18.8906***
		(2.028)
	DUMMY_ROU	-9.3748***
		(1.679)
	constant	62.5644***
		(8.492)
	Observations	1579
	Groups	101

Notes: Standard errors in parentheses. P-values marked with asterisks: *p<0.1, ** p<0.05, ***p<0.01.[26] Groups = number of countries included in the sample.

Source: EY.

[26] The p-value is a statistical measure used in econometrics to determine the strength of the evidence supporting a particular relationship between variables. It is a number between 0 and 1 that represents the likelihood of observing the given data or more extreme data, assuming that there is no relationship between the variables.

Since the explained variable (CASH_M1) is not only related to the shadow economy determinants but also to non-shadow economy related processes (e.g. shifting cash into deposits due to various macroeconomic and technological conditions) we needed to decide (based on other research and theory) which variables are (1) shadow economy determinants and which (2) control variables. According to our best understanding, the first group should include the following factors: GOV_EFFECTIVENESS, INTEGRITY, FAMILY_WORK, UNEMP and AVG__MAIN_TAXES_RATE. The remaining variables belong to the second group.[27] To control for potential differences in the impact of factors at different levels of development, we tested interaction terms with GDP_PER_CAPITA and included in the final model the ones that were statistically significant.

With the results obtained, we can calculate the theoretical value of the dependent variable CASH_M1 in Bulgaria.[28] The formula is as follows:

$$CASH\_M1_{BG,t} = MIN(-4.0273 + 0.0934 * GDP\_PER\_CAPITA_{BG,t}; 0) * GOV\_EFFECTIVENESS_{BG,t} - 0.5642 * INTEGRITY_{BG,t} + 0.7788 * FAMILY\_WORK_{BG,t} + MAX(0.4286 - 0.0090 * GDP\_PER\_CAPITA_{BG,t}; 0) * UNEMP_{BG,t} + 0.3524 * AVG\_MAIN\_TAXES\_RATE_{BG,t} - 0.2345 * GDP\_PER\_CAPITA_{BG,t} - 0.2722 * URBAN\_POPULATION_{BG,t} + 0.0183 * CREDIT\_GDP_{BG,t} - 0.1087 * INTERNET\_ACCESS_{BG,t} - 18.8906 * DUMMY\_IND_{BG,t} - 9.3748 * DUMMY\_ROU_{BG,t} - 4.3539 * FIXED\_EFFECT_{BG,t} + 62.564,$$

where *BG* denotes data for Bulgaria and *t* is time subscript. MIN and MAX denote minimum and maximum functions. They were added to the equation for variables interacted with GDP_PER_CAPITA with zeros as the second arguments. Without MIN and MAX, the interaction terms for very high level of GDP_PER_CAPITA would cause the direction of the impact of the interacted variables to reverse. In practice, Bulgaria is far from such levels of GDP_PER_CAPITA, so for the short-term analysis the MIN and MAX functions could be neglected. In the future, when the country is close to such thresholds, one should reestimate the econometric model to account for the changed specifics of the country.

DUMMY_IND and DUMMY_ROU are dummy variables controlling for specific observations for India and Romania, so for Bulgaria their values are equal to zero.

Fixed effects are country-level individual effects, which represent time-invariant, unobservable country characteristics that affect the shadow cash to M1 ratio in each country. For Bulgaria they are estimated at -4,35 with p-value = 0,6. While it should be

[27] Worth explaining is the assignment of GDP_PER_CAPITA to the control group. There are a few reasons for this. First, it is quite common approach in other currency demand research. Second, while the development of the economy can affect the role of electronic payments (e.g. through better payments infrastructure), they serve to large extent for registered transactions and only a part of additional cashless payments crowds out unregistered transactions. Third, one can argue that a large part of GDP_PER_CAPITA and shadow economy negative correlation is due to other related factors that accompany or often proceed the economic development such as improvements in government effectiveness and other aspects of public policy. Since they are among our shadow economy determinants, the economic development impact should be already adjusted for their influence in our model and, thus, mostly related to registered cash transactions. Fourth, GDP_PER_CAPITA could be moderating the impact of shadow economy determinants. For instance, in countries where the general level of development is high, a one percentage point increase in unemployment can lead to a lower increase in the shadow economy compared to the less affluent economies (e.g. there could be more alternative legal sources of income in case of losing job). As a result, we have included in the model the so-called interaction terms between the GDP and shadow economy determinants that account for the diminishing (with economic development) scale of the shadow economy determinants effect.

[28] It could also be used for other country with other parameter in front of the country-specific fixed effect and values of variables for the given country. The same applies to the next formula.

included while calculating the value from the formula above, since the related parameter is not statistically significant, we can assume that unobservable cultural factors are not among the key determinants of the cash demand and shadow economy in Bulgaria.

We can also obtain the ratio of the shadow cash (cash in circulation related to shadow economy determinants) to M1 estimate for Bulgaria from the following formula:

$$
\begin{aligned}
SHADOW\_CASH\_M1_{BG,t} &= MIN(-4.0273 + 0.0934 * GDP\_PER\_CAPITA_{BG,t}; 0) \\
&* (GOV\_EFFECTIVENESS_{BG,t} - MAX(GOV\_EFFECTIVENESS)) \\
&- 0.5642 * (INTEGRITY_{BG,t} - MAX(INTEGRITY)) + 0.7788 \\
&* (FAMILY\_WORK_{BG,t} - MIN(FAMILY\_WORK)) + MAX(0.4286 \\
&- 0.009 * GDP\_PER\_CAPITA_{BG,t}; 0) * (UNEMP_{BG,t} - MIN(UNEMP)) \\
&+ 0.3524 * (AVG\_MAIN\_TAXES\_RATE_{BG,t} \\
&- MIN(AVG\_MAIN\_TAXES\_RATE))
\end{aligned}
$$

In the equation above we include only the shadow economy determinants. To estimate their contribution to the shadow cash, we calculate the difference between their values for Bulgaria and the benchmarks included in our sample. The benchmarks are “the best values” of the shadow economy determinants present in our sample of different countries and time periods (minimum (maximum) value in case of variables that increase (decrease) the shadow economy). If a given variable reached the level of the benchmark in Bulgaria, its contribution to the shadow would be equal to zero.

With some additional assumptions and operations the ratio of shadow cash to M1 could be further translated into the share of the cash shadow economy in the total economy (GDP) (for details see section A1.1 of the technical appendix[29]).

The estimated coefficients in the econometric model should be interpreted in the way described in Table 3.

Table 3 – Interpretation of the coefficients in the final econometric model

name of the variable	variable interpretation
GOV_EFFECTIVENESS	Due to inclusion of the interaction term the effect of GOV_EFFECTIVENESS is different across countries and for a given country over time (if its income level changes). For example, in the country where the GDP_PER_CAPIA equals to 10 (thousands of international PPP dollars in constant 2017 prices) an increase in GOV_EFFECTIVENESS by 1 unit is associated with 3.1 (= 4.03-10*0.093) pp decrease in the shadow cash to M1 ratio. In the case of Bulgaria in 2022 an increase in GOV_EFFECTIVENESS by 1 unit is associated with 1.65 pp decrease in the shadow cash to M1 ratio which is equivalent of a decrease in shadow economy by 1.5% of GDP.

[29] In short, it is related to the fact that basic cash shadow economy estimates from the currency demand model should be interpreted as percent of monetary economy (i.e. economy related to monetary transactions), not percent of the total economy (GDP). To move from one terms to the other, one need to multiply the initial results by the share of the monetary economy in the total economy. The monetary economy is estimated as the total economy minus non-monetary shadow economy and so called imputed rents of owners-occupiers.

INTEGRITY	An increase in INTEGRITY by 1 unit is associated on average with 0.56 pp decrease in the shadow cash to M1 ratio which is equivalent of a decrease in shadow economy by 0.51% of GDP.
FAMILY_WORK	An increase in FAMILY_WORK by 1 percentage point is related on average with 0.78 pp growth in the shadow cash to M1 ratio which is equivalent of a increase in shadow economy by 0.71% of GDP.
UNEMP	Due to inclusion of the interaction term the effect of UNEMP is different across the countries and for a given country over time (if its income level changes). For example, in the country where the GDP per capita equals to 10 (thousands of international PPP dollars in constant 2017 prices) an increase in unemployment rate by 1 percentage point is associated with 0.34 (= 0.43-10*0.009) pp increase in the shadow cash to M1 ratio. In the case of Bulgaria in 2022 an increase in unemployment rate by 1 percentage point is associated with 0.2 pp increase in the shadow cash to M1 ratio which is equivalent of an increase in shadow economy by 0.18% of GDP.
AVG_MAIN_TAXES_RATE	An increase in AVG_MAIN_TAXES_RATE by 1 pp is associated on average with 0.35 pp increase in the shadow cash to M1 ratio which is equivalent of a increase in shadow economy by 0.32% of GDP.

Source: EY.

We can see that the signs of the obtained estimates for the shadow economy variables are in accordance with the theory and/or other research:

- Higher values for variables **GOV_EFFECTIVENESS** and **INTEGRITY** have a limiting effect on the shadow economy, likely through their multichannel impact on taxpayers' behaviour and attitudes.
- Elevated role of **FAMILY_WORK** reflects the popularity of specific relations on the labour market that likely support activity in the shadow economy.
- Increased **UNEMP** captures the worse situation on the labour market, resulting among others from the business cycle, which may encourage people to increase their unregistered activity.
- Higher tax rates (**AVG_MAIN_TAXES_RATE**) increase the costs and stimulate avoidance of reported business operations. Yet, it is worth noting that a growth in tax rates, despite leading to some expansion of the shadow economy, is still likely to increase collected tax revenues (net effect depends on both changes in the shadow economy and non-shadow-economy activity due to the higher rates). In theory, especially in the long term, additional government revenues may support the government effectiveness and integrity (such potential link is not captured by our model).
- Last but not least, in our model, the impact of **GOV_EFFECTIVENESS** and **UNEMP** declines with the economic development level. For joblessness in higher income countries, it could be linked with lower incentives or opportunities to engage in unregistered activity despite turbulences on the labour market, e.g., due to more accumulated savings and wealth, more available social security, better options to borrow money, etc. For improvements in GOV_EFFECTIVENESS, one can speculate that their impact on the shadow economy is lower in more developed countries, e.g., due to the structural differences in the economy (e.g. higher role of

large enterprises that are less likely to not report their operations) or the fact that more affluent people are less interested in risky behaviour.

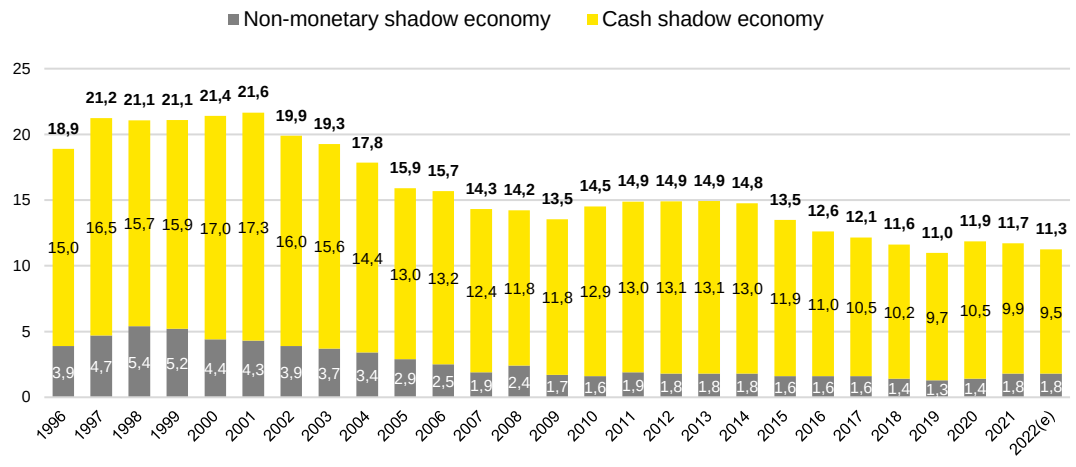
- It is worth mentioning that we also tested in our model so called time effects for years 2020 (and 2021 in some specifications) – which were the years of the COVID-19 pandemic. Yet, at least for the average cash demand and shadow economy in the analysed sample of countries, these years were not significantly different than the other years (after controlling for the level of variables included in our model).

3.4 Shadow economy estimates and role of different factors

3.4.1 Total, cash and non-monetary shadow economy

It is worth recalling that in our approach the total shadow economy is the sum of the cash shadow economy from the CDA model and the non-monetary shadow economy.[30] Chart 1 shows such estimates for Bulgaria and their evolution over time.

Chart 1 – Total, cash and non-monetary shadow economy in Bulgaria (% of GDP)



Notes: Generation of the chart above for all periods required some additional assumptions. The share of agriculture in GDP in 1997, which was the basis to non-monetary shadow economy calculations, was interpolated from years 1996 and 1998 as we detected that the original value in the database was an outlier. For the year 2021 we imputed missing data for INTEGRITY and AVG_MAIN_TAXES_RATE – we know that for the latter there was no change vs. the previous year, while for the former variable we assumed that. We also provide initial estimates of the shadow economy in 2022. The value for UNEMP comes from International Labour Organization modelled estimates. Variables FAMILY_WORK, and the amount of the imputed rents (used in translating our results into % of GDP) were calculated on the basis of the dynamics from previous years. The GDP_PER_CAPITA in 2022 was calculated on the basis of the change of real GDP per capita expressed in PPP forecasted in the IMF World Economic Outlook (October edition). For GOV_EFFECTIVENESS, INTEGRITY, AGRI_GDP and AVG_MAIN_TAXES_RATE we took the same value as in the last available year.

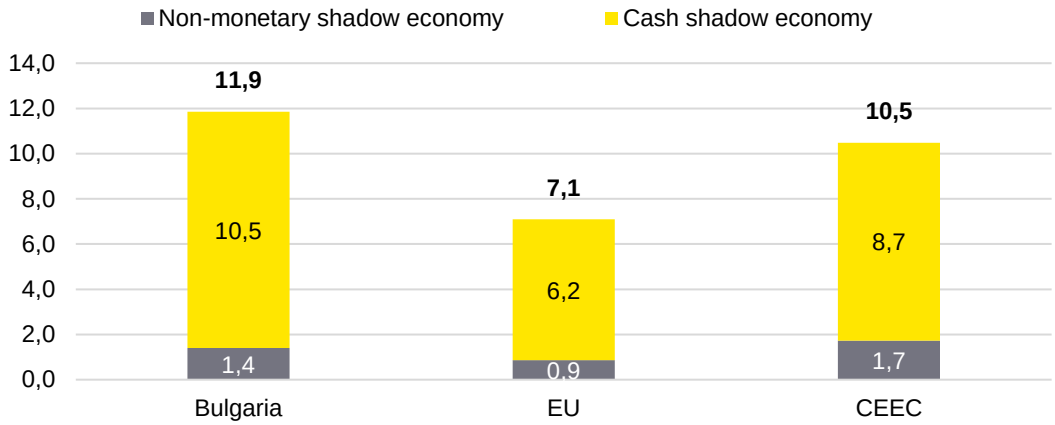
Source: EY.

We estimate that in 2022 the total shadow economy in Bulgaria amounted to 11.3% of GDP, with the majority of the value related to the cash shadow economy (9.5% of GDP vs 1.8% of GDP for the non-monetary shadow economy). In general, we can see a downward long-term trend in the non-observed economic activity in Bulgaria since 2001 as well as some cyclical fluctuations (e.g. after the 2009's recession and in the pandemic year of 2020). In the years 1996-2004 the total shadow economy was in

[30] A non-monetary shadow economy is a specific component that depending on the applied definition may sometimes be or not be included in the scope of the shadow economy.

between 17 and 22% of GDP, then it dropped below 16% of GDP and after the year 2016 it was around 11-12% of GDP. For the two components, the tendency was often similar. The contribution of the different factors to the cash shadow economy estimates is discussed in the next section.

Chart 2 – Total, cash and non-monetary shadow economy in Bulgaria, EU and CEEC in 2020 (% of GDP)



Source: EY.

For the last year with fully available data for multiple countries, that is 2020[31], the average total shadow economy in the European Union was estimated at 7.1% of GDP (cash shadow economy – 6.2% of GDP, non-monetary shadow economy – 0.9% of GDP). In turn, for the CEEC countries[32] such average amounted to 10.7% of GDP (cash shadow economy – 9% of GDP, non-monetary shadow economy – 1.7% of GDP). In Bulgaria in 2020, the total shadow economy estimate was equal to 11.9% of GDP (cash shadow economy – 10.5% of GDP and non-monetary shadow economy - 1.4% of GDP), so it was slightly higher than in the CEEC region and significantly larger than in the EU (see Chart 2).

Due to the fact that the shadow economy is not directly observable, it is hard to find other reliable figures that could be compared with our results. Unfortunately, during the data collection process for this project, we have not succeeded in obtaining any up-to-date non-observed economy estimates from the National Statistical Institute in Bulgaria. Fernandes (2022) summarizes this kind of data collection effort for various statistical offices in the EU Member States over many years.[33] For Bulgaria, the newest non-observed economy estimate is for 2001 and amounts to 10.2% of GDP. Yet, in the same research, the estimate for Bulgaria in 2000, collected during a different round of the study, amounted to 16.3% of GDP. As some additional reference points, we could mention the estimates of the statistical offices in Czechia in 2018 (9.0% of GDP), Italy in 2016 (14.9% of GDP), Romania in 2019 (27.5% of GDP) and Slovakia in 2018 (18.9%). The author of the summary concludes that “these figures depend heavily on national accounts compilation particularities in each Member State”.

[31] For 2021 there was no data for variables INTEGRITY and AVG_MAIN_TAXES_RATE.
[32] According to the Organisation for Economic Co-operation and Development: "Central and Eastern European Countries (CEECs) is an OECD term for the group of countries comprising Albania, Bulgaria, Croatia, the Czech Republic, Hungary, Poland, Romania, the Slovak Republic, Slovenia, and the three Baltic States: Estonia, Latvia and Lithuania”.
[33] Fernandes A. (2022), The non-observed economy in the national accounts, KU Leuven Working Paper, October 2022

When it comes to other shadow economy estimates conducted with macroeconomic approaches, there are many issues related to their methodology (see Dybka et al. (2018)[34]). Yet, e.g. Medina and Schneider (2018), depending on the selected approach, obtained the shadow economy estimate for Bulgaria in 2017 in the range of 19.2-29.6% of GDP (macro and adjusted MIMIC estimates).[35] The MIMIC shadow economy estimates in this (also other similar) research are generally high. For example, they amounted to 7.9-12.1% of GDP in Sweden in 2017, while Fernandes (2022) shows the statistical office's estimate of the non-observed economy for this country in 2015 at 3.0% of GDP. Medina and Schneider also show for Bulgaria the average shadow economy estimate over the 1991-2015 period obtained with the Predictive Mean Matching (PMM), which amounts to 23.3% of GDP for Bulgaria.

It is worth noting that the likely higher share of unregistered employment in the total employment in Bulgaria (vs the share of the shadow economy in GDP) does not imply the same share of the shadow economy in GDP. The reasons include, among others, relatively low value added generated by unregistered employees and the fact that some of this value may be finally registered, e.g. a new building (see section 2.3 for more detailed discussion).

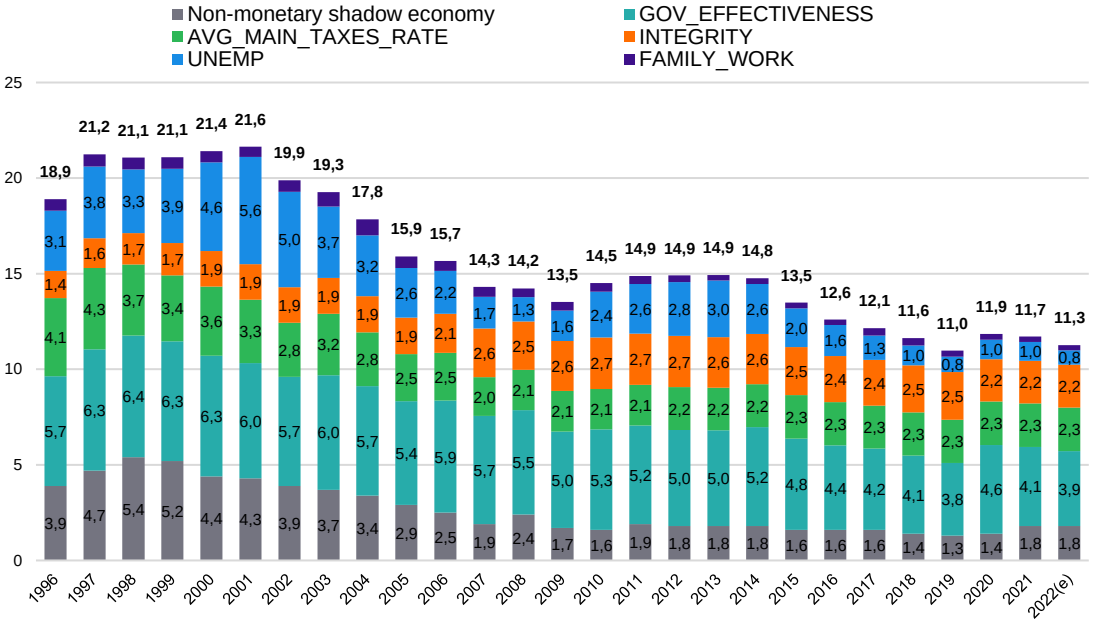
When considering shadow economy estimates as percent of GDP one should remember that a significant share of GDP is generated by the public sector and public companies as well as various large private companies that are unlikely to not report their economic activity (they may generate some tax gap in other ways, though). This means that almost the whole shadow economy should be included in the remaining part of the economy, accounting there for a much larger share of the value added than was reported for the total GDP. Similarly, looking from the expenditure side of GDP, shadow economy transactions are not likely to happen within most government, investment and foreign expenditure (exports), meaning that they are likely concentrated mostly within consumption expenditure.

3.4.2 Contribution of factors to the cash shadow economy

Using the second equation from the section on the final econometric model, we estimated the contribution of different factors to the cash and total shadow economy in Bulgaria[36]. They are presented in Chart 3.

[34] Dybka, P., Kowalczuk M., Olesiński B., Rozkrut M., Torój A. (2019), Currency demand and MIMIC models: towards a structured hybrid method of measuring the shadow economy, International Tax and Public Finance, vol. 26(1), pages 4-40.
[35] Medina, L., Schneider F. (2018), Shadow Economies Around the World: What Did We Learn Over the Last 20 Years?, IMF Working Paper No. 2018/017
[36] For the non-monetary shadow economy we only know that its majority is most often related to agricultural outputs for own final use.

Chart 3 – Contribution of different factors to the total and cash shadow economy in Bulgaria (% of GDP)



Notes: The estimates for 2022 and the earlier years with the missing data obtained in the same way as described in the notes to the chart with the total, cash and non-monetary shadow economy above.

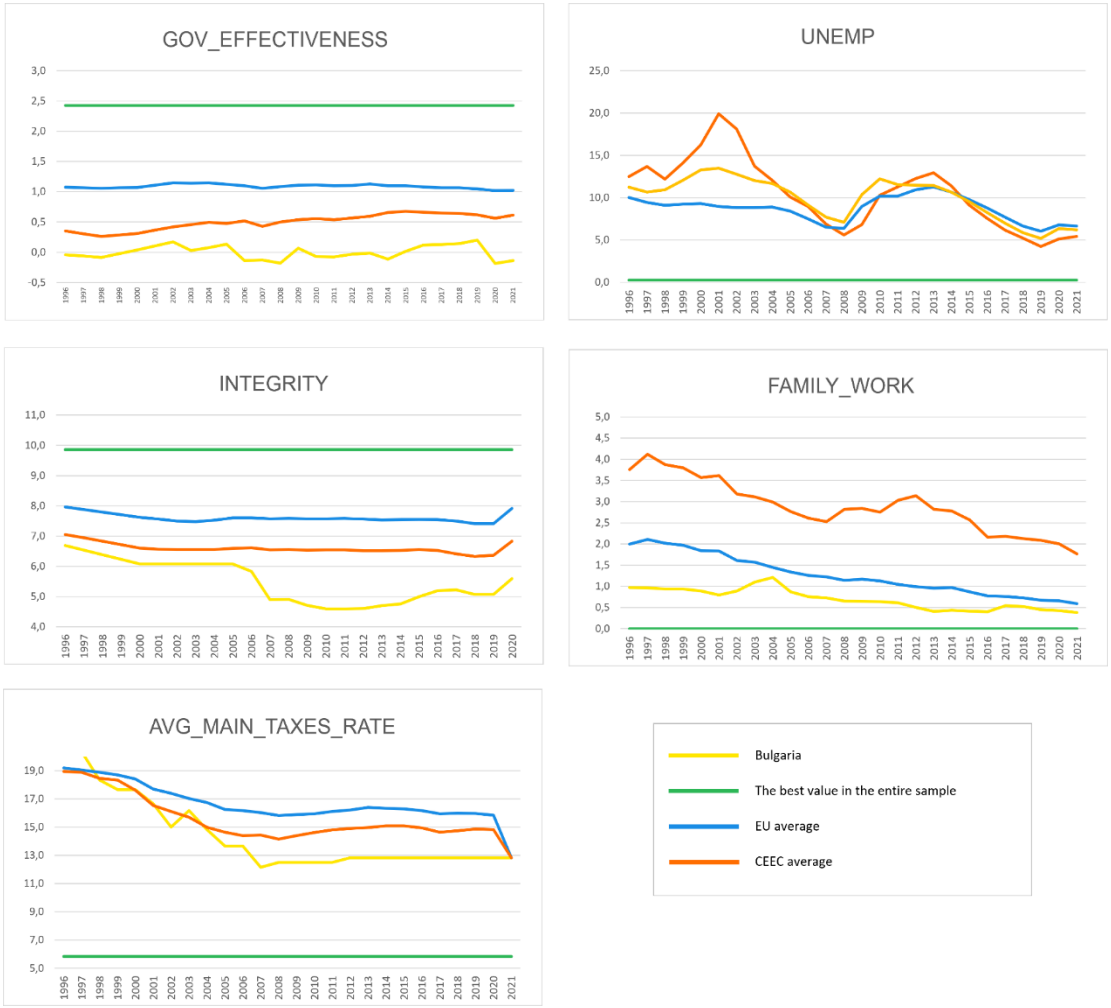
Source: EY.

We can make the following observations:

- In 2022, the key factor contributing to the shadow economy size in Bulgaria included GOV_EFFECTIVENESS (3.9% of GDP), followed by AVG_MAIN_TAXES_RATE (2.3% of GDP) and INTEGRITY (2.2% of GDP) as well as small contributions of UNEMP and FAMILY_WORK.
- The ranking of factors was similar over time, with the exception of higher role of UNEMP than INTEGRITY in the past, especially before 2007.
- In the long term, the GOV_EFFECTIVENESS and UNEMP contributions were in the downward trend. For GOV_EFFECTIVENESS it was mostly due to the interaction with GDP_PER_CAPITA, which was growing over most of the time and, as the result, decreasing the role of GOV_EFFECTIVENSS. For UNEMP it was both the effect of the similar interaction mechanism as well as improvement of the state in the labour market after 2013.
- The role of AVG_MAIN_TAXES_RATE and INTEGRITY in explaining the shadow economy level in Bulgaria has been relatively stable sine about 2005-2006.

To further analyse the impact of various factors, one may also see Chart 4 with the variables evolution in Bulgaria, their distance to the benchmarks in our sample as well as similar values for the EU and CEEC regions. They are also useful in the context of future scenarios of different variables (see further). For example, they suggest that in the future, without significant reforms/structural changes, there may be no reason to assume a positive trend (improvement) in GOV_EFFECTIVENESS or INTEGRITY in Bulgaria, while there is some support in data to assume a continuation of the downward trend in FAMILY_WORK.

Chart 4 – Historical values of the variables



Source: EY.

3.4.3 Passive and committed components of the cash shadow economy

While it is not the main focus of this research, our team has a large experience in the analysis of the role of cash and potential promotion of electronic payments (registration of cash payments) in the combat of the shadow economy.[37] Taking advantage of this background, we provide below an additional analysis of the cash shadow economy in Bulgaria.

In general, cash allows the seller not to report the transaction. With only a few exceptions, if an electronic payment was used instead of cash, it would be difficult to hide a transaction. While approximating the size of the cash shadow economy by estimating the value of unreported cash transactions, we distinguish two categories of the cash shadow economy, each to be addressed by different measures. The key differentiating factor between these two components of the cash shadow economy is the causal relationship between cash payments and the shadow economy. In the first category, cash payments contribute to the expansion of the shadow economy, while

[37] E.g. see EY (2019), Reducing the Shadow Economy in Albania through Electronic Payments (and technical appendices). Also many similar studies for other countries.

in the second component the increased cash payments are simply a result of shadow economy activities. We therefore distinguish situations where:

- Cash is a **cause** (or one of the causes) of the shadow economy,

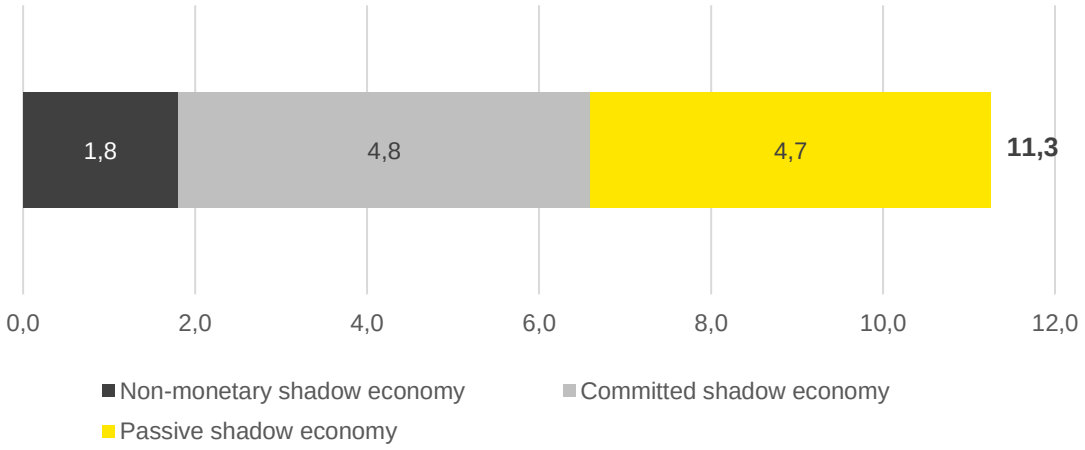
from situations where:

- Cash is a **consequence** of the shadow economy.

We call the first component of the cash shadow economy '**passive shadow economy**' and refer to transactions where consumer pays with cash (e.g., due to personal preference or lack of other payment infrastructure) and seller uses this opportunity to benefit from not reporting the transaction (consumer is often unaware of it). In such case cash is the cause of the shadow economy and policies that limit cash payments or increase their registration may help. The second component, '**committed shadow economy**' is the remaining part of the cash shadow economy, where it is not the cash payment that influences the decision of the seller not to report the transaction, but the motivation of both sides of the transaction to benefit from evading tax liabilities or to sell/buy illegal products/services. The cash form of payment is (usually) still required to hide the transaction, but it is no longer the source of illegal activity, but rather its outcome. As a result, the committed shadow economy requires different approach than passive that includes labour inspections, tax controls or reduction of administrative burden related to compliance with the regulations.

Obtained results indicate that the committed part of the shadow economy is equal to 4.8% of GDP (constitutes 50.8% of the cash shadow economy) and the passive component is equal to 4.5% of GDP (49.2% of the cash shadow economy), which indicates that promotion of electronic payments (or registration of additional cash payments) can play an important role in limiting the shadow economy in Bulgaria (see Chart 5).

Chart 5 – Decomposition of the shadow economy in Bulgaria into non-monetary, committed and passive components in 2022 (% of GDP)

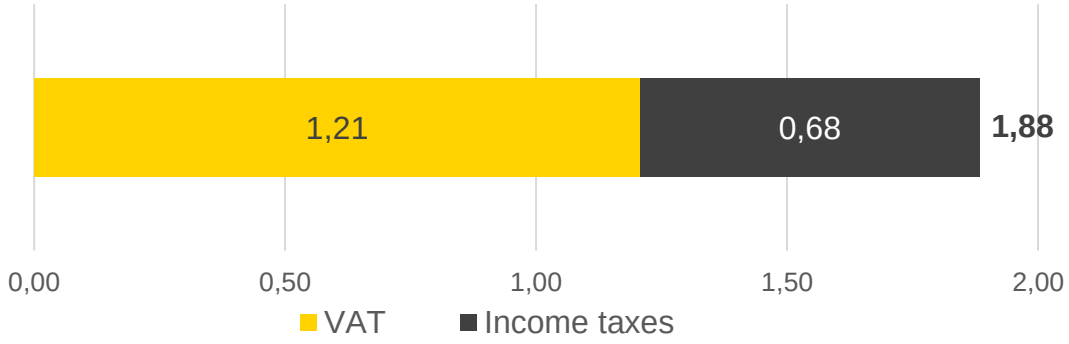


Notes: The estimation for 2022 is the same as in the chart with the total, cash and non-monetary shadow economy described above.

3.5 Lost government revenues due to the shadow economy

The obtained results show that potential government revenues from eliminating the cash shadow economy in Bulgaria amounted in 2022 to 1.88% of GDP, out of which 1.21% of GDP was related to VAT, whereas 0.68% of GDP was related to income taxes (PIT and CIT) (Chart 6). Consequently, even a partial success in dealing with unregistered transactions can significantly improve the public finance situation in Bulgaria.

Chart 6 – Lost government revenues due to cash shadow economy in Bulgaria in 2022 (% of GDP)



Source: EY elaboration.

One point of reference for the obtained results is the European Commission (2022)[38] research on the VAT gap in the EU. First, it is worth noting that the VAT gap is related to lost VAT revenues due to the shadow economy (not the shadow economy per se) but also to other sources (e.g. VAT frauds, bankruptcies, etc.). Second, non-monetary shadow economy rather does not generate lost VAT revenues. Anyway, in line with our shadow economy estimates, the authors show a downward trend in the VAT compliance gap (in % of VAT total tax liability) in Bulgaria since 2016. Yet, the difference in their VAT gap estimates between 2020 (6.3%) and 2019 (9.7%) is quite significant and in contrast to our results with a growth in the shadow economy in the first year of the pandemic, which seems quite intuitive. Our shadow economy estimates indicate that the VAT lost due to shadow economy in Bulgaria amounted to 11.7% of total VAT that should be collected[39] in 2019 and 12.6% in 2020, which is above the European Commission's study results. Nevertheless, it is worth noting that the authors of this study comment their figures for Bulgaria as "estimates based on some very outdated information or very large unexplained volatility of estimates". In other words, they are not very reliable.

[38] European Commission, Directorate-General for Taxation and Customs Union, Poniatowski, G., Bonch-Osmolovskiy, M., Śmietanka, A., et al. (2022), VAT gap in the EU: report 2022, Publications Office of the European Union,

[39] We have divided our estimate of lost VAT revenues due to cash shadow economy by the sum of collected VAT revenues (obtained from the Bulgarian Ministry of Finance) and our estimate of lost VAT revenues due to cash shadow economy.

4. Unregistered income and the PIT gap

In this chapter we describe our analysis of unregistered household income and the PIT (also social security contributions) gap. First, we discuss the main idea and background of our analytical method. Second, we summarize the used dataset and results of econometric models. Third, we translate such results into country-level estimates of tax non-compliance. Finally, we extend our analysis and discuss differences in income underreporting between various socio-economic groups.

Section A2 of the technical appendix explains the data preparation process, applied classification of households and method for estimation of the econometric model as well as derivation of various results and some of the technical terms mentioned in the text below (often names of sections in the technical appendix correspond to related parts of the main report).

