



**COLLEGIUM OF ECONOMIC ANALYSIS
WORKING PAPER SERIES**

Evaluation of the underreporting of income
across households in Bulgaria: extending
the Pissarides-Weber approach

Piotr Dybka and Magdalena Karska
and Maciej Łopusiński and Andrzej Torój

January 2026

Evaluation of the underreporting of income across households in Bulgaria: extending the Pissarides-Weber approach.

Piotr DYBKA¹ * , Magdalena KARSKA², Maciej ŁOPUSIŃSKI¹ , and Andrzej TORÓJ¹ 

¹SGH Warsaw School of Economics, Poland

²EY Economic Analysis Team

*Corresponding author (pdybka@sgh.waw.pl)

2026

Abstract

We estimate the personal income tax (PIT) gap in Bulgaria using the Pissarides and Weber (1989) methodology (“traces of true income” approach), which compares the relationship between food expenditure and income of self-employed and other employees. Our analysis relies on a unique anonymized dataset prepared by the National Statistical Institute from Household Budget Survey and National Revenue Administration records, providing a more reliable measure of income than survey data alone. Extending the standard PW framework, we estimate under-reporting not only among the self-employed but also among private-sector employees. Our results show that unreported labour income averaged 6.37% of GDP during 2017–2021 (excluding 2020), with private-sector employees contributing 5.36% and the self-employed 1.01%. The PIT gap amounted to 13.8% of theoretical revenues, while the social security contribution gap reached 16.5%, corresponding to lost revenues of 0.54% and 1.71% of GDP, respectively. Moreover, our analysis also shows that households with children and younger earners are more prone to under-reporting. These findings underscore the importance of accounting for household characteristics when designing policies to mitigate tax non-compliance.

Keywords: Shadow economy, Tax gap, PIT

JEL: C51, E26, H21, H26, O17

Declaration of interest: The present article is based on the results obtained during the Project REFORM/SC2021/045 “Strengthening the Compliance Management by Assessing External Context and Taxpayers Behaviour in Bulgaria” funded by the European Commission (EC) through DG REFORM. The beneficiary of the Project was the National Revenue Agency (NRA) in Bulgaria.

1 Introduction

Shadow economy, tax non-compliance and income underreporting represent persistent challenges for tax administrations worldwide (see, e.g. Dybka et al., 2019, 2023; Goel & Nelson, 2016; Medina & Schneider, 2018; Schneider & Enste, 2000), undermining fiscal sustainability, distorting economic incentives (Buehn et al., 2018; Torgler et al., 2010), and eroding public trust in tax systems (D’Hernoncourt & Méon, 2012). There are various factors affecting the non-compliance such as economics of crime factors (costs and benefits) (see, e.g. Becker, 1968; Friedman et al., 2000; Gërxhani, 2004), the sense of justice (Verboon & van Dijke, 2007), social norms (Alm et al., 2019) or relations with the authorities (Braithwaite, 2003).

Bulgaria, like many post-communist economies, faces particular challenges in tax compliance. The legacy of institutional transformation and evolving taxpayer attitudes toward compliance make enforcement more challenging. The goal of this study was to evaluate the Personal Income Tax (PIT) gap, defined as the difference between theoretical tax revenues and actual collections (Dybka et al., 2024), in Bulgaria. Even though there is a substantial share of self-employed people in Bulgaria, we also account for the non-compliance of the private sector employees that constitute majority of labour market in Bulgaria.

Our analysis relies on a unique anonymized dataset combining Bulgaria’s Household Budget Survey (HBS) with National Revenue Administration (NRA) tax records for 2017-2021 (excluding 2020 due to pandemic-related survey suspension). This matched dataset provides reliable measures of both reported income and household expenditure, overcoming key limitations of survey-only data.

Our study relies on the idea of the "traces of true income" methodology developed by Pissarides and Weber (1989) based on the assumption that if households underreport income, their observed consumption, particularly on necessity goods like food, will appear disproportionately high relative to their reported income. By comparing the relationship between food expenditure and reported income across different employment groups, the method identifies systematic underreporting patterns. The original PW study applied this approach to British data, finding evidence of substantial income concealment among the self-employed.

Our study makes several contributions to the literature on tax compliance and income underreporting. First, we provide the first comprehensive estimate of the PIT gap in Bulgaria using matched administrative and survey data, addressing limitations of previous survey-only approaches. Using income variable based on the survey data (instead of official data from tax returns) 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, e.g. Cabral et al., 2019). Second, we extend the standard PW framework to estimate separate underreporting parameters for private-sector employees and self-employed households, using public-sector employees as the fully compliant reference group. Third, we develop a transparent aggregation methodology to translate micro-level estimates into economy-wide measures of unreported income and revenue losses. Fourth, we examine heterogeneity in underreporting across demographic, household, and temporal dimensions, providing insights for compliance policy design.

The remainder of this paper is organized as follows. Section 2 describes our data sources and preparation procedures, whereas section 3 presents the Pissarides and Weber (1989) methodology in detail, explaining the theoretical foundations, estimation procedure, and our extensions to the standard framework. Section 4 reports econometric results and estimated income gaps for both private-sector and self-employed households, whereas in Section 5 develops our aggregation approach and presents macro-level estimates of unreported income and associated PIT and social security contribution gaps. Section 6 examines heterogeneity in underreporting across socioeconomic characteristics, providing insights into which household types exhibit higher non-compliance. Section 7 concludes with policy implications and suggestions for future research.

2 Data

We estimate income underreporting using the traces-of-true-income (Pissarides-Weber, PW) method, which relates food expenditure to reported income. The analysis relies on a unique, anonymized micro-dataset prepared by the National Statistical Institute (NSI), combining the Household Budget Survey (HBS) and administrative records from the National Revenue Administration (NRA).

2.1 Datasets

The HBS, conducted annually with a temporary suspension in 2020 due to the COVID-19 pandemic, provides information on household consumption, income, and socio-demographic characteristics. The survey employs a two-stage probability sampling design, following international standards aiming at providing the highest possible level of representativeness for the population. Between 2017 and 2021, the dataset covered around 2,900 households per year. Information includes both individual-level variables (e.g., sex, age, education, marital status, employment status, industry, and sector of employment) and household-level indicators (e.g., settlement type and size, housing status, household size and composition, income, and expenditures across 12 consumption categories).

The survey data may suffer from measurement error, especially related to the value of income. To address these limitations, survey information was matched with NRA data on tax returns and employer declarations. The administrative data provide verified records of gross and net income, tax liabilities, and social security contributions for both employees and the self-employed. Income sources are categorized as self-employment, private-sector employment, or public-sector employment. Additional NRA data at the macro level, such as average incomes and contributions by sector, gender, and age, were used to validate the micro-data and support supplementary calculations.

2.2 Data Preparation and Sample Construction

Several steps were taken to harmonize and clean the merged dataset. Households with unidentified adult members or with more than 25% of income from unknown sources were excluded. Missing or inconsistent income values were reconstructed under standard assumptions (e.g., a 10% tax rate for self-employed under patent tax), with robustness checks confirming negligible impact on results. Inconsistencies in net income reporting were corrected by recalculating values based on gross income, tax, and contributions.

Individuals with missing income sources (1.3% of the overall number of observations) were reclassified using tax return details (using information in the appendices to the Article 50 statement) or excluded from the analysis (11 households).

Household-level incomes were aggregated from individual records, while socio-demographic variables were constructed both for the household reference person (defined as the primary earner) and at the household level (e.g., number of children, elderly members, working adults, marital structure). To control for heterogeneity across years, time fixed effects were included in the econometric specifications. Pooling the four survey waves maximized the number of observations, thereby improving the reliability of estimates, particularly when examining underreporting across socio-demographic groups.

Variables in the HBS dataset were recoded from numeric to descriptive ordinal categories, with several variables (e.g. industry, education, settlement size) aggregated to reduce dimensionality. Age was grouped into 5-, 10-, and 15-year intervals. The HBS contained no missing values. The final merged dataset required several transformations of the NRA tax-return data. First, we excluded households in which at least one adult member could not be identified (2.9% of households), assuming unidentified minors had zero income. Second, for individuals reporting only gross income and social security contributions, we imputed net income using a 10% tax rate (0.6% of individuals in the original sample), alternative tax-rate assumptions had negligible effects. Third, for cases where social security contributions were not deducted when computing net income, we recalculated net income accordingly (0.3% of individuals). Fourth, income-source information (private/public/self-employed) was missing for 1.3% of individuals. We classified these cases using (i) information from the annual tax return (for 314 individuals), (ii) HBS labour market information (9 individuals), and (iii) assigned "Not available" status where classification was impossible (22 individuals). Households with more than 25% of their income from unidentified sources were excluded (0.1%).