4.1 Main idea and background of the method

Traces-of-true-income approach (otherwise called Pissarides-Weber (PW) method or expenditure method) is an indirect method for estimating the extent of income underreporting among households or individuals and related PIT gap. It is based on discrepancies in expenditure and reported income pattern identified through econometric modelling of micro data and therefore it allows for identification of socio-demographic characteristics of taxpayers – such as sex, age, level of education and sector of employment – that can be associated with lower tax-compliance.

Pissarides and Weber (1989[40]) first provided an estimation framework for assessing the scale of underreporting among self-employed in the UK by comparing the relationship between food expenditure and income of the self-employed to that of the employees who were assumed to be fully compliant. The authors assumed that how much someone spends on food is based on their true income and socio-demographic characteristics, but not on whether they work as self-employed or employees. In addition, the opportunity to hide income was considered to be much greater for self-employed than employees. Therefore, if food expenditure was higher for self-employed than for employees for a given level of income, this would indicate underreporting of income by self-employed. Using the results from 1982 Family Expenditure Survey, the authors estimated that the average true self-employed income in the UK was 1.55 times as much as what was reported.

Traces-of-true-income approach is now well-established in the literature and several changes to the original framework have been tested, including:

- **Using data on reported income from tax returns instead of surveys.** Feldman and Slemrod (2007[41]) conducted the analysis for the US relying solely on unaudited tax returns data by using charitable contributions reported for tax purposes as an expenditure variable. However, the assumption that charity expenditure does not depend on the employment status is considered by the authors to be stronger than the respective assumption about food. Tax returns data directly matched with household budget survey data were used by Paulus

[40] Pissarides, C. A., & Weber, G. (1989). An expenditure-based estimate of Britain's black economy. Journal of public economics, 39(1), 17-32.
[41] Feldman, N. E., & Slemrod, J. (2007). Estimating tax noncompliance with evidence from unaudited tax returns. The Economic Journal, 117(518), 327-352.

(2015[42]) in the case of Estonia and by Cabral, Gemmell and Alinaghi (2021[43]) in the case of New Zealand. Thanks to matching two data sources, the authors could use the most reliable source of reported income data while having access to a wide set of expenditure variables which are considered to be fairly well reported in surveys. The former study even provided the comparison of the results under different measures of income. The estimated level of underreporting by self-employed turned out to be about two times lower when using income estimates from surveys instead of tax returns. Therefore, measurement error typical for survey income estimates and associated attenuation bias may lead to significant underestimation of the level of noncompliance.

- **Using public sector employees instead of all employees as a reference group less prone to tax evasion.** The assumption that all employees honestly report their income is considered untenable in some countries where employees have strong incentives and many opportunities to hide their income (e.g., by so-called envelope wages). If this is the case, comparing income and expenditure patterns of self-employed to those of employees would lead to underestimated or insignificant results. Instead, some authors (see Ekici and Besim, 2014[44], in the case of North Cyprus), decided to treat public sector employees as a reference group and estimate the level of income underreporting and tax evasion for self-employed as well as private sector employees.
- **Using other expenditure categories than food as a "trace of true income".** Food expenditure is the baseline option; however, other expenditure categories can be considered in the case of lack of data (this was the reason for using charity contributions in the 2007 study by Feldman and Slemrod[45] and for using spending on utilities in the 2015 study by Paulus[46]). Other expenditure categories are also used for the purposes of sensitivity analysis.

A comparative study by Kukk, Paulus and Staehr (2020[47]) was the first in which the method was used for a large number of countries (14 EU countries including Bulgaria) using common specification of the model and harmonized microdata (2010 wave of the EU Household Budget Survey). The results indicated that the level of underreporting of income by self-employed varies between those countries from under 10% to over 40% of declared income and that those differences are not associated with the level of countries' development. One of the potential reasons for relatively low estimates for Southern European countries (including Bulgaria) – as explained by the authors – was using all employees as a reference group while in the case of those countries private sector employees may be to a larger extent engaged in tax evasion.

4.2 Dataset and considered factors

Traces-of-true-income approach (otherwise called Pissarides-Weber (PW) method[48] or expenditure method) to estimating the level of income underreporting by individuals is based on micro-data covering socio-demographic characteristics (sex, age, education, marital status, economic activity, employment status etc.), incomes, and

[42] Paulus, A. (2015). Income underreporting based on income expenditure gaps: Survey vs tax records (No. 2015-15). ISER Working Paper Series.
[43] Cabral, A. C. G., Gemmell, N., & Alinaghi, N. (2021). Are survey-based self-employment income underreporting estimates biased? New evidence from matched register and survey data. International Tax and Public Finance, 28(2), 284-322.
[44] Ekici, T., & Besim, M. (2016). A measure of the shadow economy in a small economy: Evidence from household-level expenditure patterns. Review of Income and Wealth, 62(1), 145-160.
[45] Feldman, N. E., & Slemrod, J. (2007). Ibid.
[46] Paulus, A. (2015). Ibid
[47] Kukk, M., Paulus, A., & Staehr, K. (2020). Cheating in Europe: underreporting of self-employment income in comparative perspective. International Tax and Public Finance, 27(2), 363-390.
[48] Pissarides, C. A., & Weber, G. (1989). An expenditure-based estimate of Britain's black economy. Journal of public economics, 39(1), 17-32.

expenditures of Bulgarian households. The dataset that we used for this particular analysis is not publicly available. It was prepared in an anonymised form by the National Statistical Institute. Specifically, the dataset contains anonymized individual-level and household-level data merged from two sources: Household Budget Survey data (standard approach) extended by the information on income from National Revenue Administration data on annual tax returns for 2017, 2018, 2019, 2021.

- **The Household Budget Survey (HBS)** is carried out annually (apart from the one-year break in 2020 due to the outbreak of the Covid-19 pandemic) by the National Statistical Institute. The sample is drawn from all households in Bulgaria using probability sampling – two-stage cluster sampling method. This method of sample selection guarantees the possibility of drawing conclusions for the entire population of Bulgaria using survey results and population weights assigned to interviewed households (the sum of those weights is equal to the sum of households in Bulgaria). The great advantage of this data is therefore the possibility to look at the consumption behavior of Bulgarian households depending on their income and other socio-demographic characteristics. The risk factor, however, is primarily the measurement error associated to the greatest extent with the collection of sensitive information via questionnaires, such as income or expenses for socially undesirable products, e.g. tobacco and alcohol. Because of this measurement risk in the HBS data on household incomes, the ideal solution while estimating the level of underreporting with the use of traces-of-true income approach is to use data on income as reported to the tax authorities and the remaining data (including data on expenditure) from the HBS.
- The initial dataset can be broken down into the following categories:
    - **Individual-level socio-demographic data**
    - **Household-level socio-demographic data**
    - **Household-level data on income** based on declarations in the HBS
    - **Household-level data on expenditures,** incl. food eaten at home and expenses in restaurants and hotels
    - **Main income source** of the taxpayer broken down into self-employment, employment in public sector and employment in private sector.
    - **Information from the Annual Tax Return** including gross taxable income, taxable bases, tax reductions and exemptions, social security contributions paid in Bulgaria and abroad, amount of personal income tax due, net income calculated based on gross income, social security contributions and personal income tax
    - **Information on taxable income declared by employers as well as self-employed persons**, which was available for persons whose only source of income was employment contract, therefore, they were not obliged to file an annual tax return.
- Additionally, the NRA provided us with macro-level data on average net and gross labour income, personal income tax and social security contributions paid by private sector employees, public sector employees and self-employed broken down by sex, age and industry. In addition, we received information on the number of people classified to those groups over the years. Such data points were employed for checking consistency of some micro data with country aggregates as well as for some supplementary calculations described later. Finally, we used several publicly available macro statistics, the sources of which we cite in the related sections of the report.

Section A2.1 of the technical appendix describes in detail our data preparation process, while section A2.2 explains applied classification of households.

4.3 Results of econometric models

The list of variables included in the final version of the PW model along with their short description is presented in Table 4. We chose household spending on food eaten at home for expenditure variable and household net income from labour reported in tax returns for income variable. The last crucial variable is the ordinal variable classifying households to public sector employee households, private sector employee households and self-employed households. We included a relatively large number of control variables to the model that were significant in the first or second stage of the 2SLS regression with education and contract term of primary earner as instruments of labour net income (see the technical appendix for the reasons of using 2SLS). It should be noted that control variables are included in order to better explain food expenditure (in the second stage of the 2SLS regression) or income (in the first stage of the 2SLS regression) so estimates of their parameters do not relate to the scale of underreporting. In order to examine the impact of socio-demographic variables on the scale of non-compliance, the interactions of these variables with the classification variable NRA_sectors_3_s should be included in the model. The results for models with such interactions will be presented in section 4.5, however, to estimate the scale of labour income underreporting and PIT and social security contribution gap at the country level, we only need the base PW specification (without interactions), which we present in this section.

Table 4 – Variables included in the final econometric model

group of variables for our analysis	closest group(s) of factors from the literature review	name of the variable	description and source
Expenditure variable in Pissarides-Weber model	Not applicable	log(HBS_expenses_food)	Natural logarithm of household expenses on food eaten at home in constant 2021 prices. Source: HBS
Variable classifying households into fully compliant/ under-reporting groups	Sector and occupation	NRA_sectors_3_s	Ordinal variable classifying households into (1) public sector employee (reference category), (2) private sector employee or (3) self-employed The classification criteria were described in section 4.3 Source: NRA
Income variable in Pissarides-Weber model	Not applicable	log(hsh_NRA_net_income)	Natural logarithm of household net income reported to the NRA in constant 2021 prices. Source: HBS
Control variable	Not applicable	year	Ordinal variable for year: (1) 2017 (reference category), (2) 2018, (3) 2019, (4) 2021 Source: HBS/NRA

Control variable	Not applicable	hsh_primary_earner_sex	Ordinal variable for sex of household primary earner: (1) Female (reference category), (2) Male Source: HBS
Control variable	Not applicable	hsh_primary_earner_age_groups_5	Ordinal variable for age group of household primary earner: (1) 18-34 (reference category), (2) 35-49, (3) 50-64, (4) 65+ Source: HBS
Control variable	Not applicable	children_0_6	Number of children aged 0-6 Source: HBS
Control variable	Not applicable	children_7_12	Number of children aged 7-12 Source: HBS
Control variable	Not applicable	children_13_18	Number of children aged 13-18 Source: HBS
Control variable	Not applicable	settlement_size_agr4	Ordinal variable for settlement size of household: (1) Capital (reference category), (2) Cities over 50 thousand inhabitants, (3) Cities up to 50 thousand inhabitants, (4) Villages Source: HBS
Control variable	Not applicable	unemployment	Binary variable taking the value 1 in households with at least one person who declared unemployment in the HBS and at the same time their reported net income according to the NRA was equal to 0 Source: HBS and NRA
Control variable	Not applicable	disability	Binary variable taking the value 1 in households with at least one person who declared being the disability pensioner in the HBS and at the same time their reported net income according to the NRA was equal to 0 Source: HBS and NRA
Control variable	Not applicable	working_number_HBS	Number of household members who declared in the Household Budget Survey that they were working Source: HBS
Control variable	Not applicable	working_number_NRA	Number of household members who reported positive net income in their tax return Source: NRA
Control variable	Not applicable	housing_ownership	Ordinal variable for housing ownership of household: (1) Own with loan or mortgage (reference category), (2) Own with no loan or mortgage, (3) Rented on a vacant lease, (4) Rented on municipal rent, (5) Used without rent Source: HBS
Control variable	Not applicable	household_members_60plus_share	Share of household members who are at least 60 years old Source: HBS

Control variable	Not applicable	hsh_primary_earner_studying	Binary variable taking the value 1 in households in which the primary earner is currently studying Source: HBS
Control variable	Not applicable	housing_type	Ordinal variable for housing type of household: (1) Apartment (reference category) (2) One-family house, (3) Multi-family house, (4) Other Source: HBS
Control variable	Not applicable	household_size	Ordinal variable for number of members of a household: (1) 1 (reference category) (2) 2, […], (11) 11 Source: HBS
Control variable – instrument for income	Not applicable	hsh_primary_earner_education_agr	Ordinal variable for completed education of a household primary earner: (1) No education (reference category) (2) Primary, (3) Secondary, (4) High-school diploma, (4) Post-secondary vocationally training or Bachelor, (5) Master or Ph.D Source: HBS
Control variable – instrument for income	Not applicable	hsh_primary_earner_contract_term	Ordinal variable for contract term of a household primary earner: (1) Permanent (reference category) (2) Temporary, (3) Reported not working in Household Budget Survey Source: HBS

Source: EY.

Table 5 presents the results of the final model (the first column of results) that was estimated on the full sample. In addition, we present the results of the model estimated on the sample restricted to households with two adults as in the original PW framework (2) and the model estimated on the sample restricted to households with at least two adults (3). In line with the literature standard, we summarize the results of the second-stage of 2SLS procedure (expenditure equation), and present average income gaps $\overline{IG}$ that were calculated based on (i) the value of parameter for the income variable (0.159 in the final model), (ii) the values of parameters for dummy variables for sectors (0.048 for households classified as private sector employee households and 0.113 for households classified as self-employed households) and – for the purpose of estimating lower and upper bound of $\overline{IG}$ – (iii) the variances of residuals from the income equations (0.279, 0.396 and 0.878 for households classifies as public sector employee, private sector employee and self-employed, respectively). Under the estimation results we present a table with the interpretation of the key results from the model and equations used for calculation of underreporting parameters $\bar{k}$ and $\overline{IG}$ (see details in the technical appendix).

Table 5 – Results of the PW model: baseline results for the full sample (1) in comparison with the results of models estimated for restricted samples: (2) – households with two adults and (3) – households with at least two adults

	Dependent variable:		
	log(HBS_expenses_food)		
	all households (1)	household adults = 2 (2)	household adults >= 2 (3)
NRA_sectors_3_sPrivate sector employee	0.048*** (0.014)	0.074*** (0.020)	0.046*** (0.016)
NRA_sectors_3_sSelf-employed	0.113*** (0.027)	0.191*** (0.036)	0.138*** (0.030)
log(hsh_NRA_net_income)	0.159*** (0.017)	0.202*** (0.020)	0.140*** (0.018)
year2018	0.015 (0.014)	-0.012 (0.019)	0.006 (0.015)
year2019	0.037*** (0.014)	0.016 (0.020)	0.034** (0.015)
year2021	0.136*** (0.015)	0.100*** (0.021)	0.136*** (0.016)
hsh_primary_earner_sexMale	-0.020* (0.010)	-0.044*** (0.014)	-0.029*** (0.011)
hsh_primary_earner_age_groups_535-49	-0.014 (0.015)	0.036 (0.022)	0.010 (0.016)
hsh_primary_earner_age_groups_550-64	0.001 (0.016)	0.063*** (0.024)	0.038** (0.017)
hsh_primary_earner_age_groups_565+	-0.017 (0.030)	-0.003 (0.045)	0.033 (0.036)
children_0_6	-0.006 (0.016)	-0.099 (0.092)	-0.004 (0.016)
children_7_12	-0.042*** (0.015)	-0.115 (0.092)	-0.045*** (0.015)
children_13_18	-0.010 (0.015)	-0.104 (0.089)	-0.015 (0.015)
settlement_size_agr4Cities over 50 thousand inhabitants	-0.020 (0.015)		0.015 (0.017)
settlement_size_agr4Cities up to 50 thousand inhabitants	-0.083*** (0.017)		-0.056*** (0.019)
settlement_size_agr4Villages	-0.130*** (0.022)		-0.115*** (0.024)
unemployment	-0.084*** (0.021)		-0.097*** (0.021)
disability	-0.076** (0.030)	-0.057 (0.044)	-0.082*** (0.029)
working_number_HBS	0.098*** (0.010)		0.096*** (0.010)
working_number_NRA	-0.088*** (0.013)	-0.088*** (0.022)	-0.080*** (0.014)
housing_ownershipOwn with no loan or mortgage	0.071* (0.040)	0.084* (0.050)	0.066 (0.041)
housing_ownershipRented on a vacant lease	-0.001 (0.048)	0.029 (0.061)	-0.025 (0.052)
housing_ownershipRented on municipal rent	0.058 (0.058)	0.037 (0.077)	0.019 (0.061)
housing_ownershipUsed without rent	0.055 (0.045)	0.134** (0.057)	0.085* (0.047)
household_members_60plus_share	-0.010 (0.019)	-0.091*** (0.027)	-0.028 (0.023)
hsh_primary_earner_studying	0.033 (0.040)	0.122* (0.071)	0.058 (0.044)
housing_typeMulti-family house	0.037** (0.017)	0.007 (0.023)	0.036* (0.019)
housing_typeOne-family house	0.047*** (0.015)	0.008 (0.018)	0.055*** (0.016)
housing_typeOther	0.010 (0.176)	-0.143 (0.357)	-0.238 (0.245)
household_size2	0.267*** (0.017)		
household_size3	0.394*** (0.022)	0.231** (0.091)	0.131*** (0.016)
household_size4	0.511*** (0.031)	0.414** (0.181)	0.252*** (0.024)
household_size5	0.663*** (0.041)	0.614** (0.278)	0.402*** (0.035)
household_size6	0.748*** (0.057)	0.309 (0.462)	0.485*** (0.052)
household_size7	0.849*** (0.086)	0.967 (0.594)	0.592*** (0.081)
household_size8	0.901*** (0.152)		0.635*** (0.147)
household_size9	1.366*** (0.260)		1.125*** (0.254)
household_size10	0.980*** (0.287)		0.723** (0.281)
household_size11	1.643*** (0.400)		1.389*** (0.392)
Constant	6.355*** (0.159)	6.285*** (0.187)	6.758*** (0.173)
Underreporting estimates:			
IG private sector employee			
point estimate	0.260*** (0.075)	0.306*** (0.077)	0.282** (0.094)
upper bound	0.302*** (0.070)	0.360*** (0.071)	0.323*** (0.088)
lower bound	0.215** (0.079)	0.248** (0.083)	0.237** (0.099)
IG self-employed			
point estimate	0.507*** (0.098)	0.610*** (0.088)	0.627*** (0.097)
upper bound	0.635*** (0.072)	0.707*** (0.066)	0.715*** (0.074)
lower bound	0.335** (0.132)	0.481*** (0.117)	0.512*** (0.127)
Variances of residuals from income regression:			
Public sector employee	0.279	0.235	0.234
Private sector employees	0.396	0.397	0.353
Self-employed	0.878	0.808	0.772
2SLS diagnostics (p-value):			
Wald test (H0: weak instruments)	0.0000	0.0000	0.0000
Wu-Hausman test (H0: endogeneity)	0.0000	0.0000	0.0000
Sargan test (H0: valid instruments)	0.1729	0.1348	0.0721
Subsample sizes:			
Public sector employees	969	503	666
Private sector employees	3758	1941	3248
Self-employed	228	122	176
2nd stage diagnostics:			
Observations	4,955	2,566	4,090
R2	0.344	0.163	0.239
Adjusted R2	0.339	0.153	0.232
Residual Std. Error	10.897 (df = 4915)	11.043 (df = 2536)	10.676 (df = 4051)
F Statistic	66.071*** (df = 39; 4915)	17.036*** (df = 29; 2536)	33.425*** (df = 38; 4051)
Note:			*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors in parentheses. Standard errors of IG parameters were estimated using bootstrap method (10000 iterations). P-values marked with asterisks: *p<0.1, ** p<0.05, ***p<0.01. 2SLS estimator – log(hsh_NRA_net_income) treated as endogenous with instrumental variables: hsh_primary_earner_education_agr and hsh_primary_earner_contract_term. Survey weights were used in estimation.

Source: EY.

Table 6 – Interpretation of the crucial results of the final PW model

Name of the variable/parameter	Variable/parameter interpretation
log(hsh_NRA_net_income)	An increase in household reported net income from labour by 1% leads, other things equal, to an increase in household expenses on food eaten at home by around 0.16%, on average
NRA_sectors_3_s: Private sector employee	Households classified as private sector employee households spend around 4.8% more on food eaten at home relative to households classified as public sector employee households with the same reported net income from labour
NRA_sectors_3_s: Self-employed	Households classified as self-employed households spend around 11.3% more on food eaten at home relative to households classified as public sector employee households with the same reported net income from labour
Underreporting parameters: private sector employee	**Households classified as private sector employee households underreport on average between 21.5% (lower PW share) and 30.2% (upper PW share) of their net labour income (point estimate = 26.0%[49]), which does not mean that the same share of such households income is unreported at the level of the whole economy (see section 4.4)** $\bar{k}_0 = \exp\left(\frac{0.048}{0.159}\right) = 1.352$ $\overline{IG}_0 = \frac{1.352 - 1}{1.352} = 0.260$ $\bar{k}_u = \exp\left(\frac{0.048}{0.159} + \frac{1}{2}(0.396 - 0.279)\right) = 1.433$ $\overline{IG}_u = \frac{1.433 - 1}{1.433} = 0.302$ $\bar{k}_l = \exp\left(\frac{0.048}{0.159} - \frac{1}{2}(0.396 - 0.279)\right) = 1.275$ $\overline{IG}_l = \frac{1.275 - 1}{1.275} = 0.215$
Underreporting parameters: self-employed	**Households classified as self-employed households underreport on average between 33.5% (lower PW share) and 63.5% (upper PW share) of their net labour income (point estimate = 50.7%[50]), which does not mean that the same share of such households income is unreported at the level of the whole economy (see section 4.4)** $\bar{k}_0 = \exp\left(\frac{0.113}{0.159}\right) = 2.035$

[49] Once reported, the hidden net income would become gross income, therefore it should be rather compared to gross reported income. Assuming that the employer's total costs are higher by 42.6% than the employee's net income (calculations based on the NRA data for 2021 for private sector employees), the share of unreported income in the total of unreported income and reported gross income including social security contributions paid by the employer would amount to 19.8% based on $\overline{IG}_0$=26.1%.

[50] Once reported, the hidden net income would become gross income, therefore it should be rather compared to gross reported income. Assuming that gross income of self-employed (net income + PIT + social security contributions) is higher by 23.2% than the self-employed net income (calculations based on the NRA data for 2021 for self-employed), the share of unreported income in the total of unreported income and reported gross income would amount to 45.5% based on $\overline{IG}_0$=50.7%.

	$\overline{IG_0} = \frac{2.035 - 1}{2.035} = 0.507$ $\bar{k}_u = \exp\left(\frac{0.113}{0.159} + \frac{1}{2}(0.878 - 0.279)\right) = 2.744$ $\overline{IG_u} = \frac{2.744 - 1}{2.744} = 0.635$ $\bar{k}_l = \exp\left(\frac{0.113}{0.159} - \frac{1}{2}(0.878 - 0.279)\right) = 1.505$ $\overline{IG_l} = \frac{1.505 - 1}{1.505} = 0.335$

Source: EY.

The crucial result from our econometric model is that households classified as private sector employee households and households classified as self-employed households underreport significant shares of their net labour income. The average share of net labour income underreporting for private sector employee households lies between 21.5% (lower PW share) and 30.2% (upper PW share) with point estimate at 26.0%. The average share of net labour income underreporting for self-employed households is higher and lies between 33.5% (lower PW share) and 63.5% (upper PW share) with point estimate at 50.7%. A larger income gap for self-employed than for private sector employees is in line with intuition and previous literature. Self-employed have more opportunities to hide their revenues, e.g. by not registering cash transactions. Meanwhile, hiding income by private sector employees can be associated with so-called "envelope wages", i.e. not registering part of the salary by the employer and handing it over to the employee in cash. Our analysis should also capture other categories of unregistered income for private sector employees, i.e. those earned outside their main place of work (e.g. providing private lessons, housework, childcare, minor repairs, etc.).

Comparing our results to similar analyses in the literature, we see that income gaps estimated from our model are much higher than the only PW analysis result for Bulgaria that we have found (Kukk, Paulus and Staehr 2020[51]). This previous analysis was based on the 2010 European Union Household Budget Survey data. The income gap was estimated for self-employed – the results were not significantly different from zero with lower bound of mean income gap at 7.4% and upper bound at 9.8%. However, despite using the PW model, the data and approach used were significantly different from ours. First, the income variable was based on the survey data (instead of official data from tax returns) which often results in downward bias to the results. Researchers who studied the source of this bias attributed it to (1) the fact that higher average income is reported in the survey than in the tax registers and (2) to the measurement error typical in the survey data that causes so-called attenuation bias, i.e. error term in the independent variable drives the estimated parameter toward zero (see Cabral, Kotsogiannis and Myles, 2019[52]). Second, the reference group in the Kukk et al. study was all employees (instead of public sector employees) so the result concerned only the difference in the scale of underreporting between self-employed

[51] Kukk, M., Paulus, A., & Staehr, K. (2020). Cheating in Europe: underreporting of self-employment income in comparative perspective. International Tax and Public Finance, 27(2), 363-390.
[52] Cabral, A. C. G., Kotsogiannis, C., & Myles, G. (2019). Self-Employment Income Gap in Great Britain: How Much and Who?. *CESifo Economic Studies*, *65*(1), 84-107.

and employees, among whom private sector employees are also non-compliant according to our results.

The analysis that used similar framework to ours, i.e. income data was matched from tax registers and the reference group consisted solely from public sector employee households, was performed for Estonia (Paulus, 2015[53]). Classification of households in this case was based on the sector of the household head. The author estimated average income gap within the bounds of 23.2% and 34.3% for private sector employee households and within the bounds of 56.1% and 78.4% for self-employed households, so the shares of unreported true income were even higher than those calculated from our final model. However, in the analysis for Estonia, the sample was restricted to households with two adults, therefore, the numbers should be compared with our results in the second column of Table 5, to which they are very close.

Restricting the sample to households with two adults (for this model we also limited the number of control variables due to the small number of observations for self-employed) or households with at least two adults does not change the conclusions as to the fact that non-compliance is present among those households, however it affects the estimates of the scale of underreporting. The point estimate of the mean income gap among private-sector employee households amounts to 26.0%, 30.6% and 28.2% for models estimated for (1) full sample, (2) households with two adults and (3) households with at least two adults. The point estimate of the mean income gap among self-employed households amounts to 50.7%, 61.0% and 62.7% depending on the sample used.

Similar conclusion emerges when we change expenditure variable in our model. Table 7 presents the results of final model in comparison with the results of similar models in which the only change in the specification is different selection of expenditure variable. We did not include estimated parameters for control variables in the table but the set of control variables is the same as in the final model already presented in Table 8. The point estimate of the mean income gap among private-sector employee households amounts to 26.0%, 23.1% and 20.4% for models explaining expenditure on (1) food eaten at home, (2) food eaten at home and expenses in restaurants and hotels and (3) total consumption expenditures of households. Similarly, the point estimate of the mean income gap among self-employed households amounts to 50.7%, 55.6% and 43.9% depending on the choice of expenditure variable. It should be noted, however, that the model explaining full consumption expenditures on households does not pass the test for validity of instruments (p-value of Sargan test <0.05) which may bias the results (perhaps some different instruments should be used in this case).

Our choice of the final model resulted from the following reasoning:

1. We chose full sample (no restriction depending on the number of adults) in order to maximize the number of observations in our model and to ensure that the structure of households in the sample is as close as possible to that of all households in Bulgaria

2. We chose expenditures of food eaten at home as this was recommended and used in the original PW model and most often used in subsequent works

It should be noted that the selection of the model has quite a significant impact on the country-level estimates of the total unreported income and lost revenues from personal

[53] Paulus, A. (2015). Income underreporting based on income expenditure gaps: Survey vs tax records (No. 2015-15). ISER Working Paper Series.

income tax and social security contributions. Therefore, other options may be considered by those using the model in the future.

Table 7 – Results of the PW model: baseline results for the model explaining household expenses on food eaten at home (1) in comparison with the results of models estimated for different expenditure variables: (2) – household expenses on food eaten at home and expenses in restaurants and hotels and (3) – total household consumption expenditure

	Dependent variable:		
	log(HBS_expenses_food) (1)	log(HBS_expenses_food_rest_hotels) (2)	log(HBS_expenses_consumption) (3)
NRA_sectors_3_sPrivate sector employee	0.048*** (0.014)	0.066*** (0.016)	0.088*** (0.016)
NRA_sectors_3_sSelf-employed	0.113*** (0.027)	0.203*** (0.029)	0.223*** (0.030)
log(hsh_NRA_net_income)	0.159*** (0.017)	0.251*** (0.019)	0.387*** (0.019)
Underreporting estimates:			
IG private sector employee			
point estimate	0.260*** (0.074)	0.232*** (0.056)	0.204*** (0.043)
upper bound	0.302*** (0.07)	0.276*** (0.04)	0.250*** (0.04)
lower bound	0.215** (0.078)	0.185*** (0.06)	0.156*** (0.045)
IG self-employed			
point estimate	0.507*** (0.098)	0.555*** (0.065)	0.439*** (0.06)
upper bound	0.635*** (0.073)	0.671*** (0.048)	0.584*** (0.045)
lower bound	0.335** (0.133)	0.400*** (0.087)	0.242*** (0.081)
2SLS diagnostics (p-value):			
Wald test (H0: weak instruments)	0.0000	0.0000	0.0000
Wu-Hausman test (H0: endogeneity)	0.0000	0.0000	0.0000
Sargan test (H0: valid instruments)	0.1729	0.3098	0.0426
Subsample sizes:			
Public sector employees	969	969	969
Private sector employees	3758	3758	3758
Self-employed	228	228	228
2nd stage diagnostics:			
Observations	4,955	4,955	4,955
R2	0.344	0.373	0.435
Adjusted R2	0.339	0.368	0.430
Residual Std. Error (df = 4915)	10.897	12.043	12.302
F Statistic (df = 39; 4915)	66.071***	75.054***	97.006***
Note:			*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors in parentheses. Standard errors of IG parameters were estimated using bootstrap method (10000 iterations). P-values marked with asterisks: *p<0.1, ** p<0.05, ***p<0.01. 2SLS estimator – log(hsh_NRA_net_income) treated as endogenous with instrumental variables: hsh_primary_earner_education_agr and hsh_primary_earner_contract_term. Survey weights were used in estimation. Estimates for control variables were omitted from the table to save space.

Source: EY.

4.4 Country-level estimates of unreported income, lost revenues from PIT/social security contributions and related tax gaps

4.4.1 Representativeness of the results from the econometric model for the entire Bulgarian economy

As the model estimated by the PW method is not a macro-level but micro-level class of econometric model, it is necessary to adopt a number of assumptions in order to translate its results to the level of the entire economy. Unfortunately, there are no clear guidelines or accepted standards in the literature on how to do this, as usually the published articles on traces-of-true-income analyses end with recalling the results for underreporting parameters $\overline{k}$ and $\overline{IG}$. In this section, we describe our innovative approach to obtaining country-level results and explain reasoning underlying the assumptions we made.

First, according to our results (point estimates), households classified as private sector employees underreport on average 26.0% of their true income while households classified as self-employed underreport on average 50.7% of their true income. However, the analysis does not give an answer to the question of whether and how the scale of underreporting differs depending on the level of income. Thus, if the relative scale of underreporting is higher among people with lower income (they may have higher incentive to "save" on taxes and social security contributions) than among those with higher income, the use of the average income gaps introduces the upward bias to the scale of unreported income at the level of the entire economy. In addition, while more affluent people also act to decrease their tax liabilities, it may be less related to income underreporting covered by our approach and more to various forms of tax avoidance, sometimes at the edge of the law.

Second, the specification of the econometric model (the use of the natural logarithm of income) does not allow for inclusion of households with zero net labour income in the estimation. Thus, persons who hid all of their income are included in the analysis only if other persons in their household declared positive income in their tax return. Exclusion of some of the informal workers from the analysis introduces the downward bias to the scale of unreported income when translating our results to the level of the entire economy.

Third, we considered whether our sample could represent all households in Bulgaria, i.e. whether the average income gap in our sample is equal to the average income gap in Bulgaria. As discussed in the technical appendix (section A2.2), the average net labour income of individuals in our sample (weighted using survey weights) was lower than the average net labour income in the whole economy (based on the macro-level data provided by the NRA). As underrepresentation of the wealthiest households in our sample is likely to introduce the upward bias to the results from the PW model, we decided that we should not assume that the highest earners in Bulgaria hide their income in the same way as households in the HBS sample. We therefore assumed that macro estimates, wherever possible, should be based on the data available in our sample where the key variable that allows translating the conclusions to the country level is the sum of survey weights that should represent the sum of similar households in Bulgaria. As a result, we multiplied the obtained values of underreported income in each category (self-employed/ private sector employees) by the sum of weights for all households included in a given category (see Table A.3 in the technical appendix). With this transition, on the one hand, we use the sum of all Bulgarian households with people working as private sector employees or self-employed in the calculations, so everyone contributes to the results. On the other hand, in the calculations we use a lower level of average net labour income (average from our sample) than observed on the macro-level, so the wealthiest households that were underrepresented in the survey contribute to the results only to the level of contribution of the households that were well-represented in the survey (in other words, we impose a lower income gap on households not represented in the survey).

Our estimation sample differed from the initial HBS sample of all households with positive net labour income as we excluded 24.1% of households with net labour income lower that other regular income (see Table A.3 in the technical appendix for the comparison). However, we assume that underreporting level for those excluded households with very low labour income is positive and equal to underreporting level estimated from the PW model. This does not have a very large impact on the results as the income from work of these households is very low.

4.4.2 Obtained estimates

Our approach to calculations of country-level estimates of unreported income and related figures is described in section A2.4 of the technical appendix.

Table 8 summarizes obtained macro-level estimates. Again, as our traces-of-true income model was estimated on the pooled sample (4 years joined together), we present averages of macro-level estimates for those 4 years.

Table 8 – Estimated unreported labour revenues and lost PIT and social security contributions

Macro-level estimates	Average for years 2017, 2018, 2019 and 2021
Unreported labour income as % of GDP	6.37%
Unreported labour income of private sector employees as % of GDP	5.36%
Unreported labour income of self-employed as % of GDP	1.01%
Lost PIT revenues as % of GDP	0.54%
PIT gap as % potential PIT revenues	13.8%
Lost revenues from social security contributions as % of GDP	1.71%
Social security contributions gap as % of potential social security contributions revenues	16.5%

Source: EY, Eurostat and NRA (for social security contributions revenues), NSI (for GDP), Ministry of Finance (for PIT revenues)

The unreported labour income was equal to 6.37% of GDP on average in years 2017, 2018, 2019 and 2021. Despite the fact that underreporting share was higher for households classified as self-employed than households classified as private sector employee households, the larger number of households in the latter group resulted in its much higher contribution to this result - our estimate of unreported labour income of private sector employees in relation to GDP is equal to 5.36% compared to 1.01% for self-employed. PIT gap, i.e. the share of lost PIT revenues in relation to theoretical PIT revenues amounted to 13.8% while the gap in revenues from social security contributions was equal to 16.5% (average for years 2017, 2018, 2019 and 2021). Larger gap observed in the case of social security contributions stems from fact that social contributions are deducted from the PIT base. Lost revenues from personal income tax and social security contributions were equal to 0.54% of GDP and 1.71% of GDP, respectively.

It is worth noting that our estimate of the unreported labour income of private sector employees and self-employed equal to 6.37% of GDP is coherent with estimates used in the previous chapter focusing on the shadow economy (unregistered value added). Although the relation between unregistered labour income and unregistered value added of the company is more complicated (see chapter 2.3 for a discussion), the non-monetary and committed components should be the elements of the overall shadow economy that are closely linked to the unregistered labour income. The average value of the non-monetary and committed shadow economy in Bulgaria over the PIT gap model estimation period amounted to 5.1% GDP.

Although economic literature often focuses on the share of unregistered employment in the total number of employees, we can compare our results to such estimates. Since compensation of employees constitute only a part of GDP, the share of unregistered

employment in total employment should be compared to the share of unregistered labour income in the total value of compensation of employees that, according to our estimates, amounted to 14.6% on average over the years 2017, 2018, 2019 and 2021. Such result is consistent with the estimate provided by International Labor Organization indicating that unregistered employment in Bulgaria is approximately 15.9% of total employment[54]. A similar result can also be obtained by comparing the total number of employees from the Labour Force Survey (it should account for all employees) and the official data on employment (only accounting for employees with formal contracts, expressed as full-time equivalents)[55] – the difference between those two sources amounted to 15.7% (average over the PIT gap model estimation period) of total number of employees found in LFS. In general, one could expect that people who fail to register (part) of their income should earn less than people who are fully compliant so our estimate of unregistered labour income should be lower than the share of unregistered employment in total employment.

4.5 Differences in income underreporting between various socio-economic groups

To analyse differences in income underreporting between different socioeconomic groups the standard Pissarides-Webber model must be extended with so called interaction terms. Technical explanation of this approach is included in section A2.5 of the technical appendix.

Based on the results from each model with interaction, we calculated average income gaps (point estimates) for analysed subgroups. In each case the reference group is all public sector employee households. For the convenience of the readers, we have split estimated income gaps for the tested interactions into three separate tables. The first table includes the results from interactions of classification variable with socio-demographic variables (we could not analyse the effect of education as it was used as instrumental variable in our model), the second table includes the results from interactions of classification variable with variables related to economic activity of households, the third one contains the results from interactions of classification variable with year. When interpreting the results, the first step is to check whether the $\overline{IG}$ point estimates are statistically significant, which means they are likely different from zero. Then, one need to look at the estimated value, to determine the average income gap in the analysed group. For example, private sector employee households without children underreport 20.2% of their true income compared to 38.7% in the case of private sector employee households with children, etc. It is important to recognize that these differences may not necessarily be causal in nature and may be influenced by other factors, such as differences in distribution of other variables. Therefore, it is crucial to interpret the results carefully and consider other potential confounding factors that may be affecting the relationship of interest.

Table 9 – Income gaps estimated from the final PW specification extended by the interaction of the classification variable with socio-demographic variables (1 interaction = 1 model)

Variable tested in interaction model	$\overline{IG}$ point estimate	Standard error	Interpre-tation*
Baseline model			
Private.sector.employee	0.260***	0.074	

[54] See: ILO, (2018), Women and men in the informal economy: a statistical picture (third edition) / International Labour Office – Geneva
[55] https://www.nsi.bg/en/content/3953/total (online, accessed: 26.04.2023).

Self.employed	0.507***	0.098	
Children in the household (children_any = 1)			
Private.sector.employee:children any=0	0.201*	0.093	-
Private.sector.employee:children any=1	0.390***	0.112	+
Self.employed:children any=0	0.466**	0.132	=
Self.employed:children any=1	0.599***	0.147	=
Married couple in the household (household_married = 1)			
Private.sector.employee:household married=0	0.216*	0.116	=
Private.sector.employee:household married=1	0.314***	0.095	+
Self.employed:household married=0	0.306	0.227	-
Self.employed:household married=1	0.642***	0.098	+
Settlement size			
Private.sector.employee:settlement size agr4=Capital	0.140	0.229	-
Private.sector.employee:settlement size agr4= Cities.over.50.thousand.inhabitants	0.278**	0.108	=
Private.sector.employee:settlement size agr4= Cities.up.to.50.thousand.inhabitants	0.416***	0.092	+
Private.sector.employee:settlement size agr4=Villages	0.104	0.169	-
Self.employed:settlement size agr4=Capital	0.528*	0.222	=
Self.employed:settlement size agr4=Cities.over.50.thousand.inhabitants	0.508**	0.150	=
Self.employed:settlement size agr4=Cities.up.to.50.thousand.inhabitants	0.518**	0.167	=
Self.employed:settlement size agr4=Villages	0.314	0.376	-
Sex of the household head			
Private.sector.employee:hsh head sex=Female	0.223	0.124	-
Private.sector.employee:hsh head sex=Male	0.272***	0.093	=
Self.employed:hsh head sex=Female	0.362	0.225	-
Self.employed:hsh head sex=Male	0.582***	0.106	=
Sex of the household primary earner			
Private.sector.employee:hsh primary earner sex=Female	0.218*	0.102	=
Private.sector.employee:hsh primary earner sex=Male	0.308**	0.103	=
Self.employed:hsh primary earner sex=Female	0.551***	0.120	=
Self.employed:hsh primary earner sex=Male	0.443**	0.165	=
Age of the household primary earner (3 age groups)			
Private.sector.employee:hsh primary earner age groups=18-39	0.312*	0.153	+
Private.sector.employee:hsh primary earner age groups=40-59	0.291***	0.080	=
Private.sector.employee:hsh primary earner age groups=60+	0.069	0.199	-
Self.employed:hsh primary earner age groups=18-39	0.494*	0.221	=
Self.employed:hsh primary earner age groups=40-59	0.612***	0.098	+
Self.employed:hsh primary earner age groups=60+	0.056	0.399	-
Age of the household primary earner (4 age groups)			

Private.sector.employee:hsh primary earner age groups=18-34	0.373	0.225	-
Private.sector.employee:hsh primary earner age groups=35-49	0.289**	0.110	=
Private.sector.employee:hsh primary earner age groups=50-64	0.241**	0.098	=
Private.sector.employee:hsh primary earner age groups=65+	0.045	0.461	-
Self.employed:hsh primary earner age groups=18-34	0.478	0.409	-
Self.employed:hsh primary earner age groups=35-49	0.637***	0.118	+
Self.employed:hsh primary earner age groups=50-64	0.529***	0.141	=
Self.employed:hsh primary earner age groups=65+	-0.272	0.925	-

Notes: Standard errors in parentheses. Standard errors of IG parameters were estimated using bootstrap method (5000 iterations). P-values of the test against zero marked with asterisks: *p<0.1, ** p<0.05, ***p<0.01.
*The following symbols were used to help interpret the estimates:
"**=**" indicates that the estimate is significant and matches the estimate from the baseline model (+/- 5 percentage points for the private sector employee households; +/- 10 percentage points for self-employed households).
"**+**" indicates that the estimate is significant and higher by more than 5 pp in the case of private sector employee households and by more than 10 pp in the case of self-employed households than the corresponding estimates in the baseline model.
"**-**" indicates that the estimate is not significant (i.e. likely equal to zero = no unreporting) or lower by more than 5 pp in the case of private sector employee households and by more than 10 pp in the case of self-employed households than the corresponding estimates in the baseline model.

Source: EY.

Here are some patterns of underreporting related to socio-demographic characteristics that can be observed based on the results presented in Table 9:

- **Children in the household:** The higher income gap among households with children compared to households without children could be due to a higher level of expenses related to raising children. This may be because parents have more expenses to cover and may feel more pressure to reduce their reported income to minimize their tax burden. Although the difference in income gap can be observed for both self-employed and private sector households, in the case of the former, the effects are not much different from the mean effect.

- **Married couple in the household:** According to our analysis, households with married couples are more prone to underreporting than other households. This is not in line with the review of literature which suggested that married taxpayers are more compliant than others (see section 3.2.7 of the methodological report for this project). It should be noted that the effect of marriage is probably related to the effect of having children in the household hence similar results for those two interactions.

- **Settlement size:** The results shows that the settlement size of the household is to a large extent related to the scale of underreporting of net labour income. In the case of households classified as private sector employee for which underreporting is mostly associated with so-called "envelope wages", the largest income gap was estimated for smaller cities (up to 50 thousand inhabitants). The income gap is also significantly different from zero for cities over 50 thousand inhabitants but not for the capital city. The effect for villages is not significantly different from zero. The results are different for households classified as self-employed which confirms that the incentives and possibilities for underreporting are different between self-employed and private sector employees. In the case of self-employed households, the income gaps are similar for all three class of cities including the capital. Again, the effect for villages is not significantly different from zero. Possibly it can be due to the fact that some people living in the countryside can spend less on food due to their own micro-scale food production.

- **Sex of the household head & sex of the household primary earner:** Examining the gender effect on non-compliance is not straightforward in a household-level model. We tested two specifications: (1) an interaction of classification variable with the sex of household head according to the HBS and (2) an interaction of classification variable with the sex of household primary earner according to the NRA. First, income gaps are significantly different from zero for both genders if we use the sex of the household primary earner but they are not statistically significant for female households if we use the sex of the household head. The results for household head suggest that households with a male as a household head are more likely to underreport their income in line with the findings from the literature suggesting that women are more tax compliant than men (see section 3.2.3 of the methodological report). Similar effect is visible for the self-employed. When considering the sex of the household primary earner (NRA classification), although the differences between sexes are not very large, the results for private sector employees are similar to those observed for the sex of the household head. However, the result for the self-employed is the opposite – it indicates that households in which the woman is the primary earner tend to hide a bigger share of their income. A possible explanation for this effect could be that the male partner in the household earns more but underreports his income and as a result he is not the primary earner according to our classification based on reported net labour income. We conclude that in the case of private sector employees in Bulgaria, men are somewhat less compliant than woman, however, in the case of self-employed, the effect of gender is inconclusive.

- **Age of the household primary earner & age of the household primary earner:** Analysis of the effect of age (we used two variants of age groups for the household primary earner) on the level of income gap suggests again that the reasons why people underreport their income may be different depending on the sector. We focus on the variant with three age groups due to the higher number of observations for each group. In the case of households classified as private sector employee households, income gap is the highest among households with primary earners in the age group 18-39 and decreases with age (income gap is not statistically significant for households with primary earners above 60 years old). This is in line with the literature review that indicated that older generation is more compliant, and that underreporting may be decreasing with age (see section 3.2.2. of the methodological report). Meanwhile, in the case of household classified as self-employed, income gap is the highest for households with primary earners in the middle age (40-59 years old) which is often the age at which the highest income is achieved. The effect is much lower for households with the youngest (18-39 years old) primary earners and not statistically significant for the oldest (60+) primary earners. These results may suggest that the problem of non-reporting among the private sector employees is stronger for people with low incomes, while this is not necessarily the case for self-employed for whom it is relatively easy to hide income.

Table 10 – Income gaps estimated from the final PW specification extended by the interaction of the classification variable with variables related to the economic activity (1 interaction = 1 model)

Variable tested in interaction model	$\overline{IG}$ point estimate	Standard error	Interpre-tation*
Baseline model			
Private.sector.employee	0.260***	0.074	
Self.employed	0.507***	0.098	
Unemployed person in the household (unemployment = 1)			

Private.sector.employee:unemployment=0	0.212**	0.081	=
Private.sector.employee:unemployment=1	0.624***	0.118	+
Self.employed:unemployment=0	0.443***	0.113	=
Self.employed:unemployment=1	0.893***	0.099	+
Industry of household head			
Private.sector.employee:hsh head industry=Agriculture	0.679**	0.228	+
Private.sector.employee:hsh head industry= Industry	0.386**	0.162	+
Private.sector.employee:hsh head industry= Not.working	0.059	0.178	-
Private.sector.employee:hsh head industry= Services	0.264**	0.093	=
Self.employed:hsh head industry=Agriculture	0.028	1.662	-
Self.employed:hsh head industry=Industry	0.325	0.477	-
Self.employed:hsh head industry=Not.working	0.311	0.256	-
Self.employed:hsh head industry=Services	0.626***	0.095	+
Industry of primary earner			
Private.sector.employee:hsh primary earner industry=Agriculture	0.408	0.678	-
Private.sector.employee:hsh primary earner industry=Industry	0.373*	0.175	+
Private.sector.employee:hsh primary earner industry=Not.working	0.269	0.276	-
Private.sector.employee:hsh primary earner industry=Services	0.235**	0.090	=
Self.employed:hsh primary earner industry=Agriculture	0.263	1.962	-
Self.employed:hsh primary earner industry=Industry	-0.172	1.153	-
Self.employed:hsh primary earner industry=Not.working	0.277	0.407	-
Self.employed:hsh primary earner industry=Services	0.650***	0.090	+
Public/private sector employees in the household			
Private sector employee : all income from self or private = 0	0.229**	0.099	=
Private sector employee : all income from self or private = 1	0.266***	0.074	=
Self employed : all income from self = 0	0.495***	0.135	=
Self employed : all income from self = 1	0.514***	0.134	=

Notes: Standard errors in parentheses. Standard errors of IG parameters were estimated using bootstrap method (5000 iterations). P-values marked with asterisks: *p<0.1, ** p<0.05, ***p<0.01.
"=" indicates that the estimate is significant and matches the estimate from the baseline model (+/- 5 percentage points for the private sector employee households; +/- 10 percentage points for self-employed households).
"+" indicates that the estimate is significant and higher by more than 5 pp in the case of private sector employee households and by more than 10 pp in the case of self-employed households than the corresponding estimates in the baseline model.
"-" indicates that the estimate is not significant or lower by more than 5 pp in the case of private sector employee households and by more than 10 pp in the case of self-employed households than the corresponding estimates in the baseline model.

Source: EY.

Based on the results presents in Table 10, we can draw the following conclusions related to differentiation of the scale of non-compliance depending on variables related to economic activity:

- **Unemployed person in the household:** Households with an unemployed person (i.e. a person who declared in the survey to be unemployed and reported no income in his/her tax return) tend to underreport larger share of their income than households without unemployed members, whether or not classified as private sector employee household or self-employed household (the results for self-

employed should be however taken with caution due to small sample size, i.e. less than 40 households with unemployment =1). This could be due to various reasons, such as households with unemployed members having lower income and therefore being more likely to underreport to reduce their tax burden, or unemployed members engaging in informal work that is not reported.

- **Industry of household head and industry of primary earner:** In the case of interaction with **industry** we aggregated NACE categories to Agriculture, Industry and Services due to the small number of observations, especially outside services. In the case of households classified as private sector employee households, the analysis for household heads indicates statistically significant income gaps in each of the three sectors with the highest one for Agriculture and the lowest in Services. When we classify the industry based on the household primary earner instead of the household head, the estimated income gaps are similar, however, the effect is not significant for Agriculture (still, income gap for Industry if higher than for Services). Other conclusions should be drawn from the analysis for the self-employed. Here, the income gap is statistically different from zero only in Services (similar results for household heads and primary earners). However, in the case of self-employed, the sample sizes for Agriculture and Industry sectors are very small. Performing a more detailed analysis would require creating a dataset of more years pooled together to increase the number of observations and preferably mapping information on industry from tax returns as the HBS data may be not accurate (e.g. many individuals are classified as “Not working” even if they reported positive income).

- **Public/private sector employees in the household:** We also investigated whether the scale of underreporting is lower (1) in households with public sector employees in the case of households classified as private sector employees or (2) in households with public and/or private sector employees in the case of households classified as self-employed. It turns out, that the differences are rather small and not far from the mean effects, however, in line with the intuition the scale of underreporting in private sector employee households in which the share of public sector employee income is above zero is somewhat smaller than in households without public sector employees. This may be due to fewer opportunities for hiding income in the former group of households.

Table 11 – Income gaps estimated from the final PW specification extended by the interaction of the classification variable with year

Variable tested in interaction model	$\overline{IG}$ point estimate	Standard error	Interpre-tation*
Baseline model			
Private.sector.employee	0.260***	0.074	
Self.employed	0.507***	0.098	
Year			
Private.sector.employee:year=2017	0.310*	0.141	+
Private.sector.employee:year=2018	0.327**	0.135	+
Private.sector.employee:year=2019	0.215	0.150	-
Private.sector.employee:year=2021	0.185	0.163	-
Self.employed:year=2017	0.476*	0.213	=
Self.employed:year=2018	0.496*	0.223	=
Self.employed:year=2019	0.613**	0.166	+
Self.employed:year=2021	0.452	0.245	-

Notes: Standard errors in parentheses. Standard errors of IG parameters were estimated using bootstrap method (5000 iterations). P-values marked with asterisks: *p<0.1, ** p<0.05, ***p<0.01.
"=" indicates that the estimate is significant and matches the estimate from the baseline model (+/- 5 percentage points for the private sector employee households; +/- 10 percentage points for self-employed households).
+" indicates that the estimate is significant and higher by more than 5 pp in the case of private sector employee households and by more than 10 pp in the case of self-employed households than the corresponding estimates in the baseline model.
"-" indicates that the estimate is not significant or lower by more than 5 pp in the case of private sector employee households and by more than 10 pp in the case of self-employed households than the corresponding estimates in the baseline model.

Source: EY.

Finally, Table 11 summarizes income gaps estimated from model including interaction of classification variable with **year**. In the case of households classified as private sector employee households, income gap is the highest in 2018 and the lowest in 2019 and 2021 when it is also not significantly different from zero. This indicates that the scale of underreporting among private sector employees may have been decreasing in the recent years. However, it would be good to track whether this trend will continue in the coming years, as it may also be due to factors related to the HBS measurement error or the structure of the sample. When it comes to households classified as self-employed, estimated income gaps are very similar in 2017, 2018 and 2021 (although the effect for 2021 is not significantly different than zero) and significantly larger in 2019. As we do not see the reason for such a hike in non-compliance in 2019, it seems to us that it may be related to small sample sizes (52-67 per year). In addition, the largest standard errors for both private sector employee and self-employed households that were observed in 2021 may be related to higher measurement error in the survey carried out during the pandemic. Therefore, we recommend that the PW analysis for Bulgaria relying on our approach should be performed on a pooled sample (at least 3 years) to increasing probability of obtaining reliable results.

5. VAT gap

In this chapter we discuss our analysis of the VAT gap. The chapter does not include a section on the main idea and background of the method since our approach is based on quite standard econometric approach. The innovation of our analysis consists in the use of unique data that approximates sectoral VAT gaps, which is described in the section on our dataset. Next, we discuss our econometric model(s) and identification of key factors. Finally, we present the translation of obtained econometric results into our VAT gap estimates (another contribution of this part of the research).

Section A3 of the technical appendix includes the detailed list of variables considered in our VAT gap models, data preparation process and various methodological details of VAT gap analysis (often names of sections in the technical appendix correspond to related parts of the main report).

5.1 Dataset and considered factors

In this section we summarize key information on the prepared dataset and factors that we have considered.

- **Type of data:** The data consists of various sectors in Bulgaria observed over different years (panel dataset). We analysed data for 84 sectors (on account of data gaps and other issues the number of sectors in the final model is equal to 57). Due to the availability of the NRA data on VAT revenues, we covered the 2014-2020/2021 period.[56]
- **Reasons for sectoral analysis:** The first reason was the availability of data. Since the NRA could only share with us detailed data on VAT revenues for Bulgaria, an international data analysis, as in the case of the currency demand model for the shadow economy, was not possible. Analysis for the country-level data for Bulgaria only was theoretically feasible but the low number of observations in such approach would limit the scope and quality of our investigation. On the other extreme, individual-level data for VAT taxpayers in Bulgaria could be difficult and time-consuming to obtain. Second, sectoral data allowed us to test the impact of various industrial characteristics and some external factors that were important in this research. Third, having conducted other analyses in this study at the international and individual level, we believed that the analysis at the level of sectors could generate most additional insights.
- **Data sources:** Our VAT gap analysis would not be possible without various sectoral data shared by the NRA, especially in the area of VAT revenues and different characteristics of businesses. This dataset was supplemented by publicly available sources with industrial, macroeconomic, institutional and sociodemographic data, including Eurostat, European Commission, European Central Bank, International Monetary Fund, World Bank, Organisation for Economic Cooperation and Development, Fraser Institute, United Nations and National Statistical Institute. Some parts of our analyses also benefited from the information regarding VAT regulations in Bulgaria (especially VAT rates) collected mainly by the local EY office.
- **Explained variables:** To analyse the VAT gap at the sectoral level, one needs explained variable(s) (or so-called indicators in the MIMIC model framework) that

[56] 2021 period was not available for all the considered variables. Yet, under certain assumptions, the estimated econometric model (see further) can be used to provide estimates or scenarios of the VAT gap also for 2021 and the following years.

to the possibly largest extent capture (indicate) the scale of the VAT gap in sectors and over time. In theory, the sectoral VAT compliance gap variable should have the following form[57]:

$$VAT\ compliance\ gap\ (\%) = \frac{potential\ VAT\ -\ declared\ VAT}{potential\ VAT}\ (*\ 100\%)$$

where $potential\ VAT$ is the value of VAT that would be declared (or collected) under the hypothetical scenario of perfect compliance with tax regulations, while $declared\ VAT$ is the value of declared VAT in tax returns available directly from the NRA. Having analysed strengths and weaknesses of various data series, we came up with two variables that try to approximate the concept from the formula above. Theoretically, they values should be within the 0-100% range. Yet, due to various inaccuracies in the actual data points and simplifications in the applied approach, they often obtained also lower or higher values. Therefore, such variables should rather only be interpreted in relative terms (whether the value in sector X in year T is higher than in sector Y in year T and in sector X in other years), not as precise measures of the scale of VAT gap in the given sector and year. Below we describe the two considered variables.

1. Output VAT gap based on potential VAT estimate (variable name: $output\_VAT\_gap$)

$$\frac{potential\ output\ VAT\ estimate - declared\ output\ VAT}{potential\ output\ VAT\ estimate}$$

It was our main explained variable. We focused on output VAT for two reasons. First, it was easier to approximate $potential\ output\ VAT\ estimate$ than $potential\ input\ VAT\ estimate$, since the latter for the given sector depends strongly on the industries and locations (including abroad) of its suppliers and corresponding VAT regulations for such transactions (i.e. requires more data points and assumptions). Second, we did it since the second explained variable (see below) covers mostly input VAT irregularities and we wanted to have more complete picture. We calculated $potential\ output\ VAT\ estimate$ based on its three components:

$$\begin{aligned} &potential\ output\ VAT\ estimate \\ &= potential\ output\ VAT\ on\ sales\ estimate \\ &+ potential\ output\ VAT\ on\ intracommunity\ acquisitions\ estimate \\ &+ potential\ output\ VAT\ on\ import\ of\ services\ outside\ the\ EU\ estimate \end{aligned}$$

The latter two components are related to the fact that for most intracommunity acquisitions and import of services Bulgarian businesses are required to apply the reverse charge mechanism, i.e., to report both output VAT and input VAT simultaneously (instead of having the output VAT calculated by the supplier and using it as their input VAT for the purpose of VAT returns). We approximated the components of $potential\ output\ VAT\ estimate$ using the similar approach as in the methodology developed by the International Monetary Fund[58]:

[57] All formulas in this section should be read for each sector and year separately, subscripts have been omitted for the simplicity.
[58] Hutton (2017), The Revenue Administration–Gap Analysis Program: Model and Methodology for Value-Added Tax Gap Estimation, IMF, Technical Notes and Manuals No. 2017/004.

$$potential\ output\ VAT\ on\ sales\ etimate = output\ from\ national\ accounts * (1 - share\ of\ exports\ outside\ the\ EU\ and\ intracommunity\ supplies\ in\ output) * average\ VAT\ rate$$

$$potential\ output\ VAT\ on\ intracommunity\ acquisitions + potential\ VAT\ on\ import\ of\ services\ outside\ the\ EU\ estimate = (value\ of\ intracommunity\ acquisitions\ estimate + value\ of\ import\ of\ services\ outside\ the\ EU\ estimate) * average\ VAT\ rate\ on\ imports$$

In the perfect world, $output\ from\ national\ accounts$ should be the most complete measure of the scale of production (≈ sales) in the given sector, often including adjustments to capture the scale of non-observed economy.[59] [60] The adjustments for exports and intracommunity supplies were made to exclude transactions for which VAT rate was equal to zero or not applicable. Average VAT rate was determined for the given sector and year based on the information from tax regulations.[61] All components in the parentheses above were estimated with the use of OECD international input-tables. This data source was also used for calculating the role of different industries in the given sector imports to calculate the average VAT rate on imports.

Advantage of this variable is that under the assumption of completeness and correctness of data in national accounts and input-output tables it should allow to approximate the total value of the sectoral output VAT gap, including its different sources (shadow economy, tax frauds, etc.). The main drawback of the measure is the fact that the mentioned assumptions are often violated to unknown extent and there are also various inconsistencies in measurement and definitions between the different used data sources (e.g. statistical office may not accurately estimate the scale of non-observed economic activity, definitions of transactions and revenues in national accounts may be different than for the purpose of VAT regulations, assignment of companies to sectors may be somewhat different in national accounts and tax office data, etc.). In addition, this variable could also cover some aspects of the policy VAT gap (i.e. lower VAT revenues stemming not from compliance issues but various regulations and exemptions that lower such tax collections). As a result, the sectoral VAT gap estimates obtained with the described approach and various data sources have sometimes not intuitive (including negative) values for some sectors. Therefore, this variable should be rather treated as an index of the sectoral VAT gap which may suggest in which sectors and time periods the VAT gap was relatively high (not as a precise measure of the scale of the sectoral VAT gap).

2. <u>(Output and input) VAT gap based on VAT audits (variable name: *vat_gap_audit)*</u>

[59] For some specific sectors, for which output from national accounts is not an approximation of turnover (e.g. trade sectors in which it covers only so called trade margins), we substituted output with best available estimates of turnover. When possible, we also disaggregated output available for aggregates of the considered sectors to the smaller sectors of our interest.

[60] In theory, the formula above may also include some additional adjustments. One of them is related to the share of companies operating below VAT threshold in total revenues of different sectors. Due to the low availability of precise data in this area and the fact that the value of such threshold in Bulgaria was low, we resigned from such correction. Another potential adjustment is related to share of companies' revenues that are VAT exempt in total revenues for different sectors. Again, the issue was the missing high-quality data to introduce such correction. Yet, we believe that the two missing adjustments, due to the limited role of such issues in the Bulgarian VAT system, should not have a large impact on the obtained results.

[61] (Weighted) average was applied when there was more than one VAT rate in the sector.

$$\frac{after\ audit\ VAT\ -\ declared\ VAT}{after\ audit\ VAT}$$

Such variable is constructed with the use of data for businesses that were subject to VAT audits. Since we have been informed by the NRA that the value of the numerator in the ratio above stems to large extent (but not only) from the wrongly declared input VAT, we interpret this variable as a VAT gap measure of both output and input VAT but with greater emphasis on the latter. Advantage of this variable is that it is based on missing VAT estimates from actual audits. Large drawback is the fact that due to non-random selection (targeting) of companies for audits, potential differences in approach and effectiveness of audits between sectors and over time, such VAT gap measure could be significantly biased in direction that is difficult to evaluate. In addition, likely not all kinds of missing output and input VAT could be identified during tax audits. As a result, we treat this variable as a less reliable one in our analyses.

- **Explanatory variables.** Since our explained variables try to directly capture the relative scale of VAT gap in different sectors and over time, all explanatory variables included in the econometric model could be interpreted as determinants (or causes in the MIMIC model framework) of the VAT gap. In other words, there are no additional control variables (variables related to other issues than tax non-compliance) as was the case in the shadow economy and PIT gap analysis. Naturally, considered determinants (factors) could be assigned to different groups (e.g. business form, financial conditional of taxpayer, etc.). It is also worth noting that while most considered explanatory variables were at the sectoral level, some other were available only at the country level (e.g. unemployment rate). Reasonable candidates for variables from the latter category are the ones that could have a similar, material impact on VAT non-compliance in different sectors. In addition, having only a few years of data in the sample, it is difficult to accurately estimate the impact of such factors, so it is good to have some economic theory supporting their potential effects. Finally, such variables should not dominate the list of variables included in the final econometric model, which in general should be based on the sectoral data. Therefore, the preferred variables from this category were key macroeconomic indicators, especially the ones for which some external forecasts are relatively easy to obtain (to analyse future scenarios of the VAT gap).
- **Alternative variables:** For different areas often more than one variable (source) was considered. The final selection was based on the number of observations and empirical analysis.
- **Initial exclusions from the analysis**: Most often we excluded variables due to data gaps.
- **Consultations**: At the request of the NRA, after they saw the first proposition of our dataset, we considered several additional variables. They were mostly sociodemographic variables, some of them with a less direct theoretical link with the VAT gap. Most of them were available at country level.

Details about our dataset could be found in section A3.1 of the technical appendix. They contain information about to which group a given variable belongs and its closest group from the literature review. They also cover variables description and data sources. You can also find there an explained decision about excluding some variables already at the initial phase of the analysis, numbers of observations, sectors, and years available. We also included additional comments, among other to address the NRA's request to link some of our macroeconomic variables with publicly available forecasts (e.g. from the IMF).

5.2 Econometric model and identification of key factors

To our best knowledge, econometric investigation of VAT gap at the sectoral level was not earlier done by other researchers and described in the literature.

Before the project started we assumed that we would use the MIMIC (multiple indicator multiple cause) model for the sectoral analysis of the VAT gap in Bulgaria. Yet, after our assessment of the actual data, we concluded that this approach should not be followed.

The main idea for applying the MIMIC model for an analysis of VAT non-compliance is the setup in which the scale of the VAT gap in sectors is not directly observable (latent variable) but there are various (more than one) indicators of this issue. Such indicators should be strongly correlated both with the underlying latent variable (this cannot be tested) as well as with each other (this could be verified), since, despite potential measurement errors and other inaccuracies, they try to capture the same phenomenon. As described in section 5.1, we identified two main indicators of the sectoral VAT gap in Bulgaria with the use of the obtained dataset: (1) output VAT gap based on the estimate of potential VAT and (2) (output and input) VAT gap based on VAT audits. Yet, our analysis of the actual data showed that the correlation coefficient between the two variables in the analysed sample is close to zero. In addition, as discussed in section 5.1, there are good reasons to believe that the two indicators measure somewhat different aspects of the VAT gap as well as to suspect that the second variable could be more biased. As the result, such explanatory variables should not be analysed together within the MIMIC framework. Instead of this we decided to investigate their determinants separately with different panel econometric models. We chose the first variable as the explained variable of our main interest, but we also show some results for the second indicator.

5.2.1 Model of output VAT gap based on potential VAT estimate

Technical discussion of the model selection is included in section A3.3 of the technical appendix. We tried various specifications and the final model consists of 5 independent variables (and, depending on the method, additional 56 dummy variables capturing sectors' individual effects). Description of the variables included in the model can be found in Table 12.

Table 12 – Variables included in the final econometric model of the output VAT gap

group of variables for our analysis	closest group(s) of factors from the literature review in the methodological report	name of the variable	description
Dependent (explained) variable	Not applicable	vat_gap_output [62]	Output VAT gap represented by the ratio of the difference between the potential output VAT estimate and the declared output VAT (nominator) to the potential output VAT estimate (denominator), %.[63]
Cause: firm size	Business form	vat_base_micro_firms_share	Share of micro firms in total VAT base (value of all made deliveries of goods and services), %.
Cause: business bankruptcies and births	Business form / financial conditions of taxpayers / shock to financial condition	firms_death_rate	Enterprise death rate obtained by dividing the number of enterprise deaths by the number of active enterprises, %.
Cause: productivity / complexity of sector's products and services	Business form	labour_prod	Labour productivity obtained by dividing gross value added (chain linked volumes, 2015) by total employment, in constant thousand BGN.
Cause: type of clients	Business form	firms_b2g_rev_share	Share of firms' revenues coming from sales to government, %.
Cause: economic or financial situation		unem	Unemployment rate, % of total labor force (economically active population).

Source: EY.

[62] In the database this variable is called *vat_prod_gap_o_xicas*. The difference between the names used in the theoretical and econometric part of the report stem from the fact that we kept several measures of the VAT gap in the database.
[63] See also section 5.1 for more detailed discussion.

Table 13 – Coefficients in the final econometric model(s) of the output VAT gap

Dependent variable: vat_prod_gap_o_xicas

	FE	RE	FGLS_no	FGLS_ar1	PCSE_ar1	PCSE_psar1
vat_base_micro_firms_share	0.9366***	0.9592***	0.5522***	0.3515**	0.7241***	0.9638***
	(0.323)	(0.231)	(0.165)	(0.161)	(0.254)	(0.217)
firms_death_rate	0.6459**	0.5842**	0.3172**	0.1746	0.2193	0.0634
	(0.290)	(0.288)	(0.150)	(0.132)	(0.383)	(0.351)
labour_prod	0.1022**	0.0879**	0.1569**	0.1721***	0.1273	0.1418
	(0.046)	(0.045)	(0.061)	(0.063)	(0.116)	(0.098)
firms_b2g_rev_share	-0.8695	-0.8406*	-0.9373***	-0.7042***	-0.3843	-0.5647
	(0.586)	(0.479)	(0.234)	(0.239)	(0.611)	(0.515)
unem	1.2952***	1.2677***	0.8764***	0.7628***	1.0146***	0.9501***
	(0.284)	(0.280)	(0.103)	(0.108)	(0.304)	(0.265)
constant	-55.7825***	-25.3191***	-49.7541***	-47.6410***	-50.2932***	-49.5295***
	(6.042)	(7.503)	(2.914)	(2.980)	(6.280)	(6.174)
Observations	399	399	399	399	399	399
Groups		57	57	57	57	57

Notes: Standard errors in parentheses. P-values marked with asterisks: *p<0.1, ** p<0.05, ***p<0.01. Groups = number of sectors included in the sample. Individual dummies (for each sector) are not shown in the table for clarity. *vat_prod_gap_o_xicas* = *vat_gap_output*.

Source: EY.

The results of our estimations with various methods are shown in Table 13[64]. Section A3.3 of the technical appendix includes the discussion why we chose as final model the FGLS with heteroskedastic error structure (FGLS_no) that takes the form of the following equation:

$$\widehat{vat\_gap\_output}_{s,t} = 0.5522 * vat\_base\_micro\_firms\_share_{s,t} + 0.3172 * firms\_death\_rate_{s,t} + 0.1569 * labour\_prod_{s,t} - 0.9373 * firms\_b2g\_rev\_share_{s,t} + 0.8764 * unem_t + (individual\_effect_s)$$

where *s* denotes sector and *t* is time subscript. The variables with both *s* and *t* subscript are differentiated across sectors and change over time (e.g. share of micro firms in the VAT base), unemployment varies only in time and individual effects are constant in time but different for each sector. The individual effect is the sum of individual dummies and the constant that is common for all sectors (-49.7541).[65] It is worth noting that further, when calculating the theoretical values of VAT gap index based on this variable, we omit the constant and fixed effects. The reason for this is that we suspect that in relatively many cases they account for the fixed in time sector-specific inaccuracies in measurement of the VAT gap rather than fixed in time sector-specific VAT non-compliance. Alternative approach in this area would impact some of our results.

The estimated coefficients in the econometric model should be interpreted in the following way (and under condition of *ceteris paribus* - holding all other factors fixed):

- **Share of micro firms in the VAT base** ($vat\_base\_micro\_firms\_share$)**.** An increase in the share of micro firms in the VAT base by 1 pp. is associated on average with 0.55 pp increase in the output VAT gap.
- **Firm death rate** ($firms\_death\_rate$). An increase in the rate of firm deaths by 1 pp. on average leads to an increase in the output VAT gap by 0.32 pp.

[64] The table with all the estimated coefficients (including individual effects) for the preferred model (FGLS_no) are shown in section A3.4 of the technical appendix.
[65] Note that the individual effect of the first sector in the panel (C10) is equal the constant. This is because the dummy variable for the first sector is omitted due to perfect multicollinearity.

- **Labour productivity** ($labour\_prod$). An increase in the labour productivity by 1 thousand BGN per employee is associated on average with 0.16 pp increase in the output VAT gap.

- **Share of revenues from sales to the government** ($firms\_b2g\_rev\_share$). An increase in the share of the government in sector's revenues by 1 pp. on average leads to a decrease in the output VAT gap by 0.94 pp.

- **Unemployment** ($unem$). An increase in the unemployment rate in Bulgaria by 1 pp. on average leads to an increase in the output VAT gap by 0.88 pp.

- **Individual effect** ($individual\_effect$). Individual effects vary across sectors and for some are positive, while negative for the others. An average individual effect for all the estimated sectors equals to -15.01 and the median equals to -7.66 (for all the estimated individual effects, see the chart in section A3.4 of the technical appendix). Notably, almost all the individual effects (51 dummies and the constant) are statistically significant[66].

Yet, one should remember that our explained variable measures the output gap with a significant inaccuracy and that its variation is likely quite different than the variation of the true sectoral VAT gap (as % of potential VAT). Therefore, the mentioned changes in percentage points of output VAT gap should be treated as indicative only. To correct for this issue when estimating the VAT gap at the country and sector level, we link our estimates with existing estimates of the total VAT gap in Bulgaria (see section 5.3).