Consistent with NRA guidance, we used net income from the Annual Tax Return under Article 50¹ whenever available, and otherwise substituted information from monthly Form 1² filings. Because the Pissarides-Weber approach operates at the household level, individual-level data were aggregated to total household net and gross income, as well as income shares from public employment, private employment, and self-employment. HBS income and expenditure data already refer to households. Each household was assigned a reference person, defined either as the HBS household head or the primary earner; although both definitions were tested, the latter aligns better with standard practice. From individual-level HBS and NRA data, we constructed additional household-level socio-demographic controls (e.g. number and age structure of children, adult composition, gender composition, employment indicators, education and labour-market attachment, and several binary indicators for unemployment, disability, and temporary non-employment). A binary marital-status variable was also constructed.

To maximise statistical power for identifying heterogeneous underreporting patterns, we estimated the model on a pooled four-year sample. All monetary variables (HBS income/expenditure and NRA net/gross income) were deflated to 2021 prices using Eurostat HICP, and year fixed effects were included. One extreme outlier, a 2019 household with labour income exceeding 1 million BGN, was winsorised to the second-highest value

¹This is the article of the Bulgaria's Personal Income Tax Act that specifies conditions when resident individuals file an annual tax return to declare various sources of incomes

²This is a document filed by employers and self-employed individuals to report social security, health insurance, and personal income tax data to the National Revenue Agency

(210,000 BGN).

The PW method compares food-expenditure-to-income relationships of potentially non-compliant groups with a fully compliant reference group. We distinguish (i) private-sector employees, (ii) self-employed, and (iii) public-sector employees (reference), based on imputed income-source information. Households were classified according to income shares: a household is considered self-employed if at least 25% of labour income is derived from self-employment (as in the Pissarides and Weber (1989)); it is considered a private-sector household if the combined share of private-sector and self-employment income exceeds zero but self-employment income remains below 25%. Preliminary estimations with a finer subdivision of private-sector households yielded statistically similar underreporting parameters, so both sub-groups were pooled. Public-sector households were required to derive 100% of labour income from public employment. One household with non-classifiable income composition was removed. To ensure labour income is the primary economic resource, we additionally excluded households for which labour income (NRA) was lower than regular non-labour income reported in the HBS (24.1% of households with positive labour income).

The resulting estimation sample consists of 969 public-sector, 3758 private-sector, and 228 self-employed households (survey-weighted). Summary statistics (Table 1) show that average net income increases substantially after the restriction, particularly for self-employed households; income patterns across groups and years are broadly consistent with expectations. As in many household surveys, wealthy individuals, especially high-income self-employed, appear underrepresented, which implies that model estimates may not fully capture underreporting behaviour at the top of the income distribution.

Table 1: Key characteristics of households by years and assigned sector

Assigned sector	Year	Avg net income	% self	% private	% public	Avg food exp	N	Sum of weights
Initial sample, i.e. received dataset after excluding households with missing information or reported labour income = 0								
Public employee	2017	11687.9	0.0	0.0	100.0	3654.3	284	275787.0
Public employee	2018	12563.7	0.0	0.0	100.0	3696.5	294	304132.6
Public employee	2019	13881.8	0.0	0.0	100.0	3926.5	341	348150.9
Public employee	2021	16620.7	0.0	0.0	100.0	4153.9	338	321517.1
Private employee	2017	13580.6	0.5	86.4	13.1	4103.2	1222	1234554.1
Private employee	2018	14791.6	0.5	84.3	15.2	4153.2	1188	1238471.7
Private employee	2019	16432.6	0.4	87.2	12.4	4230.7	1161	1210277.8
Private 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	2.9	3809.7	120	133640.1
Self-employed	2021	7440.1	86.6	11.1	2.4	4338.4	122	132057.1
Estimation sample = initial sample after excluding 24.1% of households in which labour income was lower than other regular sources of income								
Public employee	2017	13693.9	0.0	0.0	100.0	3635.8	220	216204.9
Public employee	2018	14753.8	0.0	0.0	100.0	3726.0	232	243566.1
Public employee	2019	16617.5	0.0	0.0	100.0	3934.7	261	270460.6
Public employee	2021	20757.2	0.0	0.0	100.0	4131.6	255	247641.2
Private employee	2017	16119.0	0.6	83.9	15.5	4142.3	962	996266.8
Private employee	2018	17191.7	0.4	81.5	18.0	4256.7	961	1018163.8
Private employee	2019	19626.7	0.5	85.0	14.5	4339.7	910	967469.2
Private 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