In general, the relationships between the explanatory and explained variables established in the final model match both economic theory and intuition. The VAT gap is greater in the sectors that are characterised by (1) greater role of micro enterprises (share of micro firms in the VAT base), (2) greater relative number of bankruptcies (firm death rate), and (3) when the general economic situation in the country worsens (unemployment rate). On the contrary, the greater the role of business to government transactions, the smaller the sector's VAT gap. Although somewhat counterintuitive, greater labour productivity has positive (in the statistical sense) impact on the VAT gap. One of the reasons for such relation could be the fact that enterprises with a complex production process (that are typically more productive) have more opportunities for VAT frauds.

Section A3.3 of the technical appendix also describes our analysis of robustness of the considered econometric model, concluding that it is quite robust to various changes in sample and specification.

5.2.2 Model of output and input VAT gap based on VAT audits

In the second model, the dependent variable is the (output and input) VAT gap based on VAT audits (vat_gap_audit). Although the calculations of this variable were straightforward and based on one source only (NRA), we found a large number of outliers that could sabotage our estimates.[67] Therefore we conducted additional cleaning of the dataset and removed observations with a very small number of audited firms (the minimum reliability threshold of 10 audited firms) or a very high value of VAT

[66] The few exceptions are: Manufacture of coke and refined petroleum products (C19), Remediation activities and other waste management services (E39), Retail trade, except of motor vehicles and motorcycles (G47), Food and beverage service activities (I56) and Architectural and engineering activities (M71).
[67] There were 10 observations with a VAT gap exceeding 1000%, and the maximum reached astronomical 86140%. Such a large VAT gap was usually the case when firms declared negative VAT difference while during the audit it was found that they should have been net VAT payers.

gap (the maximum threshold of 400%, encompassing top 15 observations). Next, we linearly interpolated the removed observations. After the cleaning we were left with 50 sectors and 323 observations in total.

Prior to moving to the estimation, we need to recall our concerns about the dependent variable in the model. In general, the quality of econometric results relies, among others, on the assumption that controlled units in the sample are selected at random.[68] Violation of this principle would be a source of bias (estimates would deviate from the true parameters). Given that the NRA conducts tax controls based on certain targeting, the estimates of this model may suffer from such issue.

The strategy to find the preferred specification of this model (FGLS_no) was similar to the one adopted in the previous identification process and is described in section A3.3 of the technical appendix.

Table 14 summarizes the variables that enter the final model and the estimates of coefficients are shown in Table 15.

Table 14 – Variables included in the final econometric model of the VAT gap based on audits

group of variables for our analysis	closest group(s) of factors from the literature review in the methodological report	name of the variable	description
Dependent (explained) variable	Not applicable	vat_gap_audit	Ratio of additional VAT obligation established in audit to total VAT obligation (additional VAT + VAT declared by audited liable persons), %.
Cause: self-employment / sole trader	Business form	self_empl_share	Share of self-employed in total employment (domestic concept), %.[69]
Cause: business bankruptcies and births	Business form / financial conditions of taxpayers / shock to financial condition	firms_death_rate	Enterprise death rate obtained by dividing the number of enterprise deaths by the number of active enterprises, %.
Cause: role of foreign capital	Business form	gva_foreign	Share of value added at factor costs generated by foreign-controlled companies, %.

[68] In econometrics this is an assumption of random errors (errors have zero mean).
[69] Domestic concept refers to employment in resident production units irrespective of the place of residence of the employed person. This approach is typical for national accounts data.

Cause: role of government		gov_effectiveness	The value of the indicator measuring the government effectiveness from the Worldwide Governance Indicators. It ranges from approximately -2.5 (low government effectiveness) to 2.5 (high government effectiveness). It reflects perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies.
Cause: economic or financial situation		unem	Unemployment rate, % of total labor force (economically active population).

Source: EY.

Table 15 – Coefficients in the final econometric model of the VAT gap based on audits

Dependent variable: vat_gap_audit	
	FGLS_no
self_empl_share	-0.8718**
	(0.443)
firms_death_rate	0.6409*
	(0.372)
gva_foreign	0.5586**
	(0.224)
unem	0.7812**
	(0.313)
gov_effectiveness	-14.1788**
	(5.515)
constant	89.0900***
	(16.150)
Observations	296
Groups	47

Notes: Standard errors in parentheses. P-values marked with asterisks: *p<0.1, ** p<0.05, ***p<0.01. Groups = number of sectors included in the sample.

Source: EY.

Our final model is represented by the equation:

$$\widehat{vat\_gap\_audit}_{s,t} = -0.8718 * self\_empl\_share_{s,t} + 0.6409 * firms\_death\_rate_{s,t} + 0.5586 * gva\_foreign_{s,t} + 0.7812 * unem_t - 14.1788 * gov\_effectiveness_t + individual\_effect_s$$

where *s* denotes sector and *t* is time subscript.

The estimated coefficients in the econometric model should be interpreted in the following way (and under condition of *ceteris paribus* - holding all other factors fixed):

- **Share of self-employed** ($self\_empl\_share$)**.** An increase in the share of self-employed in the total employment by 1 pp. is associated on average with 0.87 pp. decrease in the VAT gap.
- **Firm death rate** ($firms\_death\_rate$). An increase in the rate of firm deaths by 1 pp. on average leads to an increase in the VAT gap by 0.64 pp.
- **Share of value added generated by foreign-controlled companies** ($gva\_foreign$). An increase in the share of foreign-controlled companies in generating value added by 1 pp. is associated on average with 0.56 pp. increase in the VAT gap.
- **Unemployment** ($unem$). An increase in the unemployment rate in Bulgaria by 1 pp. on average leads to an increase in the VAT gap by 0.78 pp.
- **Government effectiveness** ($gov\_effectiveness$). An increase in the government effectiveness by 1 point on average leads to a decrease in the VAT gap by 14.18 pp.[70]
- **Individual effect** ($individual\_effect$). For the majority of sectors, individual effects are not statistically significant meaning that they are not statistically different from 0.[71]

The interpretation of these outcomes is the most informative when compared to the results of the first model. The common conclusion for both models is that the VAT gap increases due to increases in the firm death rate and the unemployment rate, suggesting that bankruptcy risk and business cycle are important drivers of VAT gap. We find that the role of government has negative impact on the VAT gap (represented by the share of business to government transactions in the first model and the index of government effectiveness in the second one). When looking at differences, in the model with VAT gap based on audit controls, the share of self-employed is statistically significant and has negative impact on the dependent variable (which is somewhat contradictory to the finding of positive impact of micro firms in the first model). A novelty of the second model is that sectors with greater role of foreign companies are found to be generating greater VAT gap. The results for the role of self-employed and foreign companies are somewhat counterintuitive and we should reconsider the issue of non-randomness of the sample that may cause biases in the estimators. In other words, if the tax office targets (i.a. for efficiency reasons) larger firms (in this case firms with more employees) and foreign firms then these characteristics are biased and the respective coefficients should not be considered reliable. Alternatively, it may be driven by some specific characteristics of input VAT gap (captured only in the second model) in contrast to output VAT gap (captured in both models). For example, while the latter could be to large extent related to shadow economy transactions, the former could be more linked with other sources of the VAT gap, e.g. tax frauds or evasion, which may be more prevalent among larger and foreign-controlled companies.

[70] Due to the scale of the government effectiveness indicator an increase by 1 point is very large and unlikely in the short-term. It may be better to consider an increase by 0.1 point that on average leads to a decrease in the VAT gap by 1.41 pp.
[71] However, the individual dummy variables (for each sector) are jointly statistically significant which means that they should be included in the model.

In the next step, we tested the robustness of the model. Such analysis is included in section A3.3 of the technical appendix. To summarise, our specification for this model becomes inappropriate when put to tests.

Given that the model of (output and input) VAT gap based on VAT audits is threatened by a bias and fails the stability test, the estimated parameters should be interpreted very carefully. In our view, the model is not reliable and does not succeed in estimating the true parameters. However, it can be of some value when treated as a model supplementary to the one based on the potential VAT estimates. Notably, the second model confirms several important conclusions of the main (first) model, namely the rate of firm deaths and the unemployment rate are relevant factors that cause VAT gap while the government (through direct transaction or general effectiveness) may reduce the gap.

5.3 VAT gap estimates

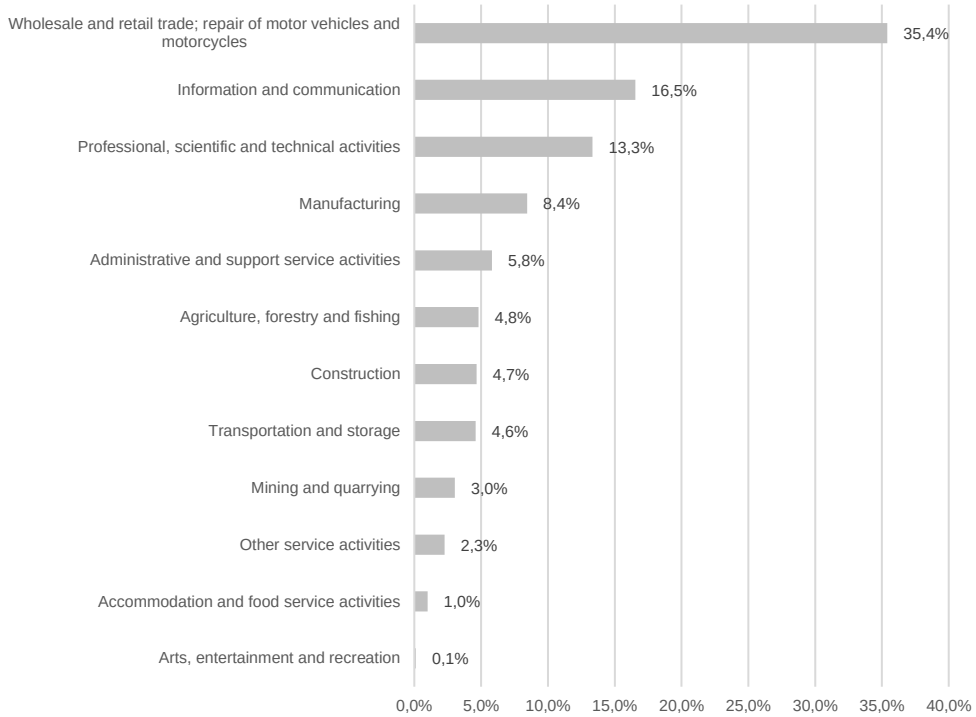
We decided to use the output VAT gap model based on potential VAT estimate (instead of the model based on VAT audits) to evaluate the sectoral level of the VAT gap. First reason was that the related dependent variable is in our opinion less biased in approximating the sectoral VAT gap in Bulgaria. Moreover, the output VAT gap model had significantly better statistical properties.

5.3.1 Contributions of sectors to the overall VAT gap

First, we calculated contributions of sectors to the overall VAT gap. For details, please see section A3.5 of the technical appendix.

Obtained results are presented in Chart 7 (NACE sections) and 8 (NACE divisions). In general, the largest contribution to the overall VAT gap stems from trade (wholesale and retail), whereas more detailed structure indicates that apart from wholesale and retail trade, computer programming and crop and animal production also contribute substantially to the VAT gap. Yet, these results are affected not only by the difference in the role of VAT gap within sectors but also by differences in the role of different sectors in our approximation of potential VAT. For the results that focus on the first aspect see section 5.3.3.

Chart 7 – Contributions of sectors (NACE sections) to the overall VAT gap in Bulgaria in 2021 (% of the total VAT gap)



Source: EY.
Notes: We omitted sectors for which we assumed zero VAT gap.

Chart 8 – Contributions of sectors (NACE divisions) to the overall VAT gap in Bulgaria in 2021 (% of the total VAT gap)

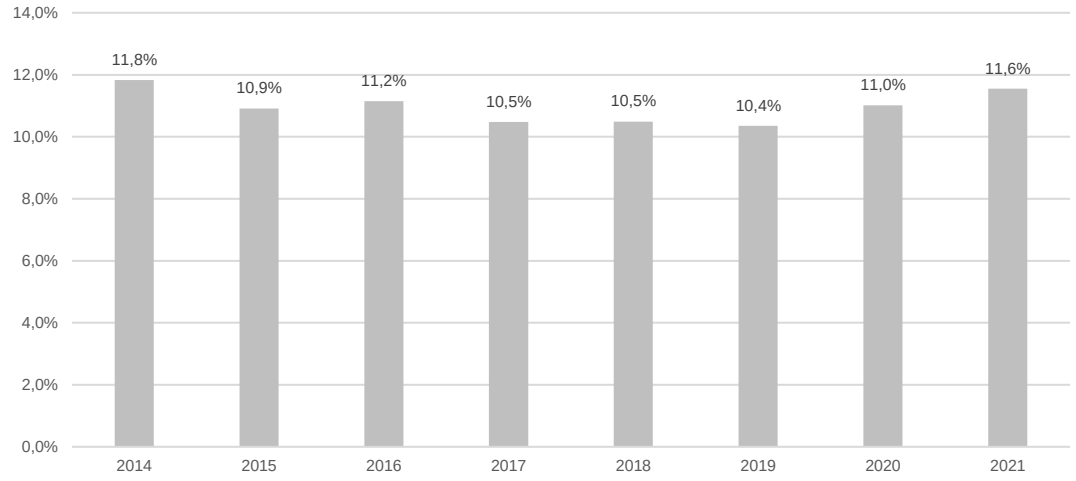
Source: EY.
Notes: We omitted sectors for which we assumed zero VAT gap.

5.3.2 VAT gap on the country level

In the previous section we described our sector contributions to the overall value of the VAT gap. Next, we took the estimates of European Commission (EC)[72] and used them to calibrate our VAT gap estimate at the country level. The EC VAT gap estimates for Bulgaria in 2020 were marked in the EC report with a red dot indicating low reliability of estimates due to unavailability of up-to-date information to conduct the research. As a result, we decided to use in our calibration the average of the EC VAT gap estimates over the 2016-2019 period.[73] Our calibration method is described in section A3.5 of the technical appendix.

According to our estimates, scaled to the 2016-2019 average results of the European Commission's study, the VAT gap (% of potential VAT) in Bulgaria slightly declined between 2014 and 2019 (from 11.8% to 10.4%), with some fluctuations during this period. Yet, in the next two pandemic years the VAT gap increased, reaching 11.6% in 2021.

Chart 9 – Model VAT compliance gap scaled to the European Commission's average VAT gap estimate over the years 2016-2019 (% of potential VAT)

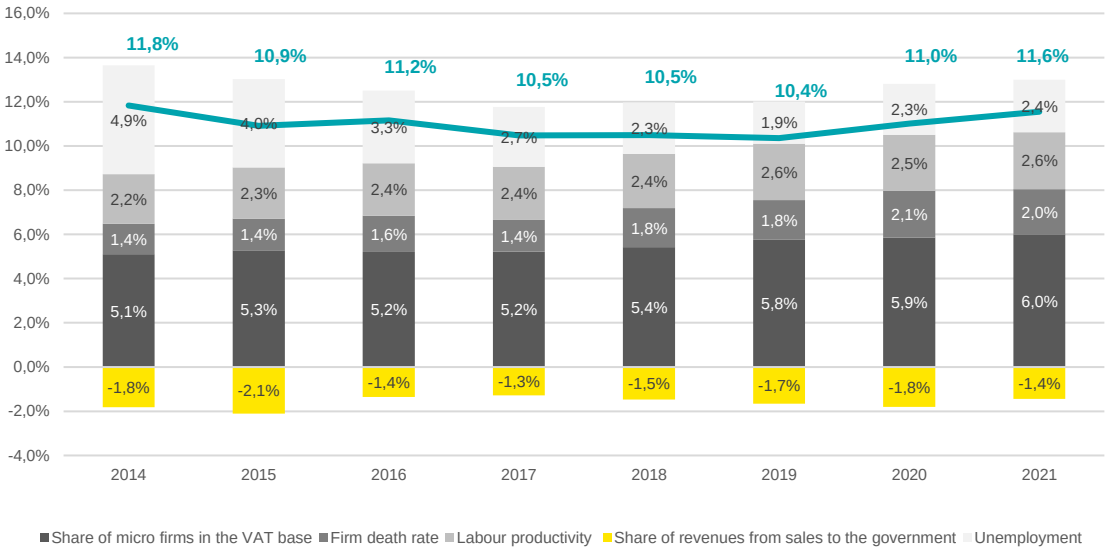


Source: EY.

In addition to this, we also calculated contributions of variables to the overall VAT gap level (see Chart 10). We can observe that share of micro firms in the VAT base has the largest contribution to the VAT gap in most years in the sample, whereas contribution of unemployment is the most volatile. The share of revenues from sales to government has a negative sign in our model which means that it contributes to a decrease in VAT gap in Bulgaria.

[72] https://op.europa.eu/en/publication-detail/-/publication/030df522-7452-11ed-9887-01aa75ed71a1 (online, accessed 18.05.2023).
[73] We excluded the year 2020 because of substantial change in the value that might stem from additional assumptions made in that study due to limitations in data availability. Moreover, the COVID-19 pandemic in 2020 could also distort the results.

Chart 10 – Contributions of variables to the model VAT compliance gap scaled to the European Commission's average VAT gap estimate over the years 2016-2019 (% of potential VAT)



Source: EY.

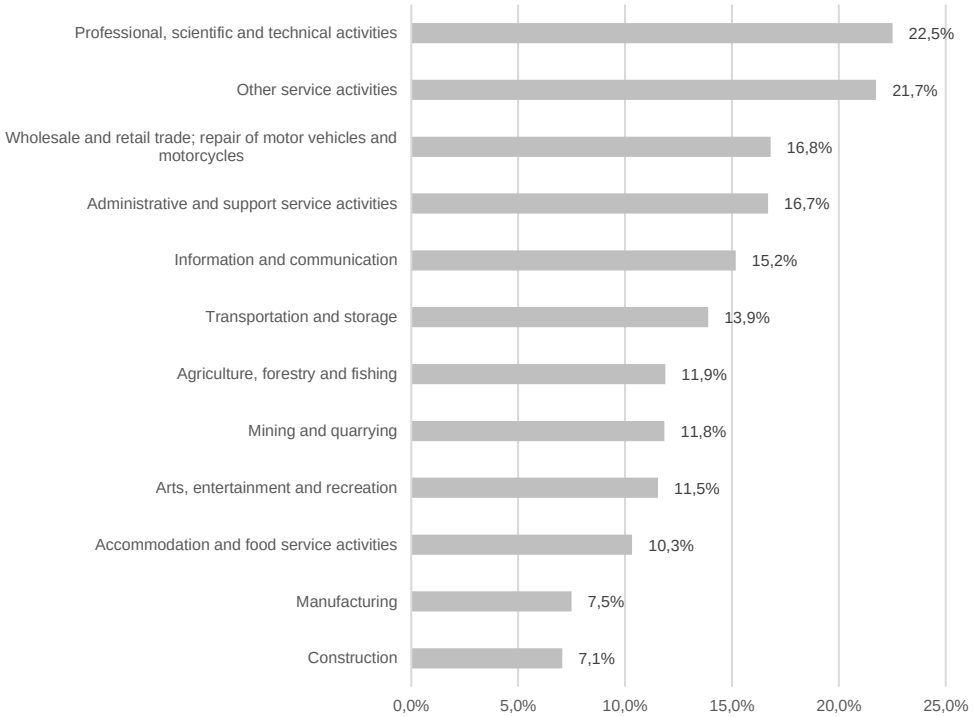
Notes: Blue line = net effect of positive and negative contributions.

5.3.3 VAT gap in sectors

Our sectoral VAT gap estimates are described in section A3.5 of the technical appendix and presented in Chart 11 (NACE sections) and 12 (NACE divisions). Sectors with the largest VAT gap (as % of potential VAT in the sector) include various professional services, other service activities and trade. The top sectors in the ranking of more detailed sectors include (1) rental and leasing activities, (2) other professional, scientific and technical activities, (3) activities of head offices; management consultancy services (4) advertising and market research. The bottom sectors include various kinds of manufacturing. A bit surprising are low positions in the ranking of the construction and accommodation and food service sectors. Such sectors in many countries are characterised by relatively large role of unregistered employment, i.e. hidden costs, that often also leads to hiding some revenues. Maybe our model has not been able to account for such specifics.[74] On the other hand, it is worth noting, at least in the context of the shadow economy, that in such sectors there are also many large companies which likely report most of their revenues and may outweigh the effects generated by some smaller companies in the sectors.

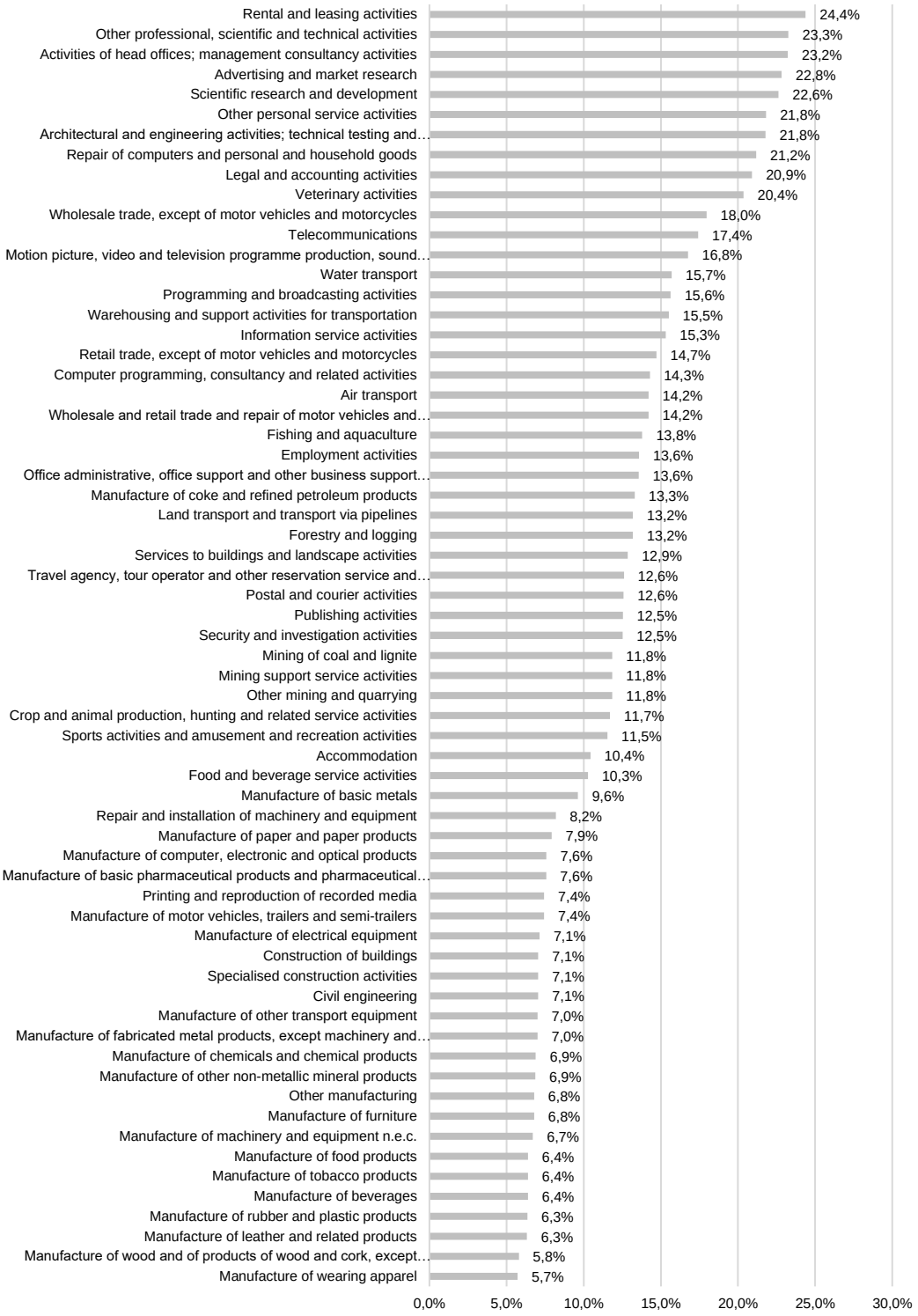
[74] If one has some estimates of unregistered employment at the sectoral level, they can be used as an additional variable in the future development of the model.

Chart 11 – VAT gap in sectors (NACE sections) in Bulgaria in 2021 (% of potential VAT in the sector under perfect compliance)



Source: EY.
Notes: We omitted sectors for which we assumed zero VAT gap.

Chart 12 – VAT gap in sectors (NACE divisions) in Bulgaria in 2021 (% of potential VAT in the sector under perfect compliance)



Source: EY.
Notes: We omitted sectors for which we assumed zero VAT gap.

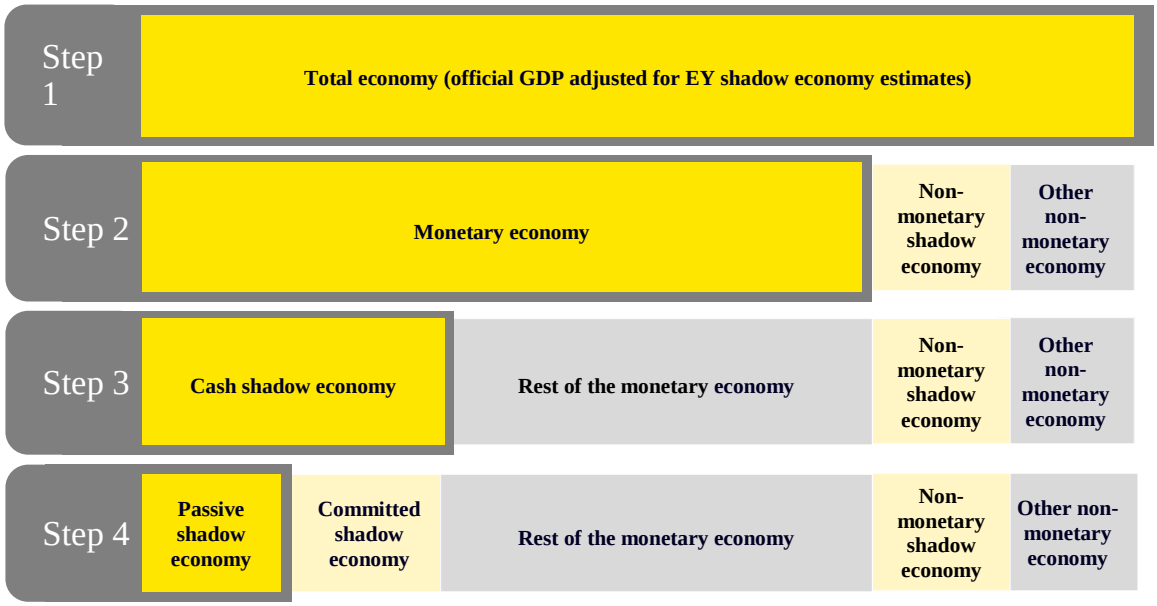
A. Technical appendix

A1. Shadow economy and related part of the tax gap

A1.1 Steps in our approach

Our approach consists of four steps decomposing the total economy into different components presented in Figure A.1 and described below.

Figure A.1 – Decomposition of the total economy into shadow and registered components



Note: The proportions of the areas above do not reflect the proportions of different components of the total economy.
Source: EY.

Step 1. Relationship between total economy and official GDP

We start with official GDP figures ($Y_{i,t}^{OFFICIAL}$ for country *i* in period *t)*. We check if information on the shadow (non-observed) economy estimates included in GDP figures of the statistical office is available. If it is, and we conclude that such estimates account for all relevant aspects of the shadow economy, we can later calculate the total economy size (total GDP, $Y_{i,t}^{TOTAL}$) by adjusting official GDP for the difference between our and statistical office's shadow economy estimates. Yet, it was not in the case of Bulgaria, so for simplicity we assume that the shadow economy included in the official GDP is equal to our estimate. This step is later needed to express our results in local currency units or as a percentage of official GDP, since our methodology returns outcomes as a percentage of total GDP.

Step 2. Splitting total economy into monetary and non-monetary components

We split the total economy into monetary ($Y_{i,t}^{MONETARY,TOTAL}$), i.e., payment-based, and non-monetary activities by estimating the latter. Non-monetary economy includes two components: 1) imputed rents of owners-occupiers that could be found in statistical offices datasets and 2) household production of goods for own final use (non-monetary shadow economy, $Y_{i,t}^{NMSE}$, mainly related to agriculture). Sometimes value of 2) is also easily available at the statistical office but it was not the case for Bulgaria. Otherwise,

we estimate it based on the role of agriculture in the economy and results of Blades (1975) who analysed its link with the non-monetary shadow economy in various countries (for details see EY (2019)[75]). Yet, for most developed countries the non-monetary shadow economy is rather small and not relevant from the perspective of policies to increase tax compliance.

Step 3. Estimating the cash shadow economy: currency demand analysis (CDA)

In step 3 we first focus on measuring the share of the monetary (or "cash") shadow economy in total monetary economy ($\frac{Y_{i,t}^{CASH,SHADOW}}{Y_{i,t}^{MONETARY,TOTAL}}$).

Inspired by the existing and our CDA research, we propose a modified approach, recognized by other shadow economy researchers.[76] We distinguish following substeps.

Substep 3.1. Estimation of CDA model

The first substep is an econometric estimation of the currency demand equation:

$$\boldsymbol{CASH\_M1}_{i,t} = \alpha_i + \boldsymbol{\beta}_{i,t}^{(1)}\boldsymbol{x}_{\mathbf{1},i,t} + \boldsymbol{\beta}_{i,t}^{(2)}\boldsymbol{x}_{\mathbf{2},i,t} + \varepsilon_{it}, \text{ (1)}$$

where i represents the analysed country and t stands for the analysed time period. In this equation, the explained (dependent) variable is the share of currency in circulation ("cash") in the M1 monetary aggregate ("total transactional money" including "cash" and overnight deposits). To explain its variation, we use two groups of explanatory variables:

Cash shadow economy determinants ($\mathbf{x_1}$). They mostly affect the willingness of agents to operate in the shadow economy (e.g. state of labour market, institutional indicators, taxation, etc.) and through this channel impact the dependent variable.

Control variables ($\mathbf{x_2}$). These variables, after controlling for the influence of $\mathbf{x_1}$ should not (directly) impact the shadow economy but may still have influence on the dependent variable. They are related to the level of the economic development, monetary conditions, etc.

$\boldsymbol{\beta}_{i,t}^{(1)}$ **and** $\boldsymbol{\beta}_{i,t}^{(2)}$ represent vectors of the regression coefficients (they may also include interactions with real GDP per capita to account for their conditionality on the development level). Finally, ε_{it} is the error term. Additionally, we include the individual effects, α_i, which represent time-invariant, unobservable country characteristics that affect the demand for cash in each country.

The construction of the coefficients α_i, $\boldsymbol{\beta}_{i,t}^{(1)}$and $\boldsymbol{\beta}_{i,t}^{(2)}$ reflects country heterogeneity which is crucial when using data for many countries. Individual effects (α_i) are estimated as fixed effects. Panel data makes it possible to incorporate such effects that can represent constant unobservable cultural factors.

We consider a wide range of potential explanatory variables from the two groups discussed above. Our preferred approach to the selection of variables and assessment of their impact is based on the frequentist and/or Bayesian model

[75] EY (2019), op. cit.
[76] See e.g. Medina L., Schneider F. (2018), "Shadow Economies Around the World: What Did We Learn Over the Last 20 Years?", *IMF Working Paper*, no. WP/18/17.

averaging procedure in which a wide array of variants of equation (1) is estimated using the Panel Corrected Standard Errors (PCSE) method[77], with different combinations of considered variables.

Substep 3.2. Using the CDA model to measure the shadow-economy-related cash

In the second substep, we set the values of $\boldsymbol{x}_1$ vector in equation (1) at their "best" (benchmark) observable levels for the countries in the sample (e.g. the lowest unemployment rate) and estimate the theoretical value of the explained variable in the case of the lowest possible cash shadow economy.

The difference between the fitted value from the model (i) calculated on the basis of the factual values of $\boldsymbol{x}_1$ in the given country and (ii) calculated on the basis of the "best" (benchmark) values of $\boldsymbol{x}_1$ in the sample may be interpreted as the share of cash related to cash shadow economy transactions in the M1 aggregate ($\frac{\boldsymbol{C}_{i,t}^{\boldsymbol{SHADOW}}}{\boldsymbol{M1}_{i,t}}$). Given the observed stock of the M1 aggregate for a given country and period, the obtained difference allows us to calculate the amount of cash that is attributable to the cash shadow economy ($\boldsymbol{C}_{i,t}^{\boldsymbol{SHADOW}}$).

Substep 3.3. Conversion of the shadow cash into the cash shadow economy

In the third substep, we estimate the size of the cash shadow economy[78]. First, we assume that the *velocity* of money in the cash shadow economy is equal to the velocity of money in the overall monetary economy:

$$\frac{\boldsymbol{Y}_{i,t}^{\boldsymbol{MONETARY,TOTAL}}}{\boldsymbol{M1}_{i,t}} = \frac{\boldsymbol{Y}_{i,t}^{\boldsymbol{CASH,SHADOW}}}{\boldsymbol{C}_{i,t}^{\boldsymbol{SHADOW}}}, \quad (2)$$

where $\boldsymbol{Y}_{i,t}^{\boldsymbol{MONETARY,TOTAL}}$ and $\boldsymbol{Y}_{i,t}^{\boldsymbol{CASH,SHADOW}}$ denote the monetary output in the total and shadow economy, respectively; $\boldsymbol{C}_{i,t}^{\boldsymbol{SHADOW}}$ stands for the amount of cash used for settling transactions in the cash shadow economy and $\boldsymbol{M1}_{i,t}$ is the M1 total transactional money.

We transform equation (2) to estimate the share of the cash shadow economy output in the total monetary output (including also the cash shadow economy) without knowing the exact value of the velocity of money*:*

$$\frac{\boldsymbol{Y}_{i,t}^{\boldsymbol{CASH,SHADOW}}}{\boldsymbol{Y}_{i,t}^{\boldsymbol{MONETARY,TOTAL}}} = \frac{\boldsymbol{C}_{i,t}^{\boldsymbol{SHADOW}}}{\boldsymbol{M1}_{i,t}}. \quad (3)$$

Note that $\frac{\boldsymbol{C}_{i,t}^{\boldsymbol{SHADOW}}}{\boldsymbol{M1}_{i,t}}$ is the endpoint of the substep 3.2. However, it is only related to those economic activities that include monetary transactions. In order to obtain the estimate of the total shadow economy $\boldsymbol{Y}_{i,t}^{\boldsymbol{TOTAL,SHADOW}}$ (as a share in total economy $\boldsymbol{Y}_{i,t}^{\boldsymbol{TOTAL}}$), we use the following formula:

$$\frac{\boldsymbol{Y}_{i,t}^{\boldsymbol{TOTAL,SHADOW}}}{\boldsymbol{Y}_{i,t}^{\boldsymbol{TOTAL}}} = \frac{\boldsymbol{Y}_{i,t}^{\boldsymbol{CASH,SHADOW}}}{\boldsymbol{Y}_{i,t}^{\boldsymbol{MONETARY,TOTAL}}} \times \frac{\boldsymbol{Y}_{i,t}^{\boldsymbol{MONETARY,TOTAL}}}{\boldsymbol{Y}_{i,t}^{\boldsymbol{TOTAL}}} + \frac{\boldsymbol{Y}_{i,t}^{\boldsymbol{NMSE}}}{\boldsymbol{Y}_{i,t}^{\boldsymbol{TOTAL}}}, \quad (4)$$

[77] The method is robust to: contemporaneous correlation of error terms between panel units, serial correlation of order 1 of the error term (a common serial correlation coefficients for all the panels is selected) as well as to heteroskedasticity.

[78] The size of the cash shadow economy corresponds to the part of monetary output / monetary GDP that is generated in the shadow economy.

in which $\frac{Y_{i,t}^{MONETARY,TOTAL}}{Y_{i,t}^{TOTAL}}$ is the output of Step 2 and the $Y_{i,t}^{NMSE}$ is the non-monetary shadow economy estimated earlier. Finally, the share of the total shadow economy in the official GDP estimate is obtained using the following adjustment:

$$\frac{Y_{i,t}^{TOTAL,SHADOW}}{Y_{i,t}^{OFFICIAL}} = \frac{Y_{i,t}^{TOTAL,SHADOW}}{Y_{i,t}^{TOTAL}} \times \frac{Y_{i,t}^{TOTAL}}{Y_{i,t}^{OFFICIAL}}, \quad (5)$$

in which $\frac{Y_{i,t}^{OFFICIAL}}{Y_{i,t}^{TOTAL}}$ is the result of the Step 1.

Step 4. Estimation of the passive and committed shadow economy

Passive shadow economy (see section 2.2 and 3.4.3 for definition) consists in underreporting of the revenues by registered, legally operating entities. We assume that the remaining part of the shadow economy, i.e., the committed shadow economy and the non-monetary shadow economy, is related to the value added generated by unregistered labour and we estimate such value in two substeps.

Our approach to empirical distinguishing the passive and committed components is based on the assumption that the output of the committed shadow economy is correlated with and mirrored by shadow labour force inputs. In order to approximate the value of committed shadow economy, we evaluate the share of unregistered employees by comparing the official number of employees under labour contract published by the National Statistical Institute of Bulgaria[79] with the number of declared employees based on the Labour Force Survey. Next, we multiply the obtained share of informal workers by three factors: (1) the share of compensation of employees in the GDP, (2) the ratio of average wage of workers performing elementary occupations (ISCO - International Standard Classification of Occupations code 09) to the average wage and (3) the share of full-time workers in the elementary occupations (to account for the fact that people working informally often earn less money and work less hours[80]). This way we obtain the adjusted estimate of the share of informal employment in the total employment, measuring their income share and approximating the share of the committed shadow economy in GDP. Finally, to obtain the passive shadow economy estimate, we subtract the committed shadow economy from our cash shadow economy estimate (calculated in substep 3.3).

Step 5. Estimation of lost government revenues

To estimate the value of additional VAT revenues due to the cash shadow economy, we multiply the value of cash shadow economy by an estimated theoretical VAT rate. Bulgaria has different VAT rates (reduced 9% and VAT exemptions/0% rate) for certain categories of goods and services, which means that we have to take into account the sectorial structure of the shadow economy to assess this rate in Bulgaria. For simplicity, we assume that the structure of consumer cash expenditure in the cash shadow economy is the same as the representative structure of household consumption, as reflected by the weights from the basket of consumer goods and services used in calculation of the CPI inflation (based on Household Budget Surveys). Accordingly, we calculate the theoretical VAT rate for the cash shadow economy transactions as a weighted average of official (standard or reduced) VAT rates applied to different goods/services in the economy, computed as if all the transactions were

[79] https://www.nsi.bg/en/content/3953/total (online, accessed: 06.04.2022).
[80] https://ec.europa.eu/social/main.jsp?catId=1322&langId=en (online, accessed: 06.04.2022).

reported. We also take into account that some services or goods might be exempted from the VAT. The formula for calculating lost VAT for Bulgaria in 2022 is as follows:

$$VAT\ (\%\ of\ GDP) = \frac{VAT\_RATE_{BG,2022}}{1 + VAT\_RATE_{BG,2022}} * CASH\_SE_{BG,2022},$$

where VAT_RATE is the estimated VAT rate and CASH_SE is the value of cash shadow economy (as % of GDP). The BG index denotes Bulgaria and 2022 is the year to which the values refer.

More specifically, VAT_RATE is calculated as the sum of contributions to the average from each commodity group[81]. Each category's contribution is determined by multiplying its weight and the relevant VAT rate (standard, reduced or zero). Weight is the ratio of consumption expenses in the given category to the total consumption expenses[82]. When the tax rates varied among one category, the relevant calculations are made. For instance, for the recreation and culture group, which has a zero rate for culture and sport subcategory that accounts for about 21.7% of the group's expenses, and 20% rate for the remaining subcategories, we calculate it as: 0.217 * 0 + 0.783 * 0.2 (which gives 15.7% rate).

Estimating the impact of cash shadow economy on income taxes is more complicated, as we are aware that some (especially small) businesses may pay PIT instead of CIT. To calculate the effective CIT and PIT income tax rate applicable to cash shadow economy, we divide the sum of CIT and PIT revenues by the gross value added[83] less imputed rents and shadow economy estimate (since the last two elements are not subject to taxation). Before applying the effective income tax rate, we deduct the value of VAT due from the cash shadow economy estimate to take into account that VAT, if applied, would reduce the income tax base (such additional cost would occur in the case of registration of the transaction). Here is the formula for calculating lost income taxes for Bulgaria in 2022:

$$INCOME\_TAXES(\%\ of\ GDP) =$$
$$= CIT\_PIT\_RATE_{BG,2022} * (1 - \frac{VAT\_RATE_{BG,2022}}{1 + VAT\_RATE_{BG,2022}}) * CASH\_SE_{BG,2022},$$

where CIT_PIT_RATE is the mixed CIT and PIT rate[84], VAT_RATE is the estimated VAT rate (we assume that the VAT paid from the newly registered transactions would be deducted from the income tax base) and CASH_SE is the value of the cash shadow economy (% of GDP). The BG index denotes Bulgaria and 2022 is the year to which the values refer.

A1.2 Variables considered in the shadow economy model

The table below presents the variables considered in our shadow economy model.

[81] Food, and non-alcoholic beverages, Alcoholic beverages and tobacco, Clothing and footwear, Housing, water, electricity, gas and other fuels, Furniture household goods and maintenance, Health, Transport, Communication, Recreation and culture, Education service, Restaurants and hotels, Miscellaneous goods and services.
[82] Source: https://www.nsi.bg/sites/default/files/files/metadata/CPIBasket_2022-ENG.pdf (online, accessed: 24.04.2022).
[83] We assume that figures on the value added of the National Statistical Institute already include shadow economy estimates equal to our estimates in this area.
[84] Calculated as the sum of PIT and CIT revenues divided by the Gross Value Added from which we subtracted shadow economy and imputed rents. The last available data for PIT and CIT revenues (on the website of the Bulgarian Ministry of Finance) were for 2021, so we have calculated the mixed PIT and CIT rate for 2021 and assumed it will be the same in 2022.

Table A.1 – Information about variables considered in the shadow economy model

group of variables for our analysis	closest group(s) of factors from the literature review in the report	name of the variable	description	decision to exclude from the analysis (if already made)	additional comments	number of observations	number of years	latest year available	earliest year available	number of countries (with any data point)	sources	hyperlinks
Dependent (explained) variable	Not applicable	CASH_M1	Share of the currency in circulation in the M1 monetary aggregate (currency in circulation + transferable deposits), %. Numerator's data series: Currency Outside Depository Corporations (from Depository Corporations data table) or - in case of missing data - Currency Outside Banking Institutions (from Non-Standardized Presentation data table). Denominator's data series: M1 or sum of data on Transferable Deposits and Currency from the same tables.			3283	27	2021	1995	149	International Monetary Fund	https://data.imf.org/?sk=B83F71E8-61E3-4CF1-8CF3-6D7FE04D0930&sId=1390030341854
Shadow economy determinant: labour market / business cycle	(Financial) condition of taxpayers (level) / shock to financial condition	NON_EMPLOYED	The share of unemployed and economically inactive in the total population aged 15-64, %.		Non-employed share was calculated as 1-employed share.	5017	27	2021	1995	186	International Labour Organization, own calculations	https://www.ilo.org/shinyapps/bulkexplorer53/?lang=en&id=EMP_2EMP_SEX_AGE_NB_A
Shadow economy determinant: labour market / business cycle	(Financial) condition of taxpayers (level) / shock to financial condition	UNEMP	Unemployment rate, % of total labor force (economically active population)		The same definition as "unemployment rate" (percent of total labor force) in the IMF, for which there are publicly available forecasts.	5049	27	2021	1995	187	World Bank - modeled ILO estimate	https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS
Shadow economy determinant: labour market / role of small and specific entities	Business form/ (financial) condition of taxpayers (level) / level of economic development	SELF_EMPLOYED	The ratio of the total number of self-employed (employers, own-account workers, members of producers' cooperatives, and contributing family workers) to the population aged 15-64, %.			5022	27	2021	1995	186	International Labour Organization, own calculations	https://www.ilo.org/shinyapps/bulkexplorer53/?lang=en&id=EMP_2EMP_SEX_AGE_NB_A
Shadow economy determinant: labour market / role of small and specific entities	Business form/ (financial) condition of taxpayers (level) / level of economic development	FAMILY_WORK	The ratio of the total number of contributing family workers to the population aged 15-64, %. Contributing family workers are own-account workers in the market-oriented business that is conducted by a related person who lives in the same household.			5022	27	2021	1995	186	International Labour Organization, own calculations	
Shadow economy determinant: labour market / role of small and specific entities	Business form/ (financial) condition of taxpayers (level) / level of economic development	OWN_ACCOUNT_WORK	The ratio of the total number of own-account workers to the population aged 15-65, %.			5022	27	2021	1995	186	International Labour Organization, own calculations	
Shadow economy determinant: institutional / regulatory	Public governance (service) quality / perceived tax service quality/ trust in government	CONTROL_OF_CORRUPTION	The value of the indicator measuring the control of corruption from the Worldwide Governance Indicators. It ranges from approximately -2.5 (low control of corruption) to 2.5 (high control of corruption). It reflects perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as "capture" of the state by elites and private interests.			4564	23	2021	1996	203	World Bank - Worldwide Governance Indicators	https://databank.worldbank.org/source/worldwide-governance-indicators/preview/on#
Shadow economy determinant: institutional / regulatory	Public governance (service) quality / perceived tax service quality/ trust in government	GOV_EFFECTIVENESS	The value of the indicator measuring the government effectiveness from the Worldwide Governance Indicators. It ranges from approximately -2.5 (low government effectiveness) to 2.5 (high government effectiveness). It reflects perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies.			4540	23	2021	1996	203	World Bank - Worldwide Governance Indicators	
Shadow economy determinant: institutional / regulatory	Public governance (service) quality / perceived tax service quality/ trust in government	POLITICAL	The value of the indicator measuring the political stability and absence of violence/terrorism from the Worldwide Governance Indicators. It ranges from approximately -2.5 (weak political stability and a greater presence of violence/terrorism) to 2.5 (strong political stability and lower presence of violence/terrorism). Political Stability and Absence of Violence/Terrorism measures perceptions of the likelihood of political instability and/or politically-motivated violence, including terrorism.			4615	23	2021	1996	205	World Bank - Worldwide Governance Indicators	
Shadow economy determinant: institutional / regulatory	Complexity of tax (or law) system / public governance (service) quality / perceived tax service quality/ trust in government	REGULATORY	The value of the indicator measuring the regulatory quality from the Worldwide Governance Indicators. It ranges from approximately -2.5 (low regulatory quality) to 2.5 (high regulatory quality). It reflects perceptions of the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development.			4542	23	2021	1996	203	World Bank - Worldwide Governance Indicators	
Shadow economy determinant: institutional / regulatory	Public governance (service) quality / perceived tax service quality/ trust in government	RULE_OF_LAW	The value of the indicator measuring the rule of law from the Worldwide Governance Indicators. It ranges from approximately -2.5 (weak rule of law) to 2.5 (strong rule of law). It reflects perceptions of the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, property rights, the police, and the courts, as well as the likelihood of crime and violence.			4654	23	2021	1996	205	World Bank - Worldwide Governance Indicators	
Shadow economy determinant: institutional / regulatory	Public governance (service) quality / perceived tax service quality/ trust in government	VOICE_AND_ACCOUNTABILITY	The value of the indicator measuring the voice and accountability from the Worldwide Governance Indicators. It ranges from approximately -2.5 (fewer rights) to 2.5 (more rights). It reflects perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media.			4648	23	2021	1996	205	World Bank - Worldwide Governance Indicators	
Shadow economy determinant: institutional / regulatory	Public governance (service) quality / perceived tax service quality/ trust in government	INDEPENDENCE	Judicial independence, index with values from 0 (worst) to 10 (best)		2A; the first source of this component is from the Global Competitiveness Report question: "Is the judiciary in your country independent from political influences of members of government, citizens, or firms? No—heavily influenced (= 1) or Yes—entirely independent (=	3548	22	2020	1995	164	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=3
Shadow economy determinant: institutional / regulatory	Public governance (service) quality / perceived tax service quality/ trust in government	COURTS	Impartial courts, index with values from 0 (worst) to 10 (best)		2B; The first source of this component is the Global Competitiveness Report question: "The legal framework in your country for private businesses to settle disputes and challenge the legality of government actions and/or regulations is inefficient and subject to manipulation	3548	22	2020	1995	164	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=4
Shadow economy determinant: institutional / regulatory	Public governance (service) quality / perceived tax service quality/ trust in government	PROPERTY_RIGHTS	Protection of property rights, index with values from 0 (worst) to 10 (best)		2C; The first source of this component is the Global Competitiveness Report question: "Property rights, including over financial assets, are poorly defined and not protected by law (= 1) or are clearly defined and well protected by law (= 7)". The second source is Property	3511	22	2020	1995	164	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=5
Shadow economy determinant: institutional / regulatory	Public governance (service) quality / perceived tax service quality/ trust in government	INTEGRITY	Integrity of the legal system, index with values from 0 (worst) to 10 (best)		2E; The first source of this component is the International Country Risk Guide Political Risk Component I for Law and Order: "Two measures comprising one risk component. Each sub-component equals half of the total. The 'law' sub-component assesses the strength and	3541	22	2020	1995	164	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=6
Shadow economy determinant: institutional / regulatory	Complexity of tax (or law) system / public governance (service) quality / perceived tax service quality/ trust in government	CONTRACTS_ENFORCEMENT	Legal enforcement of contracts, index with values from 0 (worst) to 10 (best)		2F; This first source of this component is the World Bank's Doing Business estimates for the time and money required to collect a debt. The debt is assumed to equal 200% of the country's per-capita income where the plaintiff has complied with the contract and judicial	3514	22	2020	1995	163	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=7
Shadow economy determinant: institutional / regulatory	Complexity of tax (or law) system / public governance (service) quality / perceived tax service quality/ trust in government	PROTECTIVE_GOVERNMENT	Legal System & Property Rights, index with values from 0 (worst) to 10 (best)		2; the nine components in this area (A. Judicial independence, B. Impartial courts, C. Protection of property rights, D. Military interference in rule of law and politics, E. Integrity of the legal system, F. Legal enforcement of contracts, G. Regulatory costs of the sale of real	3548	22	2020	1995	164	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=8
Shadow economy determinant: institutional / regulatory	Complexity of tax (or law) system / public governance (service) quality / perceived tax service quality/ trust in government	HIRING	Hiring regulations and minimum wage, index with values from 0 (worst) to 10 (best)		5Bi; This sub-component is based on the "Employing Workers" section of the World Bank's Doing Business and uses the following components: (1) whether fixed-term contracts are prohibited for permanent tasks; (2) the maximum cumulative duration of fixed-term	3040	22	2020	1995	164	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=9
Shadow economy determinant: institutional / regulatory	Complexity of tax (or law) system / public governance (service) quality / perceived tax service quality/ trust in government	HIRING_FIRING	Hiring and firing regulations, index with values from 0 (worst) to 10 (best)		5Bii; This sub-component is based on the Global Competitiveness Report question: "The hiring and firing of workers is impeded by regulations (= 1) or flexibly determined by employers (= 7)". The question's wording has varied over the years.	2814	22	2020	1995	161	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=10
Shadow economy determinant: institutional / regulatory	Complexity of tax (or law) system / public governance (service) quality / perceived tax service quality/ trust in government	LABOR_MARKET	Labor market regulations, index with values from 0 (worst) to 10 (best)		5B; The labor-market component (5B) is designed to measure the extent to which these restraints upon economic freedom are present. In order to earn high marks in the component rating regulation of the labor market, a country must allow market forces to	3072	22	2020	1995	164	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=11
Shadow economy determinant: institutional / regulatory	Complexity of tax (or law) system / public governance (service) quality / perceived tax service quality/ trust in government	ADMINISTRATIVE	Administrative requirements, index with values from 0 (worst) to 10 (best)		5Ci; This sub-component is based on the Global Competitiveness Report question: "Complying with administrative requirements (permits, regulations, reporting) issued by the government in your country is (1 = burdensome, 7 = not burdensome)". The	2729	21	2020	2000	161	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=12
Shadow economy determinant: institutional / regulatory	Complexity of tax (or law) system / public governance (service) quality / perceived tax service quality/ trust in government	REGULATORY_BURDEN	Regulatory Burden, index with values from 0 (worst) to 10 (best)		5Cii; This sub-component is based on the "Regulatory Burden Risk Ratings" from IHS Markit, which measures "[t]he risk that normal business operations become more costly due to the regulatory environment. This includes regulatory compliance and	2857	22	2020	1995	164	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=13
Shadow economy determinant: institutional / regulatory	Complexity of tax (or law) system / public governance (service) quality / perceived tax service quality/ trust in government	STARTING_A_BUSINESS	Starting a business, index with values from 0 (worst) to 10 (best)		5Ciii; This sub-component is based on the World Bank's Doing Business data on the amount of time and money it takes to start a new limited-liability business. Countries where it takes longer or is more costly to start a new business are given lower	3051	22	2020	1995	164	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=14
Shadow economy determinant: institutional / regulatory	Complexity of tax (or law) system / public governance (service) quality / perceived tax service quality/ trust in government	PUBLIC_ADMINISTRATION	Impartial Public Administration, index with values from 0 (worst) to 10 (best)		5Civ; This sub-component is based on the "Rigorous and Impartial Public Administration" data from the V-Dem dataset. If nepotism, cronyism, and discrimination are widespread in the application of public administration, countries receive a lower score. The rating for this	3481	22	2020	1995	161	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=15
Shadow economy determinant: institutional / regulatory	Complexity of tax (or law) system / public governance (service) quality / perceived tax service quality/ trust in government	LICENSING_RESTRICTIONS	Licensing restrictions, index with values from 0 (worst) to 10 (best)		5Cv; This sub-component is based on the World Bank's Doing Business data on the time in days and monetary costs required to obtain a license to construct a standard warehouse. Zero-to-10 ratings were constructed for (1) the time cost (measured in number of calendar	2514	17	2020	2004	162	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=16

Shadow economy determinant: institutional / regulatory	Complexity of tax (or law) system / public governance (service) quality / perceived tax service quality/ trust in government	BUSINESS_REGULATIONS	Business regulations, index with values from 0 (worst) to 10 (best)		5C; The sub-components of 5C are designed to identify the extent to which regulations and bureaucratic procedures restrain entry and reduce competition. In order to score high in this portion of the index, countries must allow markets to determine prices and refrain from	3079	22	2020	1995	164	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=17
Shadow economy determinant: institutional / regulatory	Complexity of tax (or law) system / public governance (service) quality / perceived tax service quality/ trust in government	REGULATION	Regulation, index with values from 0 (worst) to 10 (best)		5; The fifth area of the index focuses on regulatory restraints that limit the freedom of exchange in credit, labor, and product markets.	3216	22	2020	1995	164	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=18
Shadow economy determinant: institutional / regulatory	(Perceived) corruption	CORRUPTION	Corruption Perceptions Index; level of corruption ranges from 0 (highly corrupt) to 100 (no corruption)	rather exclude, low number of observations		1749	10	2021	2012	179	Transparency International	https://www.transparency.org/en/cpi/2022
Shadow economy determinant: institutional / regulatory	Trust in government	APP_TRUST_GOVERN	An indicator measuring the trust in the government, based on answers to the following question: *how much confidence you have in the government? [1 - a great deal; 4 - none at all]*	rather exclude, low number of observations	Different variants considered: interpolation with naive forecast, interpolation with dynamic forecast and interpolation without forecast	1580	27	2021	1995	109	World Values Survey	https://www.worldvaluessurvey.org/WVSContents.jsp
Shadow economy determinant: institutional / regulatory	Social norms / ethics/ conscience	APP_GOVERN_RESP	An indicator based on answers to the question comparing government's vs individual's responsibility *[1 - The government should take more responsibility to ensure that everyone is provided for, 10 - People should take more responsibility to provide for themselves]*	rather exclude, low number of observations	Different variants considered: interpolation with naive forecast, interpolation with dynamic forecast and interpolation without forecast	1580	27	2021	1995	109	World Values Survey	
Shadow economy determinant: institutional / regulatory	Social norms / ethics/ conscience	APP_AVOID_FARE	An indicator based on answers to the question related with avoiding a fare on public transport *[1 - never justifable, 10 - always justifiable]*	rather exclude, low number of observations	Different variants considered: interpolation with naive forecast, interpolation with dynamic forecast and interpolation without forecast	1580	27	2021	1995	109	World Values Survey	
Shadow economy determinant: institutional / regulatory	Social norms / ethics/ conscience	APP_CHEAT_TAX	An indicator based on answers to the question related with cheating on taxes *[1 - never justifable, 10 - always justifiable]*	rather exclude, low number of observations	Different variants considered: interpolation with naive forecast, interpolation with dynamic forecast and interpolation without forecast	1580	27	2021	1995	109	World Values Survey	
Shadow economy determinant: institutional / regulatory	Public governance (service) quality / perceived tax service quality/ trust in government	APP_DEMOCRATIC	An indicator based on answers to the question related with having a democratic political system *[1 - very goode, 4 - very bad]*	rather exclude, low number of observations	Different variants considered: interpolation with naive forecast, interpolation with dynamic forecast and interpolation without forecast	1580	27	2021	1995	109	World Values Survey	
Shadow economy determinant: taxation	Tax rate / tax at risk	TAX_REVENUE	Tax revenue, % of GDP		The year 2021 for all countries & data for Spain (1995-2020), Kazakhstan 2005-2009, Kiribati 2010, Madagascar 1995-1996, Senegal 2019-2020 and Kyrgyz 2013 were calculated on the basis of dynamics obtained from IMF data; data for Kuwait (2001-2015), Japan (1995-2020) were taken directly from IMF data	3048	27	2021	1995	155	World Bank - International Monetary Fund, Government Finance Statistics Yearbook and data files, and World Bank and OECD GDP estimates, own calculations based on data	Government Finance Statistics (GFS), Revenue: https://data.imf.org/?sk=a0867067-d23c-4ebc-ad23-d3b015045405&sId=1435757954964
Shadow economy determinant: taxation	Tax rate / tax at risk	TAXES_GOODS_AND_SERVICES	Taxes on goods and services, % of value added of industry and services		The year 2021 & data for Spain 1995-2017, Kazakhstan 2005-2009, Senegal 2019-2020 and Trinidad&Tobago 1995-2000 were calculated on the basis of dynamics obtained from IMF and WB data; calculated as the share of Taxes on goods and services (domestic currency) in the sum of Industry (including construction), value added (current LCU) and Services, value	2907	27	2021	1995	151	World Bank - International Monetary Fund, Government Finance Statistics Yearbook and data files, and World Bank and OECD value added estimates, own calculations based on	https://data.imf.org/?sk=a0867067-d23c-4ebc-ad23-d3b015045405&sId=1435757954964
Shadow economy determinant: taxation	Tax rate / tax at risk	TAXES_INCOME_PROFITS_GAINS	Taxes on income, profits and capital gains, % of GDP		Calculated as the % share of Taxes on income, profits and capital gains (current LCU) in the GDP (current LCU); the year 2021 & data for Spain 1995-2016, Kazakhstan 2005-2009, Kiribati 2010, Kyrgyz 2013, Senegal 2019-2020, West Bank and Gaza 2019-2010 and Trinidad&Tobago 1995-2000 were calculated on the basis of dynamics obtained from IMF data; data for Japan	2954	27	2021	1995	152	World Bank - International Monetary Fund, Government Finance Statistics Yearbook and data files	
Shadow economy determinant: taxation	Tax rate / tax at risk	VAT	Standard VAT/GST rate, %		For two or more different values, the mean was taken	3599	27	2021	1995	178	International Monetary Fund	https://www.imf.org/external/np/fad/tpaf/pages/vat.htm; https://www.vatlive.com/; https://www.imf.org/external/np/fad/tpaf/files/VAT_historic_rates.xlsx
Shadow economy determinant: taxation	Tax rate / tax at risk	CIT	Corporate Income Tax (CIT) rate, %			4970	28	2022	1995	211	Tax Foundation	https://taxfoundation.org/publications/corporate-tax-rates-around-the-wo
Shadow economy determinant: taxation	Tax rate / tax at risk	PIT	Top marginal income tax rate, %		1Di; Countries with higher marginal tax rates that take effect at lower income thresholds received lower ratings based on the matrix available here: https://www.fraserinstitute.org/sites/default/files/uploaded/2022/economic-freedom-of-	2976	22	2020	1995	164	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=0
Shadow economy determinant: taxation	Tax rate / tax at risk	PIT_PAYROLL	Top marginal income and payroll tax rate, %		1Dii; Countries with higher marginal income and payroll (wage) tax rates that take effect at lower income thresholds received lower ratings based on the matrix available here: https://www.fraserinstitute.org/sites/default/files/uploaded/2022/economic-freedom-of-	2787	22	2020	1995	161	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=1
Shadow economy determinant: taxation	Tax rate / tax at risk	PIT_BOTH	Top marginal tax rate, %		1D; this component is based on the top marginal income tax rate (1Di) and the top marginal income and payroll tax rate (1Dii) and the income threshold at which these rates begin to apply. These two sub-components are averaged to calculate the top marginal tax rate (1D).	2976	22	2020	1995	164	Economic Freedom of the World, Fraser Institute	https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=2
Shadow economy determinant: taxation	Tax rate / tax at risk	AVG_MAIN_TAXES_RATE	Average rate of VAT, CIT and PIT, %			2976	22	2020	1995	164	see sources for VAT, CIT and PIT rates	
Shadow economy determinant: taxation	Tax system / perceived tax service quality	TAX_TIME	Time to prepare and pay taxes in hours	exclude, the reasearch project and data was discontinued		2677	15	2019	2005	189	World Bank - Doing Business project	https://data.worldbank.org/indicator/IC.TAX.DURS
Partly shadow economy determinant, partly control variable: payment card system / financial inclusion	Third party reporting available	CARDS	The sum of credit and debit cards *per capita*		Data for the end of the period	1627	18	2021	2004	137	European Central Bank, International Monetary Fund, International Bank for Settlements, national central banks	https://sdw.ecb.europa.eu/browseTable.do?org.apache.struts.taglib.html.TOKEN=6ab720d6adb4e6c43c02c84b6d69153d&df=true&ec=&dc=&oc=&pb=&rc=&DATASET=0&removeItem=&removedItemList=&mergeFilter=&activeTab=PSS&showHide=&REF_AREA.93=AT&REF_AREA.93=BE&REF_AREA.93=BG&REF_AREA.93=CY&REF_AREA.93=CZ&REF_AREA.93=D0&REF_AREA.93=DE&REF_AREA.93=DK&REF_AREA.93=EE&REF_AREA.93=ES&REF_AREA.93=FI&REF_AREA.93=FR&REF_AREA.93=GR&REF_AREA.93=HR&REF_AREA.93=HU&REF_AREA.93=IE&REF_AREA.93=IT&REF_AREA.93=LT&REF_AREA.93=LU&REF_AREA.93=LV&REF_AREA.93=MT&REF_AREA.93=NL&REF_AREA.93=PL&REF_AREA.93=PT&REF_AREA.93=RO&REF_AREA.93=SE&REF_AREA.93=SI&REF_AREA.93=SK&PSS_INFO_TYPE.93=S101&PSS_INSTRUMENT.93=I12&MAX_DOWNLOAD_SERIES=500&SERIES_MAX_NUM=50&node=9689709&legendPub=published
Partly shadow economy determinant, partly control variable: payment card system / financial inclusion	Third party reporting available	ACCOUNT	Account, % age 15+	rather exclude, low number of observations	Data available for 2011, 2014, 2017, 2021, data interpolation was made for the remaining years	1467	11	2021	2011	160	Global Findex Database, own calculations	https://www.worldbank.org/en/publication/globalfindex/Data
Partly shadow economy determinant, partly control variable: payment card system / financial inclusion	Third party reporting available	OWNING_CARDS	Owns a debit or credit card, % age 15+	rather exclude, low number of observations	Data available for 2011, 2014, 2017, 2021, data interpolation was made for the remaining years	1467	11	2021	2011	160	Global Findex Database, own calculations	
Partly shadow economy determinant, partly control variable: payment card system / financial inclusion	Third party reporting available	USING_CARDS	Used a debit or credit card, % age 15+	rather exclude, low number of observations	Data available for 2011, 2014, 2017, 2021, data interpolation was made for the remaining years	867	8	2021	2014	135	Global Findex Database, own calculations	
Partly shadow economy determinant, partly control variable: payment card system / financial inclusion	Third party reporting available	DIGITAL_PAYMENT	Made a digital payment, % age 15+	rather exclude, low number of observations	Data available for 2011, 2014, 2017, 2021, data interpolation was made for the remaining years	1037	8	2021	2014	155	Global Findex Database, own calculations	

Control variable and for interactions with selected shadow economy determinants (to show differences in the determinants' impact depending on the country's development level)	Not applicable / level of economic development	GDP_PER_CAPITA	GDP per capita based on purchasing power parity (PPP), thousands of constant 2017 international dollars		The same definition as "Gross domestic product per capita, constant prices" (purchasing power parity; 2017 international dollar) in the IMF, for which there are publicly available forecasts.	5084	27	2021	1995	194	World Bank - International Comparison Program, World Development Indicators database, Eurostat-OECD PPP Programme.	https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.KD?end=2021&start=2004
Control variable	Not applicable	DEPOSIT_INTEREST_RATE	Households deposit rate *per annum* , %	rather exclude, data not available for some European countries for which most other variables are available		3402	27	2021	1995	150	International Monetary Fund	Financial, Interest Rates, Deposit, Percent per annum: https://data.imf.org/?sk=B83F71E8-61E3-4CF1-8CF3-6D7FE04D0930&sId=1390030341854
Control variable	Not applicable	R_DEPOSIT_INTEREST_RATE	Real deposit rate, i.e. households deposit rate per annum adjusted by yoy CPI inflation rate, in %	rather exclude, data not available for some European countries for which most other variables are available	Calculated as (1+DEPOSIT_INTEREST_RATE)/(1+CPI_RATE) - 1; in our database to obtain the appropiate units: ((1+DEPOSIT_INTEREST_RATE/100)/(1+CPI_RATE/100) - 1)*100	3135	27	2021	1995	138	International Monetary Fund, own calculations	-
Control variable	Not applicable	CPI_RATE	Inflation, consumer price index, % year-on-year		The same definition as "Inflation, average consumer prices" (percent change) in the IMF, for which there are publicly available forecasts.	4998	27	2021	1995	186	Oxford Economics	https://data.oxfordeconomics.com/render-app
Control variable	Not applicable	CREDIT_GDP	Domestic credit to private sector, % of GDP			3963	26	2020	1995	185	World Bank - International Monetary Fund, International Financial Statistics and data files, and World Bank and OECD GDP estimates.	https://data.worldbank.org/indicator/FS.AST.PRVT.GD.ZS
Control variable	Not applicable	FIN_DEV	Financial development index, aggregate of financial institutions, financial depth and financial market development indices. It ranges from 0 to 1, 1=best development			4680	26	2020	1995	180	International Monetary Fund - Svirydzenka (2016)	https://data.imf.org/?sk=F8032E80-B36C-43B1-AC26-493C5B1CD33B&sId=1481126573525
Control variable	Not applicable	URBAN_POPULATION	The share of urban population in the total population, %			5805	27	2021	1995	215	World Bank - United Nations Population Division. World Urbanization Prospects: 2018 Revision	https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS
Control variable	Not applicable	AGRI_GDP	Agriculture, forestry, and fishing value added, % of GDP			5112	27	2021	1995	204	World Bank national accounts data, and OECD National Accounts data files	https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS
Control variable	Not applicable	INDUSTRY_GDP	Industry (including construction) value added, % of GDP			5098	27	2021	1995	205	World Bank national accounts data, and OECD National Accounts data files	https://data.worldbank.org/indicator/NV.IND.TOTL.ZS
Control variable	Not applicable	SERVICES_GDP	Services value added, % of GDP			5005	27	2021	1995	202	World Bank national accounts data, and OECD National Accounts data files	https://data.worldbank.org/indicator/NV.SRV.TOTL.ZS
Control variable	Not applicable	AGRI_EMPLOYMENT	The share of people employed in agriculture in the overall employment, %			4675	25	2019	1995	187	World Bank - modeled ILO estimate	https://data.worldbank.org/indicator/SL.AGR.EMPL.ZS
Control variable	Not applicable	INTERNET_ACCESS	The share of population with Internet access, %			5092	27	2021	1995	213	World Bank - International Telecommunication Union	https://data.worldbank.org/indicator/IT.NET.USER.ZS
Control variable	Not applicable	DUMMY_IND	Binary variable controlling for the effect of demonetization in India in 2016		2016=1	6293	29	2022	1995	217	Own elaboration	-
Control variable	Not applicable	DUMMY_ROU	Binary variable controlling for the credit boom in Romania starting in 2007		2007-2010=1	6293	29	2022	1995	217	Own elaboration	-
Shadow economy determinant / control variable: sociodemographic	Age	YOUNG	Share of people aged 15-34 in the working age population (15-64), %		Considered due to the client's request.	5751 (over the period 1995-2021)	72	2021	1950	213	United Nations	https://population.un.org/wpp/Download/Standard/MostUsed/
Shadow economy determinant / control variable: sociodemographic	Education	EDUCATION	considered various variables related to the number of finished school years, educational attainment, etc.	exclude, data not available after 2019 and/or missing for many countries	Considered due to the client's request.	not available (excluded after initial analysis)	not available (excluded after initial analysis)	not available (excluded after initial analysis)	not available (excluded after initial analysis)	not available (excluded after initial analysis)	Various sources considered, for example, Penn World Table (Human Capital Index), Unesco	https://www.rug.nl/ggdc/productivity/pwt/?lang=en http://data.uis.unesco.org/index.aspx?queryid=3803
Shadow economy determinant / control variable: sociodemographic	Financial condition of taxpayers (level) / level of economic development	POVERTY_WORK	Working poverty rate (employed living below US$1.90 PPP), % of employed		Considered due to the client's request. For developed economies there is no data in the ILO database, but it can be assumed that all the observations are equal to 0.	6109	32	2022	1991	188	International Labor Organization	https://ilostat.ilo.org/topics/working-poverty/
Shadow economy determinant / control variable: sociodemographic	Not applicable	MIGRATION	Net migration rate (number of immigrants minus number of emigrants in a given year), per 1,000 population		Considered due to the client's request. Data on migration stock would be better but there are issues with such data (see MIGRATION_STOCK)	5751 (over the period 1995-2021)	72	2021	1950	213	United Nations (similar/the same data as on https://www.migrationdataportal.org/)	https://population.un.org/wpp/Download/Standard/MostUsed/
Shadow economy determinant / control variable: sociodemographic	Not applicable	MIGRATION_STOCK	Number (or "stock") of international migrants, per 1,000 population	exclude: (1) data on migrants stock (cumulative number of migrants) widely available only for 5-year intervals, while data for our model needs to be available at the annual frequency, (2) there are significant differences in definitions of such data between countries	Considered due to the client's request.	not available (excluded after initial analysis)	not available (excluded after initial analysis)	not available (excluded after initial analysis)	not available (excluded after initial analysis)	not available (excluded after initial analysis)	United Nations (similar/the same data as on https://www.migrationdataportal.org/)	https://www.un.org/development/desa/pd/content/international-migrant-stock
Shadow economy determinant / control variable: sociodemographic	Not applicable	LIFE_EXP	Life expectancy at birth, total (years)		Considered due to the client's request. In our opinion this variable is better than the proposed variable related to deaths at birth, since it captures health outcomes (including death rates) for the whole age distribution.	5540	27	2021	1995	212	World Bank (United Nations)	https://data.worldbank.org/indicator/SP.DYN.LE00.IN
Shadow economy determinant / control variable: sociodemographic	Not applicable	HOSPITAL_BEDS	Hospital beds, per 1,000 people	exclude, data not available in many cases, especially for more recent time periods	Considered due to the client's request.	3099	27	2019	1995	178	World Health Organization	https://www.who.int/data/gho/data/indicators/indicator-details/GHO/hospital-beds-(per-10-000-population)
Shadow economy determinant: labour market / business cycle	Shock to financial condition	GDP_GROWTH	Gross domestic product, constant prices, percentage change from previous year		Considered due to the client's request. Similar aspect (plus level of income) is already included in the GDP_PER_CAPITA variable.	5145 (over the period 1995-2021)	47	2022	1980	196	International Monetary Fund	https://www.imf.org/en/Publications/WEO/weo-database/2022/October
Shadow economy determinant / control variable: other	Not applicable	IMPORTS	Volume of imports of goods and services, % change from previous year		Considered due to the client's request. Since we analyse determinants of currency in circulation used for transactions of residents, considering imports (rather than exports) seem to make more sense. Theoretically, instead of (a) volume of imports (percent change, y/y) it may make more sense to consider (b) imports (% of GDP), because (b) captures not only the dynamics of imports but also their relative role in the economy ((b) could also be forecasted with publicly available sources).	4588 (over the period 1995-2021)	47	2021	1980	178	IMF	https://www.imf.org/en/Publications/WEO/weo-database/2022/October

Source: EY.