Assigned sector	Year	Avg net income	% self	% private	% public	Avg food exp	N	Sum of weights
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

Finally, unlike the original PW study, we do not restrict the sample to two-adult households, as doing so substantially reduces sample size and impedes the analysis of heterogeneous underreporting. To control for heterogeneity across household sizes, binary variables were included in the econometric specifications and we report robustness checks based on estimation for only households with exactly two adults and households with at least two adults.

3 Methodology

3.1 Pissarides-Weber Methodology (PW Model)

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

In the original PW framework, all employees (working in either the public or the private sector) were treated as the reference group, while the scale of underreporting was estimated for the self-employed. 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 self-employed households.

3.1.1 Discrepancy Between Reported and True Income

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

$$Y_i^{\text{True}} = k_i Y_i^{\text{registered}}, \quad k_i \geq 1, \quad (1)$$

where k_i represents the extent of underreporting by household i .

In the PW model, it is assumed that there is no discrepancy for employees in paid employment ($k = 1$), but self-employed may underreport their true income ($k \geq 1$):

$$Y_i^{\text{True}} = \begin{cases} k_i Y_i^{\text{registered}}, & \text{if self-employed,} \\ Y_i^{\text{registered}}, & \text{if employee.} \end{cases} \quad (2)$$

Since neither k_i nor Y_i^{True} can be directly observed, indirect methods must be used for estimation. Pissarides and Weber propose using Engel curve regression coefficients to infer the extent of underreporting. The Engel curve relates household spending to household income:

$$\log(C_i) = \alpha + \beta \log(Y_i^P) + X_i \gamma + \epsilon_i, \quad (3)$$

where:

- α - constant term,
- C_i - food expenditure of household i ,
- Y_i^P - permanent income of household i ,
- β - elasticity of consumption with respect to income,
- $X_i\gamma$ - control variables and their parameters,
- ϵ_i - error term.

The assumption is that food spending depends on true income and socio-demographic factors, but not on employment status. Hence, if food expenditure is systematically higher for self-employed than for employees with the same income level, this suggests underreporting of income by the self-employed.

3.1.2 Permanent Income Considerations

Income consists of a permanent (expected) component and a transitory (unexpected) component. Permanent income, defined as the average income a household can expect to receive over a long period of time, is considered a better predictor of consumption behavior due to consumption-smoothing effects (Campbell & Mankiw, 1990). The relationship between true income and permanent income is given by:

$$Y_i^{\text{True}} = p_i Y_i^P, \quad (4)$$

where p_i is a random variable representing deviations of actual income from permanent income.

Taking logs of (1) and (4) and combining them yields:

$$\log(Y_i^P) = \log(Y_i^{\text{registered}}) - \log(p_i) + \log(k_i). \quad (5)$$

3.1.3 Distributional Assumptions

In the PW model, it is assumed that both k_i and p_i follow log-normal distributions:

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

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

For employees (EE, reference group), the no-underreporting assumption implies:

$$\log(k_{EE}) = 0, \quad \sigma_{v,EE}^2 = 0. \quad (8)$$

The underreporting is expected:

$$\log(\bar{k}_{SE}) = \mu_k + \frac{1}{2}\sigma_{v,SE}^2, \quad (9)$$

$$\log(\bar{k}_{EE}) = 0, \quad (10)$$

$$\log(\bar{p}_{SE}) = \mu_p + \frac{1}{2}\sigma_u^2, \quad (11)$$

$$\log(\bar{p}_{EE}) = \mu_p + \frac{1}{2}\sigma_u^2. \quad (12)$$

Pissarides and Weber (1989) argue that each group is characterized by the same mean of p_i , denoted by \bar{p} . Using this assumption, we can derive the relation of distribution parameters $\sigma_{u,SE}^2 \geq \sigma_{u,EE}^2$ between the groups:

$$\mu_{p,SE} - \mu_{p,EE} = -\frac{1}{2} (\sigma_{u,SE}^2 - \sigma_{u,EE}^2) \leq 0. \quad (13)$$