1. Public information – TLP-WHITE

A1.3 Data preparation

Apart from some basic operations described in table with our dataset (e.g. dividing some variables by GDP or population size to make them comparable between countries and over time), we also took additional measures to increase the sample size and improve the dataset quality.

- **Currency in circulation for the eurozone**. We decomposed the cash in circulation in the whole eurozone into the values for each of the euro area members (such estimates are not publicly available[85]). The decomposition is based on the value of cash withdrawals from ATMs in each euro area member state collected from the European Central Bank database. We have assumed that the shares of euro area members in such withdrawals in the eurozone are the same as their shares in the currency in circulation in the euro area.
- **Interpolations**. When for the given variable and country values were missing between periods with available data points, we used simple interpolation techniques to complete the time series (e.g. Worldwide Governance Indicators data did not include observations for 1995, 1997, 1999 and 2000 years and for the last three of them it allowed us to estimate missing values).
- **Outliers**. We corrected or cleaned some outliers in the data. One of the errors identified in the original data source was incorrect units for Belarus and Zambia in the data of Currency Outside Banking Institutions (used to create the dependent variable CASH_M1). We have also removed doubtful observations for Romania in 1995 for the variable CREDIT_GDP.
- **Countries selection**. After an initial investigation, we dropped specific countries from the analysis (the list of countries and reasons in the footnote[86]). Finally, we generated a common sample (a fixed set of countries and time periods) in order to effectively compare the models with different sets of variables (otherwise changes in the obtained results would be a mix of the variables impact and changes in the sample composition resulting from the selection of different variables).

A1.4 Method for estimation of the econometric model

Even for a given set of variables, there are different econometric methods of estimation (so called estimators) of unknown parameters that describe the relationship between the explanatory and explained variables (coefficients) as well as the measure of their uncertainty or variability (standard errors). The choice of the estimator should be based

[85] Among the public data we can find a variable with such a name, but it was calculated as countries' shares in the European Central Bank's capital. In such data all the eurozone members show exactly the same percent growth of currency in circulation over time, despite the fact that the actual trends in this area could be different among them.
[86] We excluded countries with substantial amount of missing data or questionable/outlying data (due to wars, high inflation, low quality of data collection or very specific conditions in the given country): Afghanistan, Angola, American Samoa, Andorra, Antigua and Barbuda, Aruba, Bahrain, Barbados, Belize, Bermuda, Bolivia, Brunei Darussalam, Bhutan, Cambodia, Cayman Islands, Central African Republic, Channel Islands, Congo, Dem. Rep., Comoros, Cuba, Curacao, Djibouti, Dominica, Ecuador, Egypt, Arab Rep., Equatorial Guinea, Eritrea, Ethiopia, Faroe Islands, Fiji, French Polynesia, Gibraltar, Greenland, Grenada, Guam, Guinea-Bissau, Iran, Iraq, Isle of Man, Kiribati, Kosovo, Kyrgyz Republic, Liberia, Liechtenstein, Maldives, Marshall Islands, Micronesia, Monaco, Mongolia, Namibia, Nicaragua, Northern Mariana Islands, New Caledonia, Nauru, Qatar, Palau, People's Republic of Korea, Puerto Rico, Rwanda, Samoa, Sao Tome and Principe, San Marino, Saudi Arabia, Seychelles, Sint Maarten (Dutch part), Solomon Islands, Somalia, South Sudan, St. Kitts and Nevis, St. Lucia, St. Martin, St. Vincent and the Grenadines, Syrian Arab Republic, Timor-Leste, Tonga, Turks and Caicos Islands, Turkmenistan, Tuvalu, United Arab Emirates, Uzbekistan, West Bank and Gaza, Vanuatu, Venezuela, Virgin Islands (U.S.), Yemen (Rep.), Zimbabwe

on various characteristics of the analysed dataset that are discussed below. In general, such characteristics have been similar for different combinations of variables considered in our analysis. Therefore, we first chose the estimator based on a few initial sets of variables and then applied the same rule of estimating the coefficients to different set of variables.[87]

Having a ready set of data, we have performed a series of statistical tests. First, we have verified if there exists a problem of heteroskedasticity (i.e. we can observe changes in the variance of errors from the model across different countries) on the basis of likelihood ratio test, where we compared the likelihood of the model estimated using Feasible Generalized Least Squares (FGLS) estimator that takes into account heteroskedasticity with a simple Least Squares model. The results showed that there exists heteroskedasticity. Second, we have performed a serial correlation test[88] that showed that autocorrelation (i.e. correlation of errors from the model) is also present. Those results indicate that we have to use the family of FGLS estimators that take into account presence of heteroskedasticity and autocorrelation. Third, we have performed the Hausmann test that indicated that we should use the fixed effects (i.e. binary variables representing each country that take into account specific characteristics of each country included in the panel dataset).

Finally, we chose 'Panel-Corrected Standard Error' (PCSE) estimator as it accounts for the heteroskedasticity and autocorrelation, produces stable results (regarding exclusion of random countries or changes in the specification) and provides reliable evaluation of standard errors of each parameter[89]. We selected panel-specific autocorrelation structure option (psar1) which identifies that there is first-order autocorrelation and its coefficients are specific to each country.

The tool that we used to conduct the investigation of the currency demand and the shadow economy is Stata software[90], as it is well-designed for the econometric analysis of panel data. In particular, it has a well-programmed function for estimation of the model's coefficients with the PCSE estimator. For the part related to the selection of the variables (BMA analysis), which will be described in the next subsection, we used the R software[91], as it is faster and better suited to this type of analysis.

A1.5 Initial selection of variables

In preliminary part of the analysis we apply Bayesian model averaging (BMA) procedure in which a wide array of variants of CDA model is estimated using the PCSE method, with different combinations of variables from Table A.2 (for their detailed description see Table A.1). The goal of this procedure is to estimate the Posterior Inclusion Probability (PIP) of each variable[92], that is a measure indicating which variables should be included in the model.

It needs to be pointed out that the total number of combinations of models is equal to 2^k, where k is the number of considered variables. Since we have a preliminary list of more than 40 variables (after excluding some variables for which not enough data is

[87] In practice, while conducting the econometric analysis, we looked also at some additional estimators to observe the robustness of our analysis to a different choice of estimator.
[88] See Drukker, D. M. (2003), Testing for serial correlation in linear panel-data models. Stata Journal 3, pp. 168–177.
[89] For a discussion of a selection of the estimator for panel data setting see: Reed W.R. & Ye H. (2011), Which panel data estimator should I use?, Applied Economics, 43:8, pp. 985-1000
[90] For the estimation we used version Stata/IC 16.0 for Windows (64-bit x84-64)
[91] R version 3.5.3
[92] For BMA application in the context of shadow economy estimation see: Dybka, P., Olesiński, B., Rozkrut, M. and Torój, A. (2022), Measuring the model uncertainty of shadow economy estimates, International Tax and Public Finance, https://doi.org/10.1007/s10797-022-09737-x

available) we had to use some additional assumptions to further decrease the number of analysed models. As such we have divided variables into groups consisting of variables that represent similar concepts of shadow economy determinants:

- Worldwide Governance Indicators (WGI) that measure general level of institutional quality
- Other institutional and regulatory indicators
- Labour market structure indicators
- Business cycle indicators
- Taxation level indicators
- Other social factors that can affect the shadow economy development

We have assumed that in each analysed model there can only be one variable from each of the groups. This assumption has decreased substantially the number of potential combinations. Moreover, we have also assumed that in each model there must be at least one shadow economy determinant and one control variable (measuring the demand for cash used in legal transactions). We have observed that some variables kept very low levels of PIP so we have excluded them in the initial iterations of the BMA analysis. The overall number of models analysed in the final iteration of the BMA analysis amounted to 737 152 models. It is worth noting that our initial extensive analysis of millions of models with different combinations of variables with the Bayesian model averaging techniques (BMA) will not be needed in the future while re-estimating the CDA model, since many variables that performed very poorly in this approach are not likely to become relevant for the shadow economy in the future.

In addition to this, we have also imposed sign restrictions on each of the shadow economy determinant. For example, an increase in the institutional quality measure should decrease the shadow economy level. As a result, each WGI variable should have a negative sign. If there is a positive sign in the analysed model it indicates a problem regarding estimation (e.g. due to omitting important variables in the specification, the so-called “omitted variable bias”) and therefore we excluded such model from the analysis in the “restricted” variant of the BMA analysis.

Table A.2 – Summary of the BMA analysis

Variable name	PIP	PIP (sign restrictions)
Worldwide Governance Indicators		
GOV_EFFECTIVENESS	46.3%	70.8%
REGULATORY	7.0%	10.8%
POLITICAL	6.8%	4.6%
CONTROL_OF_CORRUPTION	6.8%	2.9%
VOICE_AND_ACCOUNTABILITY	8.0%	0.0%
RULE_OF_LAW	17.8%	0.0%
Probability that a variable from this group should be included:		89.1%
Other institutional and regulatory indicators		
INTEGRITY	58.1%	54.7%
COURTS	24.8%	24.6%
CONTRACTS_ENFORCEMENT*	23.1%	0.2%

BUSINESS_REGULATIONS*	0.0%	0.0%
LABOR_MARKET_REGULATIONS*	0.0%	0.0%
REGULATORY_BURDEN*	54.1%	0.0%
Probability that a variable from this group should be included:		79.3%
Labour market structure indicators		
FAMILY_WORK	100.0%	100.0%
SELF_EMPLOYED*	0.1%	0.1%
OWN_ACCOUNT_WORK*	0.0%	0.0%
Probability that a variable from this group should be included:		100.0%
Business cycle indicators		
UNEMP	79.8%	80.7%
NON_EMPLOYED	19.2%	18.7%
GDP_GROWTH	0.5%	0.3%
Probability that a variable from this group should be included:		99.7%
Taxation level indicators		
CIT	25.2%	45.9%
VAT	11.8%	20.7%
PIT	9.4%	17.1%
TAXES_INCOME_PROFITS_GAINS	4.7%	7.8%
TAXES_GOODS_AND_SERVICES	44.1%	0.0%
Probability that a variable from this group should be included:		91.6%
Other social factors		
YOUNG_LABOR_FORCE	62.2%	68.8%
LIFE_EXPECTANCY	14.7%	17.1%
MIGRATION_NET	7.2%	6.3%
POVERTY_WORK	8.2%	0.2%
Probability that a variable from this group should be included:		92.4%
Control variables		
GDP_PER_CAPITA	100.0%	100.0%
INTERNET_ACCESS	100.0%	100.0%
AGRI_GDP	98.0%	96.7%
CPI_RATE	92.7%	91.1%
IMPORTS	69.7%	72.4%
URBAN_POPULATION	59.8%	59.5%
CREDIT_GDP	55.2%	53.6%
GDP_PER_CAPITA_squared	48.1%	48.1%

Notes: *We have conducted two iterations of the BMA analysis, as we have reached such a large number of potential models that we were unable to evaluate them all at once. After the first BMA iteration we have removed variables that had a very low PIP or had a wrong sign and then we have run additional BMA iteration where we have added some new variables (e.g. “Other social factors group”). We denote variables removed after the first BMA iteration with asterisk (*).

Source: EY.

We can observe that the sum of inclusion probabilities of variables in most of the groups of shadow economy determinants is close or above 90% presenting strong evidence that a variable from the group should be present in the final model. In the case of the Other institutional and regulatory indicators group the probability is almost 80%, which still present substantial evidence[93] that a variable from that group should be considered in the final model.

Among the Worldwide Governance Indicators, the GOV_EFFECTIVENESS variable has the highest Posterior Inclusion Probability and should be considered in the final model. Moreover, the INTEGRITY variable is the best candidate for the final model from the other institutional and regulatory indicators group, whereas FAMILY_WORK should be used as a variable measuring the structure of the labour market (i.e. the share of a potentially vulnerable workers) and the UNEMP should be viewed as the variable measuring the effects of a business cycle (also general state of the labour market) on the shadow economy.

In the case of taxation level indicators, the case is less clear. To begin with, we can observe that the effective rates (i.e. variables based on the value of actually collected taxes) often have a wrong sign (due to potential endogeneity issues) - TAXES_GOODS_AND_SERVICES (measuring the income from VAT/sales tax to value added ratio) has a PIP equal to 0 after imposing restriction that its sign should be positive (i.e. higher taxes mean higher shadow economy level). Moreover, the TAXES_INCOME_PROFITS_GAINS measuring the ratio of CIT (and other entrepreneurial income taxes) ratio to value added also has a low PIP value. As a result, we conclude that the nominal rates should be considered in the final model, where the CIT rate has the highest PIP. Since VAT and PIT also show a substantial (compared to CIT) inclusion probabilities, we would also consider a simple average of CIT, VAT and PIT nominal rates in the final model.

The last group includes other types of social variables that can affect the shadow economy. In this group one demographic factor, namely share of young (15-34) people in the working age population (15-64) has the largest Posterior Inclusion Probability and should also be considered in the final model.

In the case of the so-called control variables (that account for factors affecting demand for cash that are not related to shadow economy), we did not impose any restrictions. Obtained results indicate that the most likely candidates for the final model include GDP_PER_CAPITA, INTERNET_ACCESS, AGRI_GDP and CPI_RATE. It is worth pointing out that for all control variables except the squared GDP_PER_CAPITA, the PIP was above 50% indicating there exists some evidence that such variables should be considered in the final model.

To our best knowledge, such thorough analysis of variables in the currency demand model has not been previously done in the economic literature by other researchers.

[93] For a discussion on the interpretation of PIP see e.g. Bierut B.K., Dybka P., (2021) Increase versus transformation of exports through technological and institutional innovation: Evidence from Bayesian model averaging, Economic Modelling, Vol. 99, 105501, https://doi.org/10.1016/j.econmod.2021.105501.

A2. Unregistered income and the PIT gap

A2.1 Data preparation

We introduced some transformations to the initial dataset such as:

- Recoding variables from numeric variables to descriptive ordinal variables and aggregating some ordinal variables (e.g. industry, completed education, settlement size) to a smaller number of categories to minimize the number of parameters estimated in econometric models;
- Removing households for which tax data of at least one member aged 18 or over was not identified. We assumed that underaged who were not identified did not generate any income;
- Filling missing values of some variables if other relevant data was provided;
- Aggregating data from the individual level to the household level as the econometric model corresponding to traces-of-true income approach of Pissarides and Weber (1989) is based on household-level;
- Assigning to each household a reference person, for whom individual-level socio-demographic variables were used in econometric analysis. A reference person can be either (1) the person indicated in the HBS as a **household head** or (2) the **primary earner**, i.e. the person with the highest net income in the household according to the data from tax returns.

We decided to estimate one model on a four-year sample in order to maximize the number of observations, which is particularly important for the possibility of differentiating the non-reporting scale by socio-demographic factors. This procedure of pooling several waves of the HBS in order to perform the traces-of-true-income analysis is also frequently used in the literature. Therefore, as our dataset contained four years pooled together, all household-level monetary data was converted from nominal prices to real prices using data on HICP from Eurostat – all monetary values were expressed in 2021 prices. In order to additionally control for the impact of differences between years on the modeling results, we added time effects (time specific dummy variables) to specification of our econometric models. In addition, we identified one outlier, i.e. a household with net labour income of more than one million BGN in 2019. As this observation significantly influenced the calculation of the average income in 2019 and we did not have similar observations in other years, we decided to apply winsorization, i.e. we replaced the net labour income for this household with the second highest value in our sample.

A2.2 Classification of households to the traces-of-true-income analysis and econometric model

Pissarides and Weber (PW) approach to estimating the level of income underreporting is based on comparing the relationship between food expenditure and income of the non-compliant group of workers to that of the reference group of workers that is assumed to be fully compliant. In our analysis, we want to estimate the level of underreporting among **(1) private sector workers** and – separately – among **(2) self-employed** in Bulgaria while the reference group is comprised of **(3) public sector workers**. We considered two different ways of classifying households into those **three sectors**:

1. Based on the share of household income from each of these three sources

2. Based on the source of income of household primary earner

Finally, we decided on the first method, which is consistent with the one used in the PW (1989) analysis. The second method would give us a smaller number of households in the self-employed group, for which we already have relatively few observations in our sample (a larger number of observations increases the stability and credibility of the results from econometric models).

We adopted the following criteria to classify each household into a sector:

- **A household is classified as self-employed household** (NRA_sectors_3_s = "Self-employed") **if the share of household income from self-employed members amounts to at least 25%.** The 25% criterion is consistent with the literature standard started with the work of PW (1989).
- **A household is classified as private-sector employee household** (NRA sectors 3 s = "Private sector employee") **if the share of household income from self-employed members is less than 25% and the share of household income from private sector employees added to the share of household income from self-employed members is higher than 0%.**

  Initially, we tested models where this group of households was additionally broken down into two sub-groups:

  - Households for which the share of household income from self-employed members is equal to 0% and the share of household income from private sector employees amounts to at least 25%
  - Households for which the share of household income from self-employed members is less than 25%, the share of household income from private sector employees is less than 25% and the combined share is higher than 0%[94]

  As it turned out that the estimated parameters related to non-reporting were similar for those sub-groups (however, they differed significantly from the parameter estimated for the self-employed group), we decided to combine these sub-groups into one.
- **A household is classified as public-sector employee household** (NRA_sectors_3_s = "Public sector employee") **if the share of household income from public-sector employees is equal to 100%.** In this way, we assume that there are no private-sector workers in the reference group of households.

In this way, one of three sectors was classified for each household with positive net income except one – a household in which income was generated by a public sector employee and by a person with unidentified source of income, excluded from the analysis.

In line with the PW model, only households with positive income and positive food expenditures can be included in the analysis. In addition, it is standard in literature to restrict the sample to households in which the primary earner works full time (e.g.

[94] Inclusion in the analysis of as many households that may underreport income as possible was important for us from the perspective of calculating unreported income at the macro level (see section A2.4). Therefore, we did not want to exclude from the model the households with low shares of income from work as private sector employee or self-employed.

Paulus, 2015[95]) or in which labour income is the main source of household income (e.g. Kukk, Paulus and Staehr 2020[96]). Accordingly, we decided to exclude from the econometric model those households in which net income (based on the NRA) is not the main source of regular household income. Other sources of regular household income were based on the HBS data and included (1) household income from social security benefits and (2) other regular income of households (e.g. child allowances). In this way, we excluded from the econometric analysis 24.1% of households with positive reported net income.

Table A.3 – Key characteristics of households by years and assigned sector

a) Initial sample, i.e. received dataset after excluding households with missing information or reported labour income = 0

Assigned sector	year	Avg net income	% SELF	% PRIVATE	% PUBLIC	Avg food exp	N	Sum of weights
Public sector employee	2017	11687.9	0.0	0.0	100.0	3654.3	284	275787.0
Public sector employee	2018	12563.7	0.0	0.0	100.0	3696.5	294	304132.6
Public sector employee	2019	13881.8	0.0	0.0	100.0	3926.5	341	348150.9
Public sector employee	2021	16620.7	0.0	0.0	100.0	4153.9	338	321157.1
Private sector employee	2017	13580.6	0.5	86.4	13.1	4103.2	1222	1245541.1
Private sector employee	2018	14791.6	0.5	84.3	15.2	4153.2	1188	1238471.7
Private sector employee	2019	16432.6	0.4	87.2	12.4	4230.7	1161	1210277.8
Private sector employee	2021	17488.7	0.3	87.0	12.7	4670.9	1157	1147116.1
Self-employed	2017	8100.5	88.1	9.4	2.4	3649.8	165	167788.1
Self-employed	2018	8417.5	89.6	7.5	2.8	3545.2	134	138801.8
Self-employed	2019	10151.0	85.2	11.8	3.1	3809.7	120	133640.1
Self-employed	2021	7440.1	86.6	11.1	2.4	4338.4	122	132057.1

b) Estimation sample = initial sample after excluding 24.1% of households in which labour income was lower than other regular sources of income

Assigned sector	year	Avg net income	% SELF	% PRIVATE	% PUBLIC	Avg food exp	N	Sum of weights
Public sector employee	2017	13693.9	0.0	0.0	100.0	3635.8	220	216204.9
Public sector employee	2018	14753.8	0.0	0.0	100.0	3726.0	232	243566.1
Public sector employee	2019	16617.5	0.0	0.0	100.0	3934.7	262	270640.6
Public sector employee	2021	20175.2	0.0	0.0	100.0	4313.6	255	247641.2
Private sector employee	2017	16119.0	0.6	83.9	15.5	4142.3	962	996266.8
Private sector employee	2018	17191.7	0.4	81.5	18.0	4256.7	961	1018163.8
Private sector employee	2019	19626.7	0.5	85.0	14.5	4339.7	910	967469.2
Private sector employee	2021	20680.9	0.3	84.7	15.0	4789.4	925	917941.2
Self-employed	2017	15322.5	77.5	17.6	4.9	3964.8	67	73450.4
Self-employed	2018	16490.3	79.4	14.5	6.1	4044.5	53	59859.0
Self-employed	2019	19974.7	68.6	24.3	7.1	4718.6	52	57829.1
Self-employed	2021	13063.3	72.5	22.6	5.0	4751.7	56	63435.6

Notes: All monetary values are expressed in BGN in 2021 prices. N = number of observations. All averages are weighted using sample weights.

Source: EY.

In Table A.3 we summarize key characteristics of our sample after classifying households into sectors. In the initial sample of all household with positive net income, average net income was the highest among households classified as private sector employee households and the lowest in households classified as self-employed households. However, especially in the self-employed group, there were many households with very low income declared to the NRA, which affected the average. After excluding 24.1% of households in which labour income was lower than other regular sources of income, averages of net income increased significantly. For the self-employed households they became closer to those of the private sector employee households with the exception of 2021 when net income of households classified as self-employed decreased significantly.

What particularly caught our attention is that the average net labour income of individuals in our sample (weighted using survey weights) was lower than the average

[95] Paulus, A. (2015). Income underreporting based on income expenditure gaps: Survey vs tax records (No. 2015-15). ISER Working Paper Series.
[96] Kukk, M., Paulus, A., & Staehr, K. (2020). Cheating in Europe: underreporting of self-employment income in comparative perspective. International Tax and Public Finance, 27(2), 363-390.

net labour income in the whole economy (based on the macro-level data provided by the NRA). We suspect that this is related to the fact that the wealthiest people in the country, whose income has a significant impact on the average, are underrepresented in the Household Budget Survey.[97] It is also suggested by the fact that the averages for income in the public sector, where wages may be more evenly distributed, were similar between the survey data and macro data, while the greatest disparities were for the self-employed. Consequently, the results of econometric Pissarides-Weber model may not be fully representative for the whole Bulgarian economy (if the scale of underreporting of the wealthiest households is lower than average, their exclusion introduces upward bias in the scale of underreporting estimated from the econometric model).

Next, the table shows the average shares of the three considered sources of income by years and sectors. In line with our assumptions, households classified as public sector employee households generate 100% of their labour income from work in the public sector. About 84% of labour income in households classified as private sector employee households comes from employment in the private sector whereas about 75% of labour income in households classified as self-employed households comes from the work of self-employed members. Average expenditure on food is the lowest in households classified as public sector employee households in every year and the highest in households classified as private sector employee households in 2017, 2018 and 2021 while in 2019 self-employed households spent the biggest amount on food, on average. In the estimation sample, there are 969 households classified as PUBLIC, 3,758 households classified as PRIVATE and 228 households classified as SELF. The sum of survey weights can be interpreted as the sum of similar households in Bulgaria (it is a result of application of the two-stage cluster sampling in the HBS).

The sample in the original PW model was further restricted to households of two adults. This assumption was then adopted by other authors drawing on PW methodology, or expanded to include households with at least two adults (e.g. Turgut and Tratkiewicz 2023[98]). The reason why the sample is often limited in this way is to ensure that households are as similar as possible so that differences in their composition do not affect the conclusions regarding the level of non-compliance. However, as we want to draw conclusions about the level of income underreporting and the corresponding PIT gap in the entire Bulgarian economy based on our model, eliminating one-person households from the analysis also raises doubts as to the representativeness of the study. Moreover, it reduces the number of observations in the econometric model, which is most problematic when testing the interactions of the classification variable with other socio-demographic variables. Therefore, we decided not to restrict the sample in this way, but when presenting the results, we compare those from the (1) base model estimated on the sample with all possible household compositions with the results for the models estimated for the sample limited to households with (2) two adults and (3) at least two adults.

[97] In general, weights in the survey are constructed based on the probability of selecting a given household for the study, which in the cluster method consists in constructing weights based on the number of households in a given cluster (territorial unit). The weights can be later adjusted (if post-stratification is performed), however, they are usually not adjusted for household income in the HBS studies. Therefore, the weighted HBS shares are probably representative for the number of households in the population, but not necessarily for their total spending.
[98] Turgut, M. B., & Tratkiewicz, T. (2023). Estimate of the Underground Economy in Poland Based on Household Expenditures and Incomes. Central European Journal of Economic Modelling and Econometrics, 1-29

A2.3 Method for estimation of the econometric model

The methodology used in our study heavily relies on the "traces-of-true-income" approach developed by Pissarides and Weber (1989), which is outlined below. In the following section, we describe the details of the PW model and then delve into the specifics of our study.

Pissarides-Weber methodology (PW model)

The derivation of equations of the PW model is crucial to understanding why a specific estimation procedure is used and how to interpret the underreporting parameters: scaling factor k, income gap IG and their ranges. This section relies on methodological notes in the works of Pissarides, C. A., & Weber, G. (1989)[99] and Kukk, Paulus and Staehr (2020)[100].

In the PW original framework, all employees (working in the public or in the private sector) were treated as the reference group while the scale of underreporting was estimated for the self-employed.[101] For the sake of simplicity, we will stick to this notation in this section. However, in our analysis, the reference group will be households classified as public sector employee households, and the parameters k and IG will be estimated separately for households classified as private sector employee households and the self-employed households.

The discrepancy between reported income ($Y_i^{registered}$) and true income (Y_i^{True}) can be formally expressed as:

$$Y_i^{True} = k_i Y_i^{registered}, \qquad k \geq 1, \tag{1}$$

where k_i represents the extent of underreporting by a given household i. In the PW model, it is assumed that there is no discrepancy for the employees in employment ($k = 1$), but self-employed can underreport their true income ($k \geq 1$):

$$Y_i^{True} = \begin{cases} k_i Y_i^{registered} & if\ self-employed \\ Y_i^{registered} & if\ employee \end{cases} \tag{2}$$

The value of k or Y_i^{True} cannot be directly observed, hence indirect methods are needed to estimate it.[102] Pissarides and Webber propose to use the coefficients from the Engel curve regression to estimate the extent of underreporting. The Engel curve formula relates the household spending (in the PW model food expenditures are used as they are considered to be relatively well measured in surveys) to the household income:

$$\log(C_i) = \alpha + \beta \log\left(Y_i^P\right) + X_i\alpha + \epsilon_i, \tag{3}$$

where:

α - constant term

[99] Pissarides, C. A., & Weber, G. (1989). Ibid.
[100] Kukk, M., Paulus, A., & Staehr, K. (2020). Cheating in Europe: underreporting of self-employment income in comparative perspective. International Tax and Public Finance, 27(2), 363-390.
[101] See section 3.7.3 of the methodological report for the description of extensions to the original PW model from which our analysis draws.
[102] The earning function for employees and self-employed individuals is likely to be different, and therefore, we cannot use a direct regression on income to directly estimate under-reporting parameter. The consumption function is more likely to be similar across groups.

C_i- food expenditure of household i
Y_i^P- permanent income of household i

β - the elasticity of consumption with respect to income

$X_i\alpha$ - control variables and their corresponding parameters

ϵ_i- white noise error term

The authors assumed that food spending is influenced by true income and socio-demographic factors, but not by employment status. Hence, if food expenditure is higher for self-employed than employees with the same income level it suggests underreporting of income by self-employed.

It is important to note the choice of income measure used in the regression formula. Income consists of a permanent (expected) component and a transitory (unexpected) component. According to the economic theory, the permanent income, defined as the average income a household can expect to receive over a long period of time, is a better predictor of consumption behavior. This is because permanent income provides a more stable and reliable measure of a household's economic resources and takes into account consumption smoothing strategies (Campbell and Mankiw, 1990[103]). The permanent and observed labor income are related by:

$$Y_i^{True} = p_i Y_i^P \tag{4}$$

where p_i is a random variable, which represents the extent to which a household's actual income in a given time period differs from its expected or permanent income level.

Taking the logarithm of equations (1) and (4) and combining them allows us to express permanent income using a single formula, which demonstrates two sources of bias entering the parameter β in equation (3):

$$\log\left(Y_i^P\right) = \log\left(Y_i^{registered}\right) - \log(p_i) + \log(k_i). \tag{5}$$

In the PW model, it is assumed that both k_i and p_i follow log-normal distribution, and can be expressed as:

$$\log(p_i) = \mu_p + u_i, \qquad E(u_i) = 0, \qquad Var(u_i) = \sigma_u^2, \tag{6}$$

$$\log(k_i) = \mu_k + v_i, \qquad E(v_i) = 0, \qquad Var(v_i) = \sigma_v^2. \tag{7}$$

The no-underreporting assumption of employees (2) implies that their reporting rate variance $(\sigma_{v,EE}^2)$ is equal to zero and $\log(k_i)$ is equal to zero. Underreporting can however occur in the self-employed group and we expect both $\mu_{k,SE}$ and $\sigma_{v,SE}^2$ to be positive. By assumed log-normality, the respective means for each group are:

$$\log\left(\bar{k}_{SE}\right) = \mu_k + \frac{1}{2}\sigma_{v,SE}^2 \tag{8}$$

$$\log\left(\bar{k}_{EE}\right) = 0 \tag{9}$$

[103] Campbell, John Y., and N. Gregory Mankiw. (1990) "Permanent income, current income, and consumption." Journal of Business & Economic Statistics 8.3: 265-279.

$$\log(\bar{p}_{SE}) = \mu_{p,SE} + \frac{1}{2}\sigma^2_{u,SE} \tag{10}$$

$$\log(\bar{p}_{EE}) = \mu_{p,EE} + \frac{1}{2}\sigma^2_{u,EE} \tag{11}$$

Pissarides and Webber argue that each group is characterized by the same mean of p_i, denoted by $\bar{p}_i$. Under this assumption, we can derive the relation of distribution parameters between the groups:

$$\mu_{p,SE} - \mu_{p,EE} = -\frac{1}{2}\left(\sigma^2_{u,SE} - \sigma^2_{u,EE}\right) \leq 0 \tag{12}$$

The assumption of unequal variances ($\sigma^2_{u,SE} \geq \sigma^2_{u,EE}$) leads to a difference between the means of the log of p_i. Those results will be later used to obtain the mean under-reporting factor.

By substituting equation (6) and equation (7) into equation (5), we can express the permanent income as:

$$\log\left(Y_i^P\right) = \log\left(Y_i^{registered}\right) + \left(\mu_k - \mu_p\right) + (v_i - u_i). \tag{13}$$

The equation for permanent income can now be substituted into the Engel curve formula (3):

$$\log(C_i) = \alpha_0 + \beta\log\left(Y_i^{registered}\right) + \beta\left(\mu_k - \mu_p\right) + \beta(v_i - u_i) + \alpha X_i + \epsilon_i \tag{14}$$

The term $\beta\left(\mu_k - \mu_p\right)$ can be replaced by γSE_i, where SE_i is a dummy variable indicating a household with self-employed. We can also replace $\beta(v_i - u_i) + \epsilon_i$ into one random variable of zero mean η_i. This leads to the final regression formula:

$$\log(C_i) = \alpha_0 + \beta\log\left(Y_i^{registered}\right) + \gamma SE_i + \alpha X_i + \eta_i\,. \tag{15}$$

The income variable $Y_i^{registered}$ is treated as endogenous, meaning that it is correlated with the error term (it follows directly from the fact that $\beta(v_i - u_i)$ enters the error term). The above equation is therefore estimated using two-stage least squares (2SLS) which is the estimator commonly used for instrumental variable estimation that deals with endogenous explanatory variables. This serves two purposes: (i) obtaining an unbiased estimate of β, and (ii) obtaining an estimate of the income variance for each group, which will be discussed in more detail later.

The equity of $\beta\left(\mu_k - \mu_p\right)$ and γSE_i yields the equation (16). Note that deriving the expression requires both the assumption of no-underreporting of the reference group (employees in the original PW framework) and the assumption of unequal variances of reported income between the groups.

$$\gamma = \beta(\mu_k - \frac{1}{2}(\sigma^2_{u,SE} - \sigma^2_{u,EE})) \tag{16}$$

By assumed log-normality of k_i, the parameter of interest - the mean scaling factor for the self-employed $\bar{k}_{SE}$ - is given by:

$$\log(\bar{k}_{SE}) = \mu_k + \frac{1}{2}\sigma^2_{v,SE} \tag{17}$$

Substituting (8) in the above formula, we derive:

$$\bar{k}_{SE} = \exp\left(\frac{\gamma}{\beta} + \frac{1}{2}\left(\sigma^2_{v,SE} + \sigma^2_{u,EE} - \sigma^2_{u,SE}\right)\right) \tag{18}$$

The first term of the inner sum can be obtained using the estimated regression. However, the variances involved in the calculation of the scaling factor resulting from underreporting are not observed, which means that it is not possible to calculate it exactly. To address this issue, PW propose a method for calculating a range of values within which the mean scaling factor ($\bar{k}_{SE}$) is likely to lie. The approach requires estimates of the total income variance for each group, which are obtained from the first stage of the 2SLS estimation method:

$$\log(Y_i^{registered}) = \delta_0 + \delta_1 Z_i + \delta_2 X_i + \zeta_i. \tag{19}$$

where X_i is a set of control variables (the same in both stages), Z_i is a set of instrumental variables, δ_i are respective parameters. The error term ζ_i is again a combination of three random variables: (i) unexplained variation in permanent income ε_i, (ii) deviations of true from permanent income, u_i, (iii) deviations of registered from true income, v_i. The first-stage regression is estimated under the assumption of unequal variances in each group. We can express the variances of the error term ζ_i as a composite of the three variables:

$$\sigma^2_{\zeta,SE} = var(\zeta_{SE}) = var(u_{SE} - v_{SE} + \varepsilon_{SE}), \tag{20}$$

$$\sigma^2_{\zeta,EE} = var(\zeta_{EE}) = var(u_{EE} - v_{EE} + \varepsilon_{EE}). \tag{21}$$

Furthermore, we make the assumption that the variance of permanent income is equal for both groups ($\varepsilon_{SE} = \varepsilon_{EE}$), and that ε_i is independent of both v_i and u_i. Given that v_{EE} is equal to zero, we can express the difference between the variances of the error term as follows:

$$\sigma^2_{\zeta,SE} - \sigma^2_{\zeta,EE} = \sigma^2_{u,SE} + \sigma^2_{v,SE} - 2Cov(u_{SE}, v_{SE}) - \sigma^2_{u,EE}. \tag{22}$$

The above relation links the error term in the first stage regression to the unobserved random components. The estimates of residual variance are solely insufficient to retrieve the underreporting parameter. However, if we assume that there is no relationship between the deviation of true income from permanent income u_{SE} and the deviation of reported income from true income v_{SE} (23)[104], we can use the equation to derive lower and upper bound for the parameter $\bar{k}_{SE}$. Moreover, we consider $\sigma^2_{u,EE}$ as a parameter.

$$u_{SE} \perp v_{SE} \Rightarrow Cov(u_{SE}, v_{SE}) = 0 \tag{23}$$

Under the above assumptions $\sigma^2_{v,SE}$ and $\sigma^2_{u,SE}$ are negatively related. When the former increases, the latter decreases, and vice versa. The parameter $\bar{k}_{SE}$ (18) is at its minimum level when $\sigma^2_{v,SE}$ takes its lowest value, which is zero. This implies that the

[104] PW shows that a small positive correlation between variables has little effect on the estimate of $\bar{k}_{SE}$.

underreporting rate is constant across all individuals. The expression (22) simplifies to:

$$\sigma^2_{\zeta,SE} - \sigma^2_{\zeta,EE} = \sigma^2_{u,SE} - \sigma^2_{u,EE}. \tag{24}$$

Setting $\sigma^2_{v,SE} = 0$ and substituting the result to equation (18) yields the lower bound formula:

$$\bar{k}_{SE,L} = \exp\left(\frac{\gamma}{\beta} - \frac{1}{2}\left(\sigma^2_{\zeta,SE} - \sigma^2_{\zeta,EE}\right)\right). \tag{25}$$

By the same reasoning, we can derive the upper bound formula. The parameter $\bar{k}_{SE}$ (18) is highest when $\sigma^2_{u,SE}$ = $\sigma^2_{u,EE}$. PW argues that employees permanent income has at most as much variance as permanent income of self-employed, which implies the lower bound condition. Thus, the expression (22) can we written as:

$$\sigma^2_{\zeta,SE} - \sigma^2_{\zeta,EE} = \sigma^2_{v,SE} + \sigma^2_{u,EE} - \sigma^2_{u,SE}. \tag{26}$$

Substituting the above to (18) yields the upper bound formula:

$$\bar{k}_{SE,U} = \exp\left(\frac{\gamma}{\beta} + \frac{1}{2}\left(\sigma^2_{\zeta,SE} - \sigma^2_{\zeta,EE}\right)\right). \tag{27}$$

Along the lower and upper bounds, most analyses typically also report the point estimate of the scaling factor $\bar{k}_{SE}$. The calculation is based on assumptions required for both the lower and upper bound estimates. The value lies between $\bar{k}_{SE,L}$ and $\bar{k}_{SE,U}$ and serves as a useful summary measure for reporting purposes.

$$\bar{k}_{SE} = \exp\left(\frac{\gamma}{\beta}\right) \tag{28}$$

To calculate the average income gap $\overline{IG}$ (share of unreported income in reported and unreported income), we use the

$$\overline{IG} = 1 - \frac{1}{\bar{k}_{SE}} \tag{29}$$

where $\bar{k}_{SE}$ can be substituted by $\bar{k}_{SE,L}$ or $\bar{k}_{SE,U}$ to calculate lower or upper bound of the income gap.

Procedure for selection of final econometric model

In the following section, we explain our application of the Pissarides-Webber framework. We also discuss selection of instrumental variables, specification testing, post-estimation diagnostics, and calculation of confidence intervals and p-values for our estimates.

Estimation procedure

Our approach to estimating underreporting coefficients γ goes beyond the standard PW procedure, which assumes that only self-employed individuals can underreport their income. We use binary variables to differentiate between self-employed (SE_i) and

private sector employees (PE_i) and estimate separate coefficients for underreporting for each group (respectively γ_{SE} and γ_{PE}). The reference group in our case are public sector employees. Furthermore, we consider interactions between the control variables and the classification variable to identify differential effects on underreporting behaviour (see section 4.5).

Our estimation procedure utilizes the widely-used two-stage least squares (2SLS) method, which enables us to obtain an unbiased estimate of β thanks to the use of instrumental variables (IV). The instrumental variables are used in econometrics to estimate causal relationships between variables when there is concern about potential endogeneity or omitted variable bias so that (some) explanatory variables are correlated with the error term. The 2SLS method consists in estimating the regression of interest in two steps:

1. In the first stage, an endogenous variable (income variable in our case) is explained by instrumental variables and control variables. The 1st stage regression is estimated using ordinal least squares (OLS). In addition, in the PW approach the first-stage regression provides an estimate of income residual variance, which is essential for calculating the lower and upper bound of the scaling factor k and income gap IG . For each group (i.e. households classified as (1) public sector employee, (2) private sector employee and (3) self-employed), we estimate the first equation separately but keep the same variables across all three groups. This approach enables us to account for potential differences in income functions across groups while maintaining consistency in the set of variables used. Using the notation from the section 4.4.1, the “income equation” or 1st stage equation takes the form of:

$$\log\left(Y_i^{registered}\right) = \delta_0 + \delta_1 Z_i + \delta_2 X_i + \zeta_i$$

2. In the second stage, our main-interest explanatory variable (expenditure variable in our case) is explained by theoretical values $\hat{Y}_i$ (i.e. the values predicted from the 1st stage regression) of the income variable instead of its actual values. With this procedure, we remove from the income variable the component correlated with the 2nd stage error term which is the source of endogeneity and corresponding bias to the results. The 2nd stage regression is similarly to the 1st stage regression estimated using ordinal least squares (OLS), hence the name of the 2SLS method. Using the notation from the section 4.4.1, the “expenditure equation” or 2nd stage equation takes the form of:

$$\log(C_i) = \beta \log\left(\hat{Y}_i^{registered}\right) + \gamma_{SE} SE_i + \gamma_{PE} PE_i + \alpha X_i + \eta_i$$

Instrumental variables

In our study, guided mainly by the review of traces-of-true-income literature, we've identified several variables that could be used as instrumental variables for income. Good instruments for estimating the causal relationship between income and food expenditure should be “strong” and “exogenous”. "Strong" means that they're closely related to income, while "exogenous" means that they're not correlated with anything else expect income that could be affecting food expenditure. Specifically, we have considered:

- **Primary earner industry**[105]: Different industries offer different earning opportunities. While it may affect income, it may not affect consumption in a direct way. For example, an individual working in the manufacturing industry may have higher income than someone working in the retail industry, but this may not necessarily lead to differences in consumption patterns.

- **Primary earner education level**: Education is often positively correlated with income, as people with higher levels of education may have better job prospects and earn higher salaries. However, education may not have a direct effect on consumption.

- **Primary earner contract term** (permanent/temporary): Contract term is strongly related to the job security and earning potential of individuals, which in turn affects income. The length of contract term can be determined by internal factors (e.g. work experience of an individual) or external factors such as labour laws, union negotiations, or industry standards. What is more, contract term is unlikely to be correlated with food expenditure preferences.

- **Housing type**: Housing is often considered as a sign of a household's income and financial stability, as higher-income households may be more likely to own their housing or live in larger, more expensive properties. Housing type may be a strong instrument for income, as it is closely related to income and may not have a direct effect on food expenditure pattern.

- **Housing ownership**: Ownership of a home may be associated with higher income levels, as purchasing a house or an apartment typically requires a significant amount of financial resources. Therefore, ownership of a housing may be a strong instrument for income, as it is closely related to income and may not have a direct effect on food expenditure.

Post-estimation diagnostics

After estimating the regression model using the procedure described, we conducted several post-estimation diagnostics to assess the validity of our results. The IVs used in the estimation process were selected based on the Wu-Hausman test, the Sargan test, and the Wald test. Below, we describe the role and interpretation of each test.

- **The Wald test** is used to test for the presence of weak instruments, which can result in biased and inconsistent estimates. If the test does not reject the null hypothesis, it suggests that the instrumental variables are not strong enough and the estimates may be unreliable.
- **The Wu-Hausman** test compares the consistency of the 2SLS estimator with the OLS estimator by testing the null hypothesis that the OLS estimator is consistent. If the null hypothesis is rejected, it suggests that the 2SLS estimator is more efficient than the OLS estimator.
- **The Sargan** test checks the validity of the instrument relevance assumption by testing the null hypothesis that the instrument matrix is uncorrelated with the errors in the second-stage equation. If the null hypothesis is not rejected, it suggests that the instrumental variables are valid and not correlated with the unobserved errors.

[105] The information on industry of primary earners and contract term of primary earners is based on the HBS data, therefore, it is unavailable for individuals who reported in the survey that they were not working even if they reported positive income in their tax returns. We do not exclude such cases from the analysis, but those variables take the value "Not working" for them.

Selection of control variables

In order to select the control variables to include in our analysis, we followed the relevant literature and utilized the Akaike Information Criterion (AIC) through a backward selection process. This method allowed us to select the variables that have the greatest explanatory power for our outcome variable while avoiding overfitting the model. By utilizing this method, we were able to build a parsimonious model that is both accurate and interpretable.

In our study, we conducted the selection of the control variables for both the first and second stage equations of the 2SLS estimator. Using a common variable selection procedure for both stages, helps us to further reduce bias and improve the precision of the estimates. The technique, called "double selection", has recently become increasingly common in empirical studies, especially those that aim to estimate causal effects with instrumental variable methods (Belloni, Chernozhukov and Hansen, 2014[106]).

Statistical significance of underreporting parameters $\bar{k}$ and $\overline{IG}$

To assess the statistical significance of the underreporting parameters, the delta method is usually used in the literature to calculate their standard errors and confidence intervals. However, the delta method can be unreliable when sample sizes are small or when the underlying distribution is non-normal. In our analysis we decided to use bootstrap method because it does not rely on any assumptions as regards distribution of underreporting parameters and can provide more reliable standard errors and confidence intervals.

In our study, we use the non-parametric bootstrap method that does not assume any particular distribution for the errors or the underlying data, making it a flexible and robust method for estimating the uncertainty of underreporting parameters. By sampling the errors from the first and second stage regressions, we simulate the variability in the error terms and estimate how this affects the underreporting parameter of interest. Repeating this procedure many (R) times allows us to obtain a distribution of the underreporting parameter and estimate its uncertainty. Below, we will refer to N as the sample size.

The procedure of obtaining underreporting parameter distribution is following:

1. Estimate the model on the original sample
2. Sample N errors from 1st stage regression (ζ_j) with replacement
3. Sample N errors from 2nd stage regression (η_j) with replacement
4. Update the expenditure variable using the reduced form formula:[107]

$$\log(C_i) = \hat{\alpha}_0 + \hat{\beta}\left(\hat{Y}_i^{registered}\right) + \hat{\gamma}_{SE}SE_i + \hat{\gamma}_{PE}PE_i + \hat{\alpha}X_i + \eta_j + \hat{\beta}\zeta_j$$

5. Estimate the model and calculate the underreporting parameters $\bar{k}$ and $\overline{IG}$

[106] Belloni A., Chernozhukov V., Hansen Ch. (2014), Inference on Treatment Effects after Selection among High-Dimensional Controls, *The Review of Economic Studies*, Volume 81, Issue 2, Pages 608–650
[107] Note that both the sampled error terms refer to the same observation.

6. Repeat steps 2-5 to obtain R bootstrap estimates of the underreporting parameters

We calculate the standard deviation of the estimates across bootstrap samples to obtain the standard error. We construct the confidence intervals by taking the appropriate quantiles of the bootstrap distribution. Based on the computed quantiles we determine p-values as follows:

- p-value < 0.01 (***) if the 99% confidence interval does not contain 0;
- p-value < 0.05 (**) if the 95% confidence interval does not contain 0;
- p-value < 0.1 (*) if the 90% confidence interval does not contain 0.

A2.4 Country-level estimates of unreported income, lost revenues from PIT/social security contributions and related tax gaps

In this section we explain our approach to calculations of country-level estimates of unreported income and related categories.

We chose point estimates of mean income gaps $\overline{IG}$ of the final PW model - 26.0% for households classified as private sector employee households and 50.7% for households classified as self-employed households – the base of those shares is total (reported and unreported) net labour income.[108] We tested whether we could obtain different shares for each year in our sample, but we did not get statistically significant estimates of some parameters, which suggests that one-year sample is too small and the PW model for Bulgaria should be estimated for the HBS samples of 3-4 years pooled together.

For each year in our sample we calculated total unreported net income in Bulgaria using the following formulas (weighted averages on net labour income and sum of weights used are the same as in the Table A.3 (a)):

$$unreported\ income_{t,PRIVATE} = weighted\ average\ of\ net\ labour\ income_{t,PRIVATE} * \frac{0.260}{1-0.260} * sum\ of\ weights_{t,PRIVATE}$$

$$unreported\ income_{t,SELF} = weighted\ average\ of\ net\ labour\ income_{t,SELF} * \frac{0.507}{1-0.507} * sum\ of\ weights_{t,SELF}$$

As we do not want differences in the HBS sample selection in different years to affect the results (for example, in 2021 we observed a significant drop in net labour income among households classified as self-employed which is not reflected in macro data), we calculate unreported income in relation to GDP as the average for four years (2017, 2018, 2019 and 2021):

$$unreported\ labour\ income\ as\ \%\ of\ GDP = \frac{\sum_{t=2017}^{t=2021}(unreported\ income_{t,PRIVATE} + unreported\ income_{t,SELF})}{\sum_{t=2017}^{t=2021}(GDP_t)}$$

[108]The share of unreported income in reported income is calculated by the formula: $\frac{IG}{1-IG}$

Next, we calculate personal income tax and social security contributions lost due to income underreporting. We assume that net labour income that was not reported would be subject to taxation if reported (unreported net income would become reported gross income). The average shares of personal income tax and social security contributions in gross labour income were again calculated from our sample as the information is matched not only on net labour income but also on gross labour income, PIT and social security contributions paid. For each year in our sample we calculated country-level lost PIT and social security contributions (SSC):

$$lost\ PIT_{t,PRIVATE} = unreported\ income_{t,PRIVATE} * \ effective\ PIT\ rate_{t,PRIVATE}$$

$$lost\ PIT_{t,SELF} = unreported\ income_{t,SELF} * \ effective\ PIT\ rate_{t,SELF}$$

In the case of social security contributions of private sector employees, we had to additionally take into account contributions paid by employers, which, based on macro data from the NRA, accounted for approximately 141% of contributions paid by employees in this period.

$$lost\ SSC_{t,PRIVATE} = unreported\ income_{t,PRIVATE} * \ effective\ SSC\ rate_{t,PRIVATE} * 2.41$$

$$lost\ SSC_{t,SELF} = unreported\ income_{t,SELF} * \ effective\ SSC\ rate_{t,SELF}$$

Finally, we calculate PIT and social security contributions lost in relation to GDP as well as PIT gap and social security contributions gap as the average for four years:

$$lost\ PIT\ revenues\ as\ \%\ of\ GDP = \frac{\sum_{t=2017}^{t=2021}(lost\ PIT_{t,PRIVATE} + \ lost\ PIT_{t,SELF})}{\sum_{t=2017}^{t=2021}(GDP_t)}$$

$$PIT\ gap = \frac{\sum_{t=2017}^{t=2021}(lost\ PIT_{t,PRIVATE} + \ lost\ PIT_{t,SELF})}{\sum_{t=2017}^{t=2021}(PIT\ revenues_t + \ lost\ PIT_{t,PRIVATE} + \ lost\ PIT_{t,SELF})}$$

$$lost\ SSC\ revenues\ as\ \%\ of\ GDP = \frac{\sum_{t=2017}^{t=2021}(lost\ SSC_{t,PRIVATE} + \ lost\ SSC_{t,SELF})}{\sum_{t=2017}^{t=2021}(GDP_t)}$$

$$SSC\ gap = \frac{\sum_{t=2017}^{t=2021}(lost\ SSC_{t,PRIVATE} + \ lost\ SSC_{t,SELF})}{\sum_{t=2017}^{t=2021}(SSC\ revenues_t + \ lost\ SSC_{t,PRIVATE} + \ lost\ SSC_{t,SELF})}$$

A2.5 Differences in income underreporting between various socio-economic groups

This is a technical introduction to the section of the report on differences in income underreporting between various socio-economic groups.

The scale of non-compliance may vary across subgroups within a category of classification variable. In such cases, it is necessary to include interactions between these subgroups and the classification variable to examine how the relationships differ among them.[109] For example, suppose we want to examine how sex affects non-compliance. We might include an interaction term between sex and the classification variable to see if the role of unreported income differs depending on whether the

[109] Control variables in the standard PW model are included in order to better explain food expenditure and do not relate to the scale of underreporting.

household primary earner (or household head) is male or female. The 2nd stage regression formula and the point estimate of the scaling factor and income gap for self-employed households with male household head would take the following form:

$$\log(C_i) = \alpha_0 + \beta \log\left(\hat{Y}_i^{registered}\right) + \gamma_{SE,Male} SE_i * Male_i + \gamma_{SE,Female} SE_i * Female_i + \gamma_{PE,Male} PE_i * Male_i + \gamma_{PE,Female} PE_i * Female_i + \alpha_1 Male_i + \alpha_2 Female_i + \alpha X_i + \eta_i$$

$$\bar{k}_{SE,Male} = \exp\left(\frac{\gamma_{SE,Male}}{\beta}\right)$$

$$\overline{IG}_{SE,Male} = 1 - \frac{1}{\bar{k}_{SE,Male}}$$

In multiple cases, including interactions between variables resulted in groups with too few observations, making it difficult to draw reliable conclusions. Therefore, it was necessary to limit the number of interactions included in the analysis to ensure that there are sufficient observations in each group to draw meaningful inferences. While this limits our ability to fully capture the complexity of the relationships between variables, it is important to balance the need for including additional variables with the need for having enough observations to make accurate conclusions.

When exploring the relationship between multiple variables, including only one interaction is not enough to fully capture the complexity of the relationship. This is because control variables are often correlated with each other and changing only one variable while holding the others constant may not provide a complete understanding of the relationship. However, it can offer valuable insights into the importance and direction of the variables' impact.

To address the limitations, we chose to run multiple regressions with one interaction added in each model. By running multiple regressions with one interaction at a time, we can carefully examine the relationships between variables and their interactions and ensure that we have enough observations in each group to draw meaningful conclusions. This approach can also help us in identifying which interactions are most important and inform future work that can explore these relationships in more detail.

When predicting outcomes using interaction coefficients from more than one model, it is important to be cautious and understand that the sum of the coefficients of interaction terms may not accurately represent the overall effect. This is because interaction terms represent the effect of the interaction between two variables on the outcome, but this effect is not necessarily additive with the effects of the individual variables.[110] Instead, the overall effect may be more complex and nonlinear. Therefore, relying solely on the sum of the interaction coefficients can lead to inaccurate predictions and interpretations of the relationship between variables.

A3. VAT gap

A3.1 Variables considered in VAT gap models

The table below presents the variables considered in our VAT gap models.

[110] Often, the way two or more variables affect an outcome is not as simple as just adding them up. For instance, when studying how (1) having children in the household and (2) having female as a primary earner in the household impact non-compliance, it's not enough to just add the effect of each factor together. That's because the impact of having female as a primary earners may be different in households with and without children.

Table A.4 – Information about variables considered in the VAT gap model

group of variables for our analysis	closest group(s) of factors from the literature review in the report	name of the variable	description	decision to exclude from the analysis	additional comments	number of observations (for divisions and sections, since 2014	number of observations for divisions, since 2014	number of observations for sections, since 2014	number of observations (for divisions and sections, all years	latest year available	earliest year available	number of divisions and sections (with any data point)	number of divisions (with any data point)	number of sections (with any data point)	sources	hyperlinks
Indicator: based on comparison of theoretical VAT revenues estimate with actual VAT revenues (explained variable in model 1)	Not applicable	vat_gap_output (vat_prod_gap_o_xcas in the database)	(output) VAT gap estimate obtained from the formula: (potential output VAT estimate - declared output VAT) / potential output VAT estimate, %. Where the potential output VAT estimate include potential VAT on sales, intracommunity acquisitions and import of services outsied the EU.	We selected the best VAT gap estimate among these five variants which was the one focusing on output VAT and adjusted for international trade.	Depending on the variable variant we focus on VAT result (VAT output minus VAT input) or VAT output. For more on VAT productivity ratios see: https://www.imf.org/external/np/fad/tpaf/pages/vat.htm For more on the estimation of theoretical VAT revenues at the sectoral level see: https://www.imf.org/en/Publications/TNM/Issues/2017/04/07/The-Revenue-Administration-Gap-Analysis-Program-Model-and-Methodology-for-Value-Added-Tax-Gap-44715	616	456	160	616	2021	2014	77	57	20	National Revenue Agency of the Republic of Bulgaria; Eurostat, National Accounts; EY calculations	-
Indicator: based on comparison of theoretical VAT revenues estimate with actual VAT revenues (variant 2)	Not applicable	vat_productivity_gap_r	VAT gap estimate obtained from the formula: 1 - (Gross Value Added * weighted VAT rate) / (Reported VAT result), %. Where reported VAT result is taken from the NRA and Gross Value Added from National Accounts in Eurostat.			616	456	160	616	2021	2014	77	57	20	National Revenue Agency of the Republic of Bulgaria; Eurostat, National Accounts; EY calculations	-
Indicator: based on comparison of theoretical VAT revenues estimate with actual VAT revenues (variant 3)	Not applicable	vat_productivity_gap_o	VAT gap estimate obtained from the formula: 1 - (reported output VAT / (output * weighted VAT rate)), %. Where reported output VAT is taken from NRA and output is taken from National Accounts in Eurostat.			616	456	160	616	2021	2014	77	57	20	National Revenue Agency of the Republic of Bulgaria; Eurostat, National Accounts; EY calculations	-
Indicator: based on comparison of theoretical VAT revenues estimate with actual VAT revenues (variant 4)	Not applicable	vat_productivity_gap_pv	VAT gap estimate obtained from the formula: 1 - (reported VAT result / potential VAT result), where potential VAT result is an estimated theoretical VAT that should be paid by each sector (output VAT plus VAT on imports minus claimable input VAT - all based on data sources outside the NRA), %.			403	256	147	403	2021	2014	51	32	19	National Revenue Agency of the Republic of Bulgaria; Organisation for Economic Cooperation and Development, Input Output tables; Eurostat, National Accounts; EY calculations	-
Indicator: based on comparison of theoretical VAT revenues estimate with actual VAT revenues (variant 5)	Not applicable	vat_productivity_gap_tb	VAT gap estimate obtained from the formula: 1 - (output VAT / (VAT output base * weighted VAT rate)), %. Where output VAT and VAT output base are taken from the NRA.			800	640	160	800	2021	2014	100	80	20	National Revenue Agency of the Republic of Bulgaria; Eurostat, National Accounts; EY calculations	-
Indicator based on VAT audits (explained variable in model 2)	Not applicable	vat_gap_audit	Ratio of additional VAT obligation established in audit to total VAT obligation (additional VAT + VAT declared by audited liable persons), %.			655	509	146	693	2021	2014	96	77	19	National Revenue Agency of the Republic of Bulgaria; EY calculations	*
Indicator: based on comparison of reported and collected VAT	Financial conditions of taxpayers / shock to financial condition	collected_VAT_remaining	Remaining VAT to be collected obtained from formula (Reported result VAT - collected VAT) / Reported result VAT * 100, %.			816	648	168	816	2021	2014	102	81	21	National Revenue Agency of the Republic of Bulgaria	
Cause or indicator: role of cash payments	Third party reporting available / probability of detection	cash_registered_income_share	Share of cash registered revenues in total revenues from operating activities, %.			752	600	152	846	2021	2013	94	75	19	National Revenue Agency of the Republic of Bulgaria	-
Cause: role of cash registers	Third party reporting available / probability of detection	businesses_with_cash_register_share	Share of businesses with at least one cash register in all active businesses, %.			752	600	152	846	2021	2013	94	75	19	National Revenue Agency of the Republic of Bulgaria	-
Cause: firm size	Business form	empl_micro_firms_share	Share of persons employed in micro firms (<10 persons employed) in total number of persons employed, %.			574	475	99	982	2021	2008	96	80	16	Eurostat, Structural Business Statistics; EY calculations	https://ec.europa.eu/eurostat/databrowser/bookmark/ec60126-caf5-4355-b413-98216ef1f483?lang=en https://ec.europa.eu/eurostat/databrowser/bookmark/167350-31c081-4e23-9074-a8bb3b295cc2?lang=en
Cause: firm size	Business form	empl_small_firms_share	Share of persons employed in small firms (10 to 49 persons employed) in total number of persons employed, %.			516	425	91	884	2021	2008	96	80	16	Eurostat, Structural Business Statistics; EY calculations	https://ec.europa.eu/eurostat/databrowser/bookmark/ec60126-caf5-4355-b413-98216ef1f483?lang=en https://ec.europa.eu/eurostat/databrowser/bookmark/167350-31c081-4e23-9074-a8bb3b295cc2?lang=en
Cause: firm size	Business form	empl_medium_firms_share	Share of persons employed in medium firms (50 to 249 persons employed) in total number of persons employed, %.			557	466	91	935	2021	2008	96	80	16	Eurostat, Structural Business Statistics; EY calculations	https://ec.europa.eu/eurostat/databrowser/bookmark/ec60126-caf5-4355-b413-98216ef1f483?lang=en https://ec.europa.eu/eurostat/databrowser/bookmark/167350-31c081-4e23-9074-a8bb3b295cc2?lang=en
Cause: firm size	Business form	empl_large_firms_share	Share of persons employed in large firms (250 and more persons employed) in total number of persons employed, %.			524	434	90	883	2021	2008	96	80	16	Eurostat, Structural Business Statistics; EY calculations	https://ec.europa.eu/eurostat/databrowser/bookmark/ec60126-caf5-4355-b413-98216ef1f483?lang=en https://ec.europa.eu/eurostat/databrowser/bookmark/167350-31c081-4e23-9074-a8bb3b295cc2?lang=en
Cause: firm size	Business form	gva_micro_firms_share	Share of value added at factor cost generated by micro firms (<10 persons employed) in total value added, %.			513	429	84	917	2020	2008	80	68	12	Eurostat, Structural Business Statistics; EY calculations	https://ec.europa.eu/eurostat/databrowser/bookmark/1b22b02f-d198-4728-b0cc-f0428a0ed5de?lang=en
Cause: firm size	Business form	gva_small_firms_share	Share of value added at factor cost generated by small firms (10 to 49 persons employed) in total value added, %.			441	364	77	801	2020	2008	80	68	12	Eurostat, Structural Business Statistics; EY calculations	https://ec.europa.eu/eurostat/databrowser/bookmark/1b22b02f-d198-4728-b0cc-f0428a0ed5de?lang=en
Cause: firm size	Business form	gva_medium_firms_share	Share of value added at factor cost generated by medium firms (50 to 249 persons employed) in total value added, %.			450	373	77	841	2020	2008	80	68	12	Eurostat, Structural Business Statistics; EY calculations	https://ec.europa.eu/eurostat/databrowser/bookmark/1b22b02f-d198-4728-b0cc-f0428a0ed5de?lang=en
Cause: firm size	Business form	gva_large_firms_share	Share of value added at factor cost generated by large firms (250 and more persons employed) in total value added, %.			442	366	76	799	2020	2008	80	68	12	Eurostat, Structural Business Statistics; EY calculations	https://ec.europa.eu/eurostat/databrowser/bookmark/1b22b02f-d198-4728-b0cc-f0428a0ed5de?lang=en
Cause: firm size	Business form / financial conditions of taxpayers	micro_firms_share	Share of micro firms (<10 persons employed) in total number of firms, %.	Data available at sections level (alphabetical codes) only. Annual averages of quarterly series.		152	0	152	152	2021	2014	19	0	19	National Revenue Agency of the Republic of Bulgaria	-
Cause: firm size	Business form / financial conditions of taxpayers	small_firms_share	Share of small firms (10 to 49 persons employed) in total number of firms, %.	Data available at sections level (alphabetical codes) only. Annual averages of quarterly series.		152	0	152	152	2021	2014	19	0	19	National Revenue Agency of the Republic of Bulgaria	-
Cause: firm size	Business form / financial conditions of taxpayers	medium_firms_share	Share of medium firms (50 to 249 persons employed) in total number of firms, %.	Data available at sections level (alphabetical codes) only. Annual averages of quarterly series.		152	0	152	152	2021	2014	19	0	19	National Revenue Agency of the Republic of Bulgaria	-
Cause: firm size	Business form / financial conditions of taxpayers	large_firms_share	Share of large firms (250 and more persons employed) in total number of firms, %.	Data available at sections level (alphabetical codes) only. Annual averages of quarterly series.		152	0	152	152	2021	2014	19	0	19	National Revenue Agency of the Republic of Bulgaria	-
Cause: firm size	Business form	vat_base_micro_firms_share	Share of micro firms in total VAT base, %.	Data available at sections level (alphabetical codes) only. Annual averages of quarterly series.		152	0	152	152	2021	2014	19	0	19	National Revenue Agency of the Republic of Bulgaria	-
Cause: firm size	Business form	vat_base_small_firms_share	Share of small firms in total VAT base, %.	Data available at sections level (alphabetical codes) only. Annual averages of quarterly series.		152	0	152	152	2021	2014	19	0	19	National Revenue Agency of the Republic of Bulgaria	-
Cause: firm size	Business form	vat_base_medium_firms_share	Share of medium firms in total VAT base, %.	Data available at sections level (alphabetical codes) only. Annual averages of quarterly series.		152	0	152	152	2021	2014	19	0	19	National Revenue Agency of the Republic of Bulgaria	-
Cause: firm size	Business form	vat_base_large_firms_share	Share of large firms in total VAT base, %.	Data available at sections level (alphabetical codes) only. Annual averages of quarterly series.		152	0	152	152	2021	2014	19	0	19	National Revenue Agency of the Republic of Bulgaria	-

group of variables for our analysis	closest group(s) of factors from the literature review in the report	name of the variable	description	decision to exclude from the analysis	additional comments	number of observations for divisions and sections, since 2014	number of observations for divisions, since 2014	number of observations for sections, since 2014	number of observations for divisions and sections, all years	latest year available	earliest year available	number of divisions and sections (with any data point)	number of divisions (with any data point)	number of sections (with any data point)	sources	hyperlinks
Cause: self-employment / sole trader	Business form	self_empl_share	Share of self-employed in total employment (domestic concept), %.		Some divisions available as aggregates (provided by Eurostat).	592	440	152	1258	2021	2005	76	55	19	Eurostat, National Accounts; EY calculations	https://ec.europa.eu/eurostat/databrowser/bookmark/a877bdec-7c0c-4e77-862c-4fbb73bfe2b?lang=en
Cause: self-employment / sole trader	Business form / financial conditions of taxpayers	sole_trader_share	Share of firms registered as sole traders, %.		Data available at sections level (alphabetical codes) only. Annual averages of quarterly series.	152	0	152	152	2021	2014	19	0	19	National Revenue Agency of the Republic of Bulgaria	-
Cause: taxation	Tax rate / tax at risk	vat_rate	Weigthed average VAT rate on goods and services produced or sold by the sector, %. The rate is weighted by the share of goods or services that are subject to the standard VAT (20%), reduced (9%), zero-rate (0%) or exemption (0%). The rate does not account for the share of exports that is typically subject to 0% VAT rate.	rather to be excluded from the analysis due to low variation between sectors and time periods (to establish the link with other variables certain amount of variability is required)	The shares of different VAT rates based on: 1) shares of turnover generated by firms that produce goods or services that are subject to non-standard VAT rate; 2) shares of consumer spending on certain goods or services and 3) shares of international transport (in transport).	840	672	168	1575	2021	2007	105	84	21	International Monetary Fund; Eurostat; EY calculations	-
Cause: taxation	Tax rate / tax at risk	vat_rate_export_adj	Weigthed average VAT rate on goods and services produced or sold by the sector including the zero-rated exports of goods and certain services, %. The rate is weighted by the share of goods or services that are subject to the standard VAT rate (20%), reduced (9%) or zero-rate or exemption (0%).	rather to be excluded from the analysis due to low variation between sectors and time periods (to establish the link with other variables certain amount of variability is required) as well as the fact that some indicator variables will be already adjusted for the role of exports/intra-community supplies in the sector	The shares of different VAT rates based on: 1) shares of turnover generated by firms that produce goods or services that are subject to non-standard VAT rate; 2) shares of consumer spending on certain goods or services and 3) shares of international transport (in transport). The share of exports in firms' revenues obtained from OECD Input-Output tables.	840	672	168	1575	2021	2007	105	84	21	International Monetary Fund; Organisation for Economic Cooperation and Development; Eurostat; EY calculations	-
Cause: business bankruptcies and births	Business form / financial conditions of taxpayers / shock to financial condition	firms_pop_net_growth	Net business population growth, %.		Some divisions available as aggregates (provided by Eurostat).	532	413	119	912	2020	2009	76	59	17	Eurostat, Business demography	https://ec.europa.eu/eurostat/databrowser/bookmark/87adcd9b-4400-4750-81b5-f930de5f1ce1?lang=en
Cause: business bankruptcies and births	Business form / financial conditions of taxpayers / shock to financial condition	firms_death_rate	Enterprise death rate obtained by dividing the number of enterprise deaths by the number of active enterprises, %.		Some divisions available as aggregates (provided by Eurostat).	532	413	119	988	2020	2008	76	59	17	Eurostat, Business demography	https://ec.europa.eu/eurostat/databrowser/bookmark/87adcd9b-4400-4750-81b5-f930de5f1ce1?lang=en
Cause: productivity / complexity of sector's products and services	Business form	labour_prod	Labour productivity obtained by dividing gross value added (chain linked volumes, 2015) by total employment, in constant thousand BGN.		Some divisions available as aggregates (provided by Eurostat).	592	440	152	1258	2021	2005	76	55	19	Eurostat, National Accounts; EY calculations	
Cause: productivity / complexity of sector's products and services	Business form	empl_prof_occup_share	Share of managers, professionals and technicians in total employment, %.		Data available at sections level (alphabetical codes) only.	144	0	144	246	2021	2008	21	0	21	Eurostat, Labour Force Survey; EY calculations	https://ec.europa.eu/eurostat/databrowser/bookmark/814ee652-7c7c-4782-9632-97bfc084f8d0?lang=en
Cause: productivity / complexity of sector's products and services	Business form	empl_elem_occup_share	Share of elementary occupations in total employment, %.		Data available at sections level (alphabetical codes) only.	99	0	99	175	2021	2008	21	0	21	Eurostat, Labour Force Survey; EY calculations	https://ec.europa.eu/eurostat/databrowser/bookmark/814ee652-7c7c-4782-9632-97bfc084f8d0?lang=en
Cause: economic or financial situation	Financial conditions of taxpayers / shock to financial condition	profitability	Profitability computed as a ratio of net operating surplus to output, %.		In few cases where output from national accounts is not an appropriate proxy for revenues, we use turnover instead, namely sections D and G (due to trade margins), and L (imputed rents). For divisions available as aggregates, we makes estimates based on (aggregate) values of output and shares in tunover.	592	440	152	1258	2021	2005	76	55	19	Eurostat, National Accounts; EY calculations	https://ec.europa.eu/eurostat/databrowser/bookmark/c540ea09-5e5d-4fe6-81f8-b3e520139fa7?lang=en
Cause: economic or financial situation	Financial conditions of taxpayers / shock to financial condition	economic_situation	Firms' perception of their economic situation in the recent months, net indicator (percentage of firms that declare growth - percentage of firms declaring decline in activity over the recent period), percentage points.		Annual averages of monthly unadjusted series. Data based on firms' answers to a survey question: How has your business activity (sales, production, building, business situation) developed over the past 3 months? Values computed as a difference between percentage of positive answers (increase in activity) and negative answers (decrease of activity). Data available at divisions level (two-digit numerical codes) only.	456	456	0	955	2021	2005	57	57	0	European Commission; EY calculations	https://economy-finance.ec.europa.eu/economic-forecast-and-surveys/business-and-consumer-surveys/download-business-and-consumer-survey-data/subsector-data_en
Cause: economic or financial situation	Financial conditions of taxpayers / shock to financial condition	limitation_demand	Share of companies that indicate demand as a main factor currently limiting production, %.		Annual averages of quarterly unadjusted series. Data based on firms' answers to a survey question: What are the main factors currently limiting your production? Data available at divisions level (two-digit numerical codes) only.	456	456	0	955	2021	2005	57	57	0	European Commission; EY calculations	https://economy-finance.ec.europa.eu/economic-forecast-and-surveys/business-and-consumer-surveys/download-business-and-consumer-survey-data/subsector-data_en
Cause: role of foreign capital	Business form	gva_foreign	Share of value added at factor costs generated by foreign-controlled companies, %.			469	388	81	867	2020	2008	79	66	13	Eurostat, FATS	https://ec.europa.eu/eurostat/databrowser/bookmark/4c184b61-874c-42a1-a667-2fbc5906b3a1?lang=en
Cause (or control variable): role of investment	Business form	inv_to_gva	Ratio of gross fixed capital formation to gross value added, %.		Some divisions available as aggregates (provided by Eurostat). This is to account for the fact that companies with high role of investment spending have often larger VAT input than VAT output.	592	440	152	1258	2021	2005	76	55	19	Eurostat, National Accounts; EY calculations	
Cause: type of clients	Business form	firms_b2b_rev_share	Share of firms' revenues coming from sales to other domestic firms, %.		Data available at sections level (alphabetical codes) only. Annual averages of quarterly series.	152	0	152	152	2021	2014	19	0	19	National Revenue Agency of the Republic of Bulgaria	-
Cause: type of clients	Business form	firms_b2g_rev_share	Share of firms' revenues coming from sales to government, %.		Data available at sections level (alphabetical codes) only. Annual averages of quarterly series.	152	0	152	152	2021	2014	19	0	19	National Revenue Agency of the Republic of Bulgaria	-
Cause: type of clients	Business form	firms_b2c_rev_share	Share of firms' revenues coming from sales to consumers, %.		Data available at sections level (alphabetical codes) only. Annual averages of quarterly series.	152	0	152	152	2021	2014	19	0	19	National Revenue Agency of the Republic of Bulgaria	-
Cause: type of clients	Business form	firms_exports_rev_share	Share of firms' revenues coming from exports, %.		Data available at sections level (alphabetical codes) only. Annual averages of quarterly series.	152	0	152	152	2021	2014	19	0	19	National Revenue Agency of the Republic of Bulgaria	-
Cause: role of government (and other types of legal forms)	Business form	various variables	Role of government entities (and other types of legal forms) in output VAT (base), %.	excluded from the analysis since government entities were covered in the "government and other" category where non-government entities were responsible for majority of the values											National Revenue Agency of the Republic of Bulgaria	-