The assumption of unequal variances, $\sigma_{u,SE}^2 \geq \sigma_{u,EE}^2$, implies a difference between the means of $\log p_i$, which will later be used to obtain the mean under-reporting factor. Substituting equations (6) and (7) into (5), permanent income can be written as

$$\log(Y_i^P) = \log(Y_i^{\text{registered}}) + (\mu_k - \mu_p) + (v_i - u_i). \quad (14)$$

Substituting (14) into the Engel curve (3),

$$\log(C_i) = \alpha_0 + \beta \log(Y_i^{\text{registered}}) + \beta(\mu_k - \mu_p) + \beta(v_i - u_i) + \gamma X_i + \varepsilon_i. \quad (15)$$

The term $(\mu_k - \mu_p)$ is different for each group (self-employed and the employees) and to estimate it we will use binary variables. The error component $\beta(v_i - u_i) + \varepsilon_i$ is combined into η_i . The final regression equation is:

$$\log(C_i) = \alpha_0 + \beta \log(Y_i^{\text{registered}}) + \gamma SE_i + \alpha X_i + \eta_i. \quad (16)$$

where SE_i is a binary variable for self-employed households. Because $Y_i^{\text{registered}}$ is endogenous (its error term contains $\beta(v_i - u_i)$), equation (16) is estimated using two-stage least squares (2SLS), which yields an unbiased estimate of β and allows for the identification of income variances by group.

Equating $\beta(\mu_k - \mu_p)$ with γSE_i yields

$$\gamma = \beta \left(\mu_k - \frac{1}{2} (\sigma_{u,SE}^2 - \sigma_{u,EE}^2) \right). \quad (17)$$

Substituting the equation (9) in the above formula, we can derive the formula for scaling factor:

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

The first term of the inner sum can be obtained from the estimated regression. However, the variances required to compute the scaling factor associated with underreporting are unobserved, implying that the factor cannot be calculated exactly. To address this limitation, PW propose a method for deriving a range of plausible values within which the mean scaling factor \bar{k}_{SE} is likely to be located. This approach requires estimates of the total income variance for each group, obtained from the first stage of the 2SLS estimation procedure:

$$\log(Y_i^{\text{registered}}) = \delta_0 + \delta_1 Z_i + \delta_2 X_i + \zeta_i, \quad (19)$$

where X_i is the set of control variables (identical in both stages), Z_i is the set of excluded instruments, and δ_i are the corresponding parameters. The error term ζ_i is a composite of three random components: (i) unexplained variation in permanent income, ε_i ; (ii) deviations of true income from permanent income, u_i ; and (iii) deviations of registered

from true income, v_i . The first-stage regression is estimated under the assumption of unequal variances across groups. Accordingly, the variance of ζ_i can be expressed as the variance of the sum of these components:

$$\sigma_{\zeta,SE}^2 = \text{var}(\zeta_{i,SE}) = \text{var}(u_{i,SE} - v_{i,SE} + \varepsilon_{i,SE}), \quad (20)$$

$$\sigma_{\zeta,EE}^2 = \text{var}(\zeta_{i,EE}) = \text{var}(u_{i,EE} - v_{i,EE} + \varepsilon_{i,EE}). \quad (21)$$

Assuming equal variances of permanent income across groups ($\varepsilon_{SE} = \varepsilon_{EE}$), independence between ε_i and both u_i , v_i , and noting that $v_{i,EE} = 0$, the difference in variances is

$$\sigma_{\zeta,SE}^2 - \sigma_{\zeta,EE}^2 = \sigma_{u,SE}^2 + \sigma_{v,SE}^2 - 2\text{Cov}(u_{SE}, v_{SE}) - \sigma_{u,EE}^2. \quad (22)$$

Following PW, we assume

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

Under the above assumptions, $\sigma_{v,SE}^2$ and $\sigma_{u,SE}^2$ are negatively related: when the former increases, the latter decreases, and vice versa. The parameter \bar{k}_{SE} in equation (19) reaches its minimum when $\sigma_{v,SE}^2$ attains its lowest possible value, which is zero. This corresponds to the case where the underreporting rate is constant across all individuals. Under this condition, expression (23) simplifies to:

$$\sigma_{\zeta,SE}^2 - \sigma_{\zeta,EE}^2 = \sigma_{u,SE}^2 - \sigma_{u,EE}^2, \quad (24)$$

Setting $\sigma_{v,SE}^2 = 0$ yields the lower bound

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

The upper bound is obtained by imposing $\sigma_{u,SE}^2 = \sigma_{u,EE}^2$, which implies

$$\sigma_{\zeta,SE}^2 - \sigma_{\zeta,EE}^2 = \sigma_{v,SE}^2 + \sigma_{u,EE}^2 - \sigma_{u,SE}^2, \quad (26)$$

leading to

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

A commonly reported point estimate assumes both limit conditions hold simultaneously, giving

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

The average income gap, i.e. the share of unreported income in true income, is

$$\overline{IG} = 1 - \frac{1}{\bar{k}_{SE}}, \quad (29)$$

where \bar{k}_{SE} may be replaced by its lower or upper bound.