group of variables for our analysis	closest group(s) of factors from the literature review in the report	name of the variable	description	decision to exclude from the analysis	additional comments	number of observations and sections, since 2014	number of observations for divisions, since 2014	number of observations for sections, since 2014	number of observations for divisions and sections, all years	latest year available	earliest year available	number of divisions and sections (with any data point)	number of divisions (with any data point)	number of sections (with any data point)	sources	hyperlinks
Cause: economic or financial situation	Financial conditions of taxayers / shock to financial condition	gva_cp_yoy	Gross Value Added (current prices), percentage change on previous year, %.		Some divisions available as aggregates (provided by Eurostat).	592	440	152	1184	2021	2006	74	55	19	Eurostat, National Accounts; EY calculations	-
Cause: economic or financial situation	Financial conditions of taxayers / shock to financial condition	gva_cvl15_yoy	Gross Value Added, (chain linked volumes, 2015), percentage change on previous year, %.		Some divisions available as aggregates (provided by Eurostat).	592	440	152	1184	2021	2006	74	55	19	Eurostat; National Accounts; EY calculations	-
Cause: role of government		gov_effectiveness	The value of the indicator measuring the government effectiveness from the Worldwide Governance Indicators. It ranges from approximately -2.5 (low government effectiveness) to 2.5 (high government effectiveness). It reflects perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies.		Available at national level.	8	0	0	17	2021	2005	0	0	0	World Bank, Worldwide Governance Indicators	-
Cause: legal system		integrity	Integrity of the legal system, index with values from 0 (worst) to 10 (best).	excluded from the analysis due to insufficient number of observations	Available at national level. The first source of this component is the International Country Risk Guide Political Risk Component I for Law and Order: "Two measures comprising one risk component. Each sub-component equals half of the total. The 'law' sub-component assesses the strength and impartiality of the legal system, and the 'order' sub-component dissolves popular observance of the law". The second source is Judicial Accountability, Compliance with the High Court, Judicial Review, Transparent Laws with Predictable Enforcement, and Access to Justice for Men from the V-Dem dataset. (An adjustment for the area as a whole is made later to account uniformly for gender disparities.) Each of the v-Dem variables is individually rated using the formula (Vi – Vmin) / (Vmax – Vmin), where Vi is the country's V-Dem score according to V-Dem, and Vmax and Vmin were set at 4.0 and 0, respectively. The five	7	0	0	16	2020	2005	0	0	0	Fraser Institute, Economic Freedom of the World	-
Cause: taxation		tax_rates	Average rate of VAT, CIT and PIT, %.	excluded from the analysis due to low variation across time periods	Available at national level.	8	0	0	17	2021	2005	0	0	0	International Monetary Fund, Tax Foundation; Fraser Institute, Economic	-
Cause: economic or financial situation		unem	Unemployment rate, % of total labor force (economically active population).		Available at national level.	8	0	0	17	2021	2005	0	0	0	World Bank, modeled ILO estimate	-
Cause: economic or financial situation		inflation	Consumer price index, change from previous year, %.		Available at national level.	8	0	0	17	2021	2005	0	0	0	International Monetary Fund, World Economic	-
Cause: economic or financial situation		gdp_growth	Gross domestic product, constant prices, change from previous year, %.		Considered due to the client's request. Available at national level.	8	0	0	17	2021	2005	0	0	0	International Monetary Fund, World Economic	-
Cause: economic or financial situation		poverty_rate	Percentage of population at risk of poverty, where the poverty threshold is defined as 40% of median equivalised income.		Considered due to the client's request. Available at national level.	8	0	0	14	2021	2008	0	0	0	National Statistical Institute	-
Cause: economic or financial situation		poverty_work_rate	Percentage of people at risk of poverty.		Considered due to the client's request. Available at national level.	8	0	0	14	2021	2008	0	0	0	National Statistical Institute	-
Cause: economic or financial situation		consumer_confidence	Indicator measuring households' perception of economic situation. It is obtained by taking arithmetic average of the balances (in percentage points) of the answers to the questions on the past and expected financial situation of households, the expected general economic situation and the intentions to make major purchases over the next 12 months.		Considered due to the client's request. Available at national level. Annual averages of quarterly series.	8	0	0	17	2021	2005	0	0	0	National Statistical Institute	-
Cause: trade		exports	Volume of exports of goods and services, change from previous year, %.		Considered due to the client's request. Available at national level.	8	0	0	17	2021	2005	0	0	0	International Monetary Fund, World Economic	-
Cause: trade		imports	Volume of imports of goods and services, change from previous year, %.		Considered due to the client's request. Available at national level.	8	0	0	17	2021	2005	0	0	0	International Monetary Fund, World Economic	-
Cause: trade		trade_ratio	Exports to imports of goods ratio, current prices.		Considered due to the client's request. Available at national level.	8	0	0	17	2021	2005	0	0	0	National Statistical Institute	-
Cause: demography		young	Share of people aged 15-34 in the working age population (15-64), %.		Considered due to the client's request. Available at national level.	8	0	0	17	2021	2005	0	0	0	United Nations, World Population Prospects	-
Cause: health situation		life_exp	Life expectancy at birth, total (years).		Considered due to the client's request. Available at national level.	8	0	0	17	2021	2005	0	0	0	United Nations, World Population Prospects	-
Cause: health situation		hospital_beds	Hospital beds per 1,000 population.	excluded from the analysis due to insufficient number of observations	Considered due to the client's request. Available at national level.	4	0	0	8	2017	2010	0	0	0	World Health Organization	-
Cause: health situation		physicians	Number of physicians in health establishment per 10,000 population.		Considered due to the client's request. Available at national level.	8	0	0	17	2021	2005	0	0	0	National Statistical Institute	-
Cause: migration		migration	Net migration rate (number of immigrants minus number of emigrants in a given year), per 1,000 population.		Considered due to the client's request. Available at national level.	8	0	0	17	2021	2005	0	0	0	United Nations, World Population Prospects	-
Cause: migration		immigrants_chg	Number of immigrants, change from previous period, %.		Considered due to the client's request. Available at national level.	8	0	0	14	2021	2008	0	0	0	National Statistical Institute	-
Cause: migration		immigrants_per_1000	Number of immigrants per 1,000 population.		Considered due to the client's request. Available at national level.	8	0	0	15	2021	2007	0	0	0	National Statistical Institute	-
Cause: opinions about the tax system and administration		app_cheat_tax	An indicator measuring social acceptance for cheating on taxes, where 1 represents the stand that cheating is never justifiable and 10 - it is always justifiable.	excluded from the analysis due to insufficient number of observations	Available at national level.	4	0	0	13	2017	2005	0	0	0	World Values Survey	-
Cause: opinions about the tax system and administration		tax_system_unfairness	Percentage of persons who define the tax and social security contribution system as unfair.		Available at national level.	8	0	0	8	2021	2014	0	0	0	National Revenue Agency of the Republic of Bulgaria	-
Cause: opinions about the tax system and administration		tax_system_unfairness_bus	Percentage of businesses that define the tax and social security contribution system as unfair.		Available at national level.	8	0	0	8	2021	2014	0	0	0	National Revenue Agency of the Republic of Bulgaria	-
Cause: opinions about the tax system and administration		tax_evasion_impunity	Percentage of persons who consider a likelihood of being detected for evading taxes and/or social security contributions to be low (sum of answers: : rather low and very low). The greater the value, the more persons believe in impunity of tax evasion.		Available at national level.	8	0	0	8	2021	2014	0	0	0	National Revenue Agency of the Republic of Bulgaria	-
Cause: opinions about the tax system and administration		tax_evasion_impunity_bus	Percentage of businesses that consider a likelihood of being detected for evading taxes and/or social security contributions to be low (sum of answers: : rather low and very low). The greater the value, the more firms believe in impunity of tax evasion.		Available at national level.	8	0	0	8	2021	2014	0	0	0	National Revenue Agency of the Republic of Bulgaria	-
Cause: opinions about the tax system and administration		tax_evasion_penalty_unsev erity	Percentage of persons who consider penalties for tax and/or social insurance evasion to be low (sum of answers: rather low and very low). The greater the value, the more persons perceive the penalties to be low.		Available at national level.	8	0	0	8	2021	2014	0	0	0	National Revenue Agency of the Republic of Bulgaria	-
Cause: role of card payments		cards_per_capita	The sum of credit and debit cards per capita		Available at national level. Data for the end of the	8	0	0	17	2021	2005	0	0	0	European Central Bank; NSI	-
Cause: role of card payments		cards_payments	The ratio of card payments to households final consumption expenditure. All card payments included except for those made e-money function.		Available at national level.	8	0	0	17	2021	2005	0	0	0	European Central Bank; Eurostat; National Accounts	-
Cause: crime rate		convicted	Persons convicted per 100,000 population aged 14 and over.		Considered due to the client's request. Available at national level.	8	0	0	17	2021	2005	0	0	0	National Statistical Institute	-

Source: EY.

A3.2 Data preparation

The process of data collection and manipulation is always fundamental in econometric modelling. The preparation of the database for the VAT gap model was a laborious task that included data transformation, estimation, disaggregation and imputation. Overall, we have collected over 60 variables at the sectorial level and over 20 at the country level.

Preparation of some variables was a multilevel process that required somewhat complex calculations (sometimes based on a set of assumptions). The most extensive work was done while preparing the dependent variable, $output\ VAT\ gap$. Similar steps were taken in order to estimate other (potential) dependent (explained) variables. In the end, we chose to model the best two out of seven considered variants.[111]

An important example of how we processed data is our estimation of sectors' VAT rates. First, we matched information on VAT rates (the standard, reduced and zero rates as well as exemptions) in Bulgaria with goods and services produced by each sector. Second, for sectors where several VAT rates apply, we took weighted averages, where estimated shares in turnover were used as weights. To obtain reliable estimates we used data at the most granular level available, including data on turnover, households' consumption expenditure and share of international transport in each type of transport. For details on our approach and assumptions, see section 5.1 and section A3.1 of the technical appendix.

Preparation of other independent (explanatory) variables mainly consisted of elementary data transformation such as computation of shares, ratios and year-on-year (percent) changes. In several cases we had to make additional assumptions (e.g. on the variable *profitability*, see section A3.1 in the technical appendix for more details). Given the panel character of our dataset, when possible, we performed sector disaggregation for some variables in order to increase the number of observations in the model (for example in the national accounts sectors such as e.g. C10, C11 and C12 are aggregated into one and we disaggregated them).

Once the initial database was ready, we took three steps to obtain the final dataset:

- **Imputation.** In our preliminary database we included all the available divisions (a second level in the NACE structure identified by a two-digit numerical code). As some variables were available at a less granular level of sections (a first level in the NACE structure identified by an alphabetical code), we decided to increase the number of observations in the dataset by imputing data to each division from its parent section, if for the former the data was not available (e.g. if the data was not available for division A01, we took values of the section A, etc.).[112] When considering the variables at the country level, all sectors take identical values. [113]
- **Exclusion of selected sectors and outliers.** First, we removed sectors whose primary production is exempt from the VAT. These include financial and insurance activities, public administration and defence, human health and social work activities, activities of households as employers and activities of extraterritorial

[111] These include measures of VAT gap based on (i) VAT audits, (ii) comparison of reported and collected VAT and (iii) comparison of theoretical VAT revenues estimate with actual VAT revenues (five variants).
[112] This procedure was applied only for variables expressed as shares, ratios and percentage changes. For some variables, mostly when based on national accounts, aggregates of two or three divisions are available. In such cases, we either assign values of the aggregates to the divisions or keep the aggregates and treat them as single sectors.
[113] Full list of NACE sections and divisions is available at ec.europa.eu.

organisations. Next, we looked at sectors where many special rules apply (real estate, transportation)[114] or where VAT refunds dominate (agriculture). Since our estimates for these sectors are potentially threatened by some level of inaccuracy, we decided to remove them from the sample. In fact, these sectors were often among the outliers in the dataset. Further data inspection (with particular focus on outlying values) led us to also exclude mining and quarrying. In the model with the measure of input VAT gap based on VAT audits, we additionally removed observations with a very small number of audited firms (less than 10) and very high values of VAT gap (*vat_gap_audit* greater than 400%[115]).

- **Interpolation.** In the final step of data preparation, we lineary interpolated all the missing values (including previously removed ones).

For most of the data preparation (especially data imputation) as well as for the econometric part of the analysis, we used Stata software[116] version Stata/IC 16.0 for Windows (64-bit x84-64) which is well suited for advanced data manipulation and panel estimations.

A3.3 Econometric model and identification of key factors

This section includes various technical information related to selection of econometric models and identification of key factors divided into two considered models.

Model of output VAT gap based on potential VAT estimate

Our dataset is a panel data characterised by two dimensions: a relatively large number of units (sectors) and relatively low number of time periods (years). In such data a researcher always needs to account for (unobserved) sectors heterogeneity. Neglecting these effects could be a source of a bias in the estimation results. In the temporal dimension, one may also have to pay attention to time series dependence, which should not be a large issue in our case, given the short time span of the data. There is also a topic of potential heteroskedasticity (unequal variance) of the residuals in the model.[117]

In the literature it is common to apply several estimation methods that differently deal with the issues of biasedness and efficiency of the estimators. Such a strategy not only allows to choose the correct and the best estimation method but also to test the stability and reliability of a model. The considered estimation methods included Fixed Effect, Random Effect, Feasible Generalised Least Squares and Panel Corrected Standard Errors models.

Prior to estimating any econometric model, we defined a set of conditions that our final specification had to meet. First, we wanted the model to primarily consist of variables available at the sectoral level so that we could possibly well explain the differences across sectors. Therefore, each tested model had more sectoral variables than the country-level ones. Second, we always included at least one variable at the country

[114] These rules include VAT exemption of revenues coming from leasing of residential buildings to individuals and certain other real estate transactions, or zero-rated international transport and related services.
[115] The threshold of 400% removes the greatest jumps and outliers. We tested other levels as well, for more details see section A3.3.
[116] Stata/IC 16.0 for Windows (64-bit x84-64).
[117] To test for autocorrelation, we performed Wooldridge test which suggests that we do have first-order autocorrelation in the panel. However, we think that this test may be unreliable because of a very short time span of the data (7 years). In addition, we performed Likelihood Ratio test for heteroskedasticity (sector specific variance) which confirmed the necessity to account for heteroskedasticity.

level as it allowed us to control for factors common to all the sectors (and that change over time) such as business cycle or quality of public institutions. Third, we also tried to maximise the number of relevant variables that enter the model and capture different factors affecting the VAT gap. And finally, we excluded cases where two independent variables were highly correlated or represented similar cause of the VAT gap (e.g. we did not include simultaneously variables representing share of micro firms in employment and gross value added).

We tested several methods to estimate the final model's specification, namely Fixed Effect (FE), Random Effect (RE), Feasible Generalised Least Squares with heteroskedastic but uncorrelated error structure (FGLS_no) and with first-order autoregressive autocorrelation structure (FGLS_ar1), and Panel Corrected Standard Errors with first-order autoregressive autocorrelation structure (PCSE_ar1) and with panel-specific first-order autoregressive autocorrelation structure (PCSE_psar1). First, we need to underline the fact that in all the models the estimated parameters have identical signs (direction of impact on the dependent variable) and relatively similar magnitude (importance of each factor). This is the first evidence that our results are stable and robust to using different estimation methods. Turning back to the issue of a short temporal dimension of our dataset, we can conclude that models which account for autocorrelation appear too restrictive (they impose a structure of panel robust to autocorrelation based on just 7 observations)[118]. Next, Random Effect is superior to Fixed Effect model as it is more efficient[119]. In the final choice between the RE and the FGLS_no, we incline to the latter as it is often the case that RE becomes inconsistent as new data arrives[120].

In the discussion included in the main part of the report, our model has been supported with theoretical considerations (direction of impact of explanatory variables) and econometric testing (applying different estimation techniques, testing for heteroskedasticity and autocorrelation). Next, we ran a few tests in order to verify the robustness of the model. In general, we can conclude that the model is fairly robust to changes in both sample and specification. First, we extended the sectors to the ones that were initially excluded in the process of data preparation (agriculture, mining and quarrying, transportation and real estate activities). Inclusion of these sectors separately as well as simultaneously does not significantly affect the estimates of the model (i.e. all the variables remain statistically significant, the coefficients have the same signs and similar magnitude in comparison to the base model)[121]. Second, the model is robust to exclusion of each independent variable. Third, the model is relatively robust to using alternative explanatory variables (e.g. different measures of the role of micro firms) or including additional variables. In these tests, we found vast majority of new variables to be statistically insignificant and to have no impact on all the other variables in the model. However, there are several exceptions when tested variables appeared significant. In the case of alternative variables, we found that $empl\_micro\_firms\_share$ (share of micro firms in employment, alternative to $vat\_base\_micro\_firms\_share$) enters the model with an opposite, negative, coefficient. Given that the analysis concerns VAT gap, we believe that the share of micro firms in

[118] Adding restrictions weighs on the efficiency of estimation method. However, as the number of years increases, these approaches might become relevant in the future.
[119] RE is more efficient because it can capture both the between variability (inter-sectoral aspect of the data) and the within variability (intra-sectoral changes), while FE focuses only on the latter. RE can only be considered if it is consistent, which in our case is proved by Hausman test.
[120] This turned out true while we performed robustness tests of the model – inclusion of the previously removed sectors or addition of new independent variables to the model made the RE inconsistent.
[121] These conclusions apply strictly to inclusion of agriculture, transportation and real estate activities. In the case of inclusion of mining and quarrying one variable, $vat\_base\_micro\_firms\_share$, falls out of the 10% significance range. This is caused by outlying values that are typical for this sector.

the VAT base is a better measure to assess the impact of firm's size on the VAT gap[122]. In the case of additional variables included to the base model, we found several variables to be statistically significant, namely (1) at the sectoral level, $cash\_registered\_income\_share$ (enters the model with "-" sign), and (2) at the country level, $inflation$ (+) and $immigrants\_per\_1000$ (+). We did not include these variables in the final model due to instability ($cash\_registered\_income\_share$), limited predictive power in the future analyses ($inflation$), and effects that are not relevant to all the sectors ($immigrants\_per\_1000$)[123]. Last but not least, we tested with time dummy whether the pandemic year, 2020, had particular impact on the results, but the variable turned out insignificant (please note that with unemployment rate we already control for the business cycle in the model).

Model of output and input VAT gap based on VAT audits

Once again, we restricted considered specifications to meet initial criteria (such as greater focus on sectorial factors rather than the country ones, capturing different factors that affect the VAT gap, exclusion of highly correlated variables). We estimated each relevant specification using six different methods (described in the previous section) that address several problems typical for panel data. Basing on the experience from the first model, as well as on the tests results (that show the presence of heteroskedasticity, but not autocorrelation), we favour the FGLS with heteroskedastic error structure (FGLS_no).[124]

We also conducted the robustness test of our model. Overall, the test results are unsatisfactory: the model turns out to be often vulnerable to changes in the sample and the specification. First, the model is only partially robust to inclusion of additional sectors such as agriculture, mining and quarrying, transportation and real estate activities. While the estimates are not significantly affected when sectors are included separately, adding all of them simultaneously heavily impacts our results. Other tests on the sample also show unambiguous results – changing the criterium for removal of outliers (moving the acceptable maximum value of VAT gap from 400% to 300% and 500%) does not seriously affect the model,[125] whereas the estimates are very vulnerable to even small changes in the threshold of minimum number of audited firms. Second, the model is not robust to inclusion of new explanatory variables (adding new variables to the model often makes other variables insignificant) and is only partially robust to exclusion of each independent variable. To summarise, our specification becomes inappropriate when put to tests.

[122] Another (significant) alternative variable is $profitability$ (alternative to $labour\_prod$) which enters the model with the same (positive) coefficient. This is not surprising since both variables are relatively highly correlated.
[123] As for $inflation$, we could observe a structural change in this variable in 2022 (from an average of 1% in the sample period 2014-2020 to 15.3% in 2022) making the (past) estimates bring little value to the future analyses. As for $immigrants\_per\_1000$, this variable implies that all the sectors are affected to the same extent by the number of immigrants per 1000 inhabitants, which does not seem likely. One should also remember that for the country-level variables there are only a few unique observations in the sample (values for different sectors in the same year are repeated). Therefore, testing the impact of such variables on the VAT gap in our framework is quite challenging and one should rather focus on the variables with a strong background in economic theory and other research.
[124] In this model Random Effect (RE) is not a considered estimation method as the Hausman test suggests that the independent variables are correlated with the individual heterogeneities leading the RE parameters to be inconsistent.
[125] However, the tested thresholds lead to removal/addition of just a few observations.

A3.4 Additional details of VAT gap models

Table A.5 – Estimated parameters in the output VAT gap model

Dependent variable: vat_gap_output	
vat_base_micro_firms_share	0.5522***
firms_death_rate	0.3172**
labour_prod	0.1569**
firms_b2g_rev_share	-0.9373***
unem	0.8764***
constant	-49.7541***

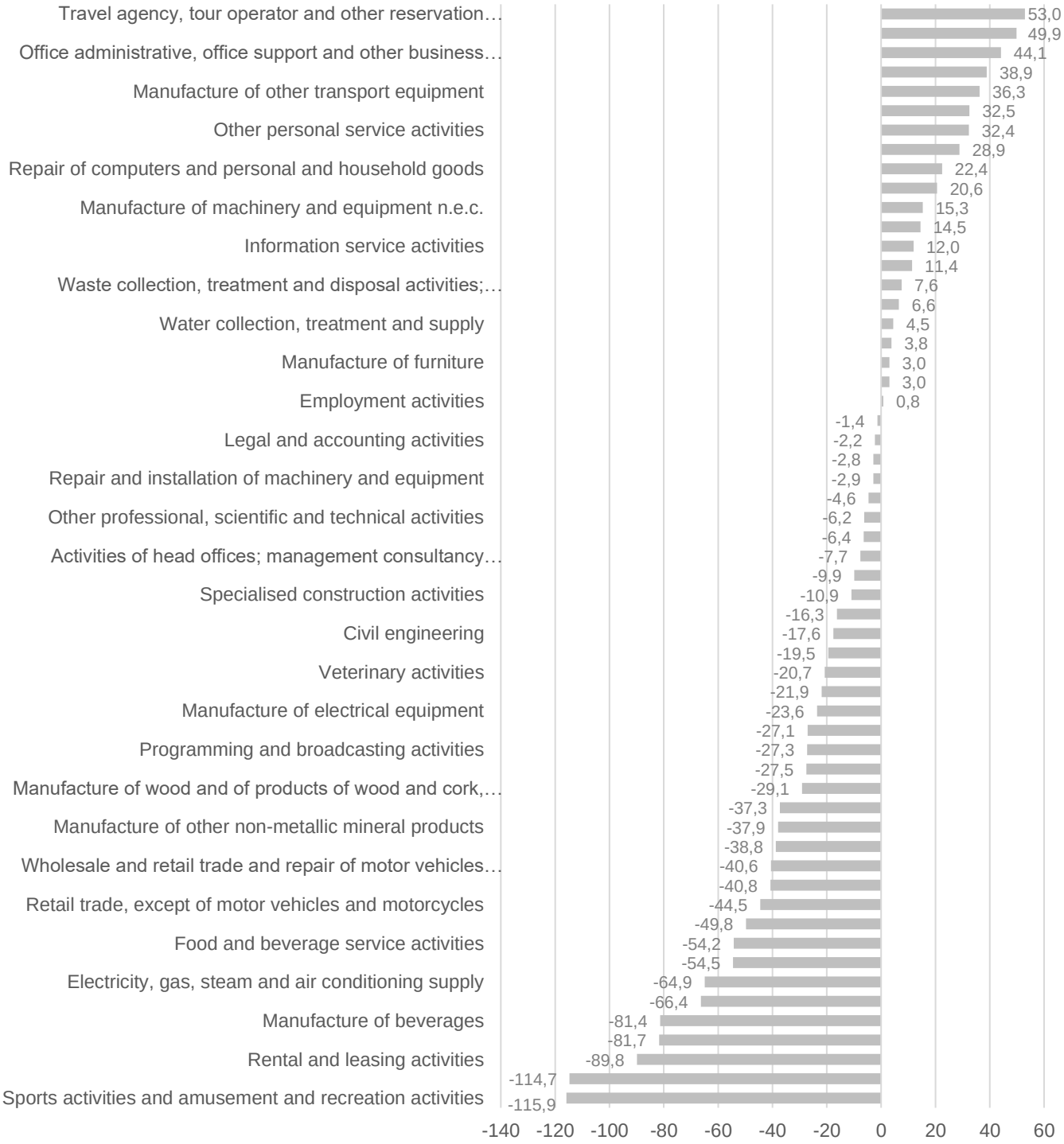
C10	0.0000	C32	78.6182***	J63	61.7514***
C11	-31.6023***	C33	46.8912***	M69	47.5121***
C12	-64.9476***	D35	-15.1906**	M70	42.0984***
C14	64.2227***	E36	54.2147***	M71	12.4829
C15	61.1518***	E37	22.6940***	M72	88.6603***
C16	20.6870***	E38	57.3264***	M73	30.2649***
C17	43.3440***	E39	-4.7316	M74	43.5077***
C19	-16.6043	F41	33.4974***	M75	29.0068***
C20	27.8775***	F42	32.1949***	N77	-40.0873***
C21	52.7910***	F43	38.8584***	N78	50.5670***
C22	48.3776***	G45	9.1829**	N79	102.7087***
C23	11.8068*	G46	10.9752***	N80	46.9184***
C24	8.9861*	G47	5.2592	N81	53.5342***
C25	39.8637***	I55	-31.9776***	N82	93.8746***
C26	99.6569***	I56	-4.4729	R93	-66.1228***
C27	26.1669***	J58	45.1042***	S95	72.1872***
C28	65.0991***	J59	56.3334***	S96	82.1213***
C29	82.2830***	J60	22.4793***		
C30	86.0398***	J61	22.2250***		
C31	52.7945***	J62	70.3975***		

Observations	399
Groups	57

Notes: Standard errors in parentheses. P-values marked with asterisks: *p<0.1, ** p<0.05, ***p<0.01. Groups = number of sectors included in the sample.

Source: EY.

Chart A.1 – Estimated fixed effects for each sector, model of output VAT gap



Source: EY.

A3.5 VAT gap estimates

In this section of the technical appendix we include methodological details related to translation of our econometric results into various VAT gap estimates.

Contributions of sectors to the overall VAT gap

The explained variable and obtained results from the output VAT gap model are not directly interpretable in monetary terms and should be viewed as relative measure (index) of the VAT gap presence in a sector – i.e. a higher value indicates that a sector is more prone to VAT non-compliance. Therefore, further, $\widehat{vat\_gap\_output}_{s,t}$[126] $= VAT\ gap\ index_{s,t}$. As a result, we need to transform obtained measures so that they would be easier to interpret.

First, we assumed that sectors: electricity, gas, steam and air conditioning supply, (NACE code: D65), financial services (NACE code: K), real estate services (NACE code: L), education (NACE code: P), human health and social work services (NACE code: Q) do not generate VAT gap (as most of those services are exempt from VAT or provided by the public sector).

In the second step, we calculated each sector contributions to the VAT gap in a given year (expressed in terms of % potential VAT in the whole economy) according to the formula:

$$VAT\ gap\ contribution_{s,t} =$$

$$= VAT\ gap\ index_{s,t} * \frac{GVA_{s,t}\ * VAT\ rate_{s,t} * (1 - export\ share_{s,t})}{\sum_s GVA_{s,t}\ * VAT\ rate_{s,t} * (1 - export\ share_{s,t})}$$

Where $VAT\ gap\ index_{s,t}$ is the index of VAT gap presence in the sector s in a year t, $GVA_{s,t}$ is the sector's *s* gross value added generated in year t, $VAT\ rate_{s,t}$ is VAT rate applied in the sector s in year t and $export\ share_{s,t}$ is share of exports (inside and outside the EU) in the total output of the sector s in year t. The numerator and denominator in the ratio in the formula approximated potential overall VAT (i.e. difference between output VAT and input VAT) in the sector *s* and whole economy, respectively. In this step and further sections, having no other reliable models apart for the output VAT gap, we implicitly assumed that our relative results obtained in terms of the gap in output VAT could be extrapolated and interpreted in terms of the gap in the overall VAT.

In the third step, we calculated each sector contributions to the overall VAT gap in year *t* as % of the total VAT gap in the economy, by dividing the result from the formula above for the sector *s* and year *t* by the sum of such values over all sectors in year *t*.

VAT gap on the country level

Our calibration is based on four steps. First, we calculated the average share of VAT gap in Bulgaria over the years 2016-2019 according to EC estimates expressed as % of collected VAT (row (a) in Table A.6). Second, we calculated the sums of $VAT\ gap\ contribution_{s,t}$ from the formula in the previous subsection over all sectors and years 2016-2019 (see row (b) in Table A.6). Third, for each year we divided the VAT gap model index for that year by the average VAT gap model index over the 2016-2019 period. Fourth, we multiplied the value obtained in step three by the average VAT

[126] This is the theoretical value from the output VAT gap model for the given sector and year (see section 5.2.1 for the formula) but without constant and fixed effects that we interpreted as accounting mostly for inaccuracies in the measurement of the VAT gap at the sectoral level (not fixed components of VAT non-compliance).

gap in Bulgaria from step one. This way we obtained the VAT gap in Bulgaria from our model in a given year expressed as % of the collected VAT revenues (row (c) in Table A.6). Such measure has its own evolution over time stemming from our results but the same 2016-2019 average as the EC VAT gap estimate.

Table A.6 – Summary of the key data used in the recalibration of the modelled VAT gap

	2014	2015	2016	2017	2018	2019	2020	2021
(a) European Commission - VAT compliance gap (% of the collected VAT)			14.66%	9.46%	12.72%	10.69%		
(b) VAT gap model index (year average)	19.96	18.22	18.67	17.40	17.43	17.18	18.40	19.42
(c) Model VAT compliance gap scaled to EC average (% of collected VAT)	11.7%	11.2%	11.8%	11.4%	11.9%	12.5%	13.9%	15.5%
(d) Collected VAT (BGN m)	7264	7740	8553	9320	10064	11086	11021	12979
(e) Model VAT compliance gap scaled to EC average (% of potential VAT)	11.8%	10.9%	11.2%	10.5%	10.5%	10.4%	11.0%	11.6%

Source: European Commission - VAT gap in the EU. Report 2022, EY.

Next, using the data from Ministry of Finance on collected VAT in BGN[127] (row (d) in Table A.6) we calculated the VAT gap expressed as % of the total potential VAT that could have been collected in Bulgaria under the assumption of perfect compliance (row (e) in Table A.6).

VAT gap in sectors

Having calculated our VAT gap model index for the whole economy (row (b) in Table A.6), for each year we divided the EC VAT gap estimate (% of potential VAT, not shown in Table A.6) by such index. This way we obtained scaling (correction) factor for our results. Next, we multiplied our initial $VAT\ gap\ index_{s,t}$ (% potential VAT in the sector but likely measured with inaccuracies) by the scaling factor and, thus, obtained our final sectoral VAT gap estimates (% of potential VAT in the sector) which in such approach are consistent with the EC VAT gap estimate at the country level.

[127] https://www.minfin.bg/en/1582 (online, accessed 18.05.2023).

EY | Assurance | Tax | Transactions | Advisory

About EY

EY is a global leader in assurance, tax, transaction and advisory services. The insights and quality services we deliver help build trust and confidence in the capital markets and in economies the world over. We develop outstanding leaders who team to deliver on our promises to all of our stakeholders. In so doing, we play a critical role in building a better working world for our people, for our clients and for our communities.

EY refers to the global organization and/or one or more of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. Information about how EY collects and uses personal data and a description of the rights individuals have under data protection legislation are available via ey.com/pl/pl/home/privacy.

EY analysis was conducted based on publicly available data, data provided by the Client as well as EY insights. EY has not performed any audit/assurance of the data used in the analysis.

EY analysis is based on statistical and/or econometric models. For this reason, the results of analyses should be taken to be an approximation to any true relationships, which is conditional on the method of analysis, the data used and any expert assumptions. If any of the assumptions, the data used or the methods of analyses are revised, the results of the project analyses may change relative to the results set out in the Report. EY assures the highest degree of professional care and conduct but regard must be taken of the uncertainties inherent in economic modeling.

Neither EY nor the employees of EY can be held responsible for the activities taken, or the lack of such activities, based on the information included in this Report. EY accepts no loss arising from any action taken or not taken by anyone using this publication.

EY responsibility for the contents of this Report is solely to the Client. EY takes no responsibility whatsoever to third parties using this Report. None of the Services or any part of the Report constitute any legal opinion or advice.

For more information about our organization, please visit ey.com.