3.2 Estimation Procedure

Our estimation of the underreporting coefficients γ extends the standard PW framework, which typically assumes that only self-employed individuals underreport income. We differentiate between self-employed (SE_i) and private-sector employees (PE_i), estimating separate coefficients γ_{SE} and γ_{PE} , respectively. Public-sector employees serve as the reference group. To account for potential heterogeneity in reporting behavior, interaction terms between control variables and sectoral classification dummies are included.

The empirical strategy relies on the two-stage least squares (2SLS) method, which addresses endogeneity concerns through the use of valid instrumental variables. The general model can be summarized as follows.

First Stage (Income Equation):

$$\log(Y_i^{registered}) = \delta_0 + \delta_1 Z_i + \delta_2 X_i + \zeta_i, \quad (30)$$

where $Y_i^{registered}$ denotes reported income, Z_i is the vector of instruments, X_i is the vector of control variables, and ζ_i represents the error term. This stage is estimated using ordinary least squares (OLS) for each labor market group (public sector, private sector, and self-employed) separately, while keeping the same specification to maintain comparability. The estimated residual variance from the first stage is used to derive the scaling factor k and the income gap IG .

Second Stage (Expenditure Equation):

$$\log(C_i) = \beta \log(\hat{Y}_i^{registered}) + \gamma_{SE} SE_i + \gamma_{PE} PE_i + \alpha X_i + \eta_i, \quad (31)$$

where C_i denotes household food expenditure, and $\hat{Y}_i^{registered}$ represents the fitted income values from the first stage. This approach purges income of the component correlated with the second-stage error term η_i , mitigating endogeneity bias.

3.3 Instrumental Variables

Following the literature on the traces-of-true-income approach, we identify several candidate instruments that are both *strong* (correlated with income) and *exogenous* (uncorrelated with the expenditure equation error term). The selected instruments include:

- **Primary earner's industry:** Reflects earning potential across sectors but does not directly affect consumption preferences.
- **Education level:** Positively correlated with income but unlikely to directly influence food expenditure.
- **Contract type (permanent/temporary):** Affects job security and income potential but not consumption patterns.
- **Housing type:** Proxy for long-term financial stability, correlated with income but not with food consumption preferences.

- **Housing ownership:** Ownership status signals wealth and income capacity without directly influencing food expenditure.

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. These variables satisfy the relevance and exclusion restrictions required for valid instrumentation in the 2SLS estimation.

3.4 Selection of Control Variables

Control variables were selected using a backward elimination procedure based on the Akaike Information Criterion (AIC), following established practices in empirical economics. This approach ensures a parsimonious model that balances explanatory power and model simplicity. To mitigate omitted variable bias, we applied the “double selection” procedure proposed by Belloni et al. (2014), whereby variable selection is performed separately for both the first and second stage equations. This method enhances the robustness and precision of the 2SLS estimates.

3.5 Statistical Inference for Underreporting Parameters

To assess the statistical significance of the underreporting parameters \bar{k} and \overline{IG} , we employ a non-parametric bootstrap approach rather than the conventional delta method. The delta method assumes asymptotic normality, which may not hold in small samples; in contrast, bootstrapping provides distribution-free estimates of standard errors and confidence intervals. The bootstrap procedure proceeds (see Chernick & LaBudde, 2011, for a discussion) as follows:

1. Estimate the model on the original sample.
2. Draw N residuals from the first-stage regression (ζ_j) and N residuals from the second-stage regression (η_j) with replacement.
3. Update the expenditure variable according to:

$$\log(C_i) = \hat{\alpha}_0 + \hat{\beta} \log(\hat{Y}_i^{\text{registered}^{(r)}}) + \hat{\gamma}_{SE} SE_i + \hat{\gamma}_{PE} PE_i + \hat{\alpha} X_i + \eta_j^{(r)} + \hat{\beta} \zeta_j^{(r)}. \quad (32)$$

4. Re-estimate the model and compute \bar{k} and \overline{IG} .
5. Repeat steps 2–4 for (r) bootstrap replications.

The standard errors of the underreporting parameters are calculated as the standard deviation of the bootstrap estimates. Confidence intervals are constructed using the empirical quantiles of the bootstrap distribution.

4 Results

4.1 Econometric model

The list of variables included in the model along with their short description is presented in Table 5. 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.

Table 4.1 reports the estimates of the final specification (Column 1), which is based on the full sample of households. For comparison, we additionally present results for the sample restricted to households with exactly two adults, as in the original PW framework (Column 2) and for the sample restricted to households with at least two adults (Column 3). In line with standard practice in the literature, we summarize the second-stage results of the 2SLS procedure (the expenditure equation) and report the average income gaps, \overline{IG} , derived from: (i) the estimated coefficient on reported labour income (0.159 in the preferred specification); (ii) the coefficients on sectoral binary variables (0.048 for private-sector employee households and 0.113 for self-employed households) and for computing the lower and upper bounds of \overline{IG} , (iii) the group-specific variances of the residuals from the first-stage income regressions (0.279, 0.396, and 0.878 for public-sector employees, private-sector employees, and the self-employed, respectively).

Table 2: Results of the PW Model for Alternative Samples

	Dependent variable: log(HBS_expenses_food)		
	(1) All household	(2) Adults = 2	(3) Adults \geq 2
NRA_sectors_3: Private sector employee	0.048*** (0.014)	0.196*** (0.026)	0.048*** (0.016)
NRA_sectors_3: Self-employed	0.113*** (0.027)	0.291*** (0.036)	0.149*** (0.030)
log(hsh_NRA_net_income)	0.159*** (0.007)	0.139*** (0.010)	0.160*** (0.009)
year2018	0.032*** (0.012)	0.016 (0.022)	0.045*** (0.015)
year2019	0.037*** (0.014)	-0.026 (0.021)	0.034** (0.016)
year2021	0.136*** (0.015)	0.163*** (0.024)	0.132*** (0.017)
hsh_primary_earner_sex: Male	0.079*** (0.013)	0.062*** (0.014)	0.091*** (0.015)
hsh_primary_earner_age 35-49	0.016 (0.015)	0.003 (0.024)	0.016 (0.017)
hsh_primary_earner_age 50-64	0.048*** (0.018)	0.062** (0.024)	0.018 (0.019)
hsh_primary_earner_age 65+ children 0-6	-0.207*** (0.022)	-0.251*** (0.029)	-0.198*** (0.024)
children 7-12	0.016 (0.015)	0.020 (0.020)	0.022 (0.017)
children 13-18	0.047*** (0.015)	0.064*** (0.022)	0.048*** (0.016)
children 13-18	0.012 (0.017)	-0.114*** (0.020)	0.015 (0.018)
settlement >50k	0.020 (0.012)	0.049** (0.020)	0.019 (0.014)
settlement \leq 50k	0.083*** (0.014)	0.080*** (0.022)	0.085*** (0.016)
villages	-0.015 (0.012)	-0.029 (0.018)	-0.009 (0.013)
unemployment	-0.020 (0.023)	-0.057 (0.044)	-0.027 (0.026)
disability	-0.023 (0.020)	-0.097** (0.039)	-0.024 (0.023)
working_number_HBS	0.098*** (0.010)	-0.083*** (0.022)	0.096*** (0.010)
working_number_NRA	-0.071*** (0.011)	-0.011 (0.024)	-0.092*** (0.012)
housing_ownership (no loan)	0.071*** (0.014)	0.068** (0.027)	0.082*** (0.017)

Continued on next page

	Dependent variable: log(HBS_expenses_food)		
	(1) All house- holds	(2) Adults = 2	(3) Adults \geq 2
housing rented (vacant)	-0.084** (0.037)	-0.082 (0.077)	-0.091** (0.041)
housing rented (municipal)	0.085 (0.136)	0.397 (0.277)	0.126 (0.165)
housing without rent	0.055 (0.045)	0.134* (0.057)	0.085* (0.051)
share of members 60+	0.103 (0.071)	-0.021 (0.083)	0.095 (0.074)
primary earner studying	0.036 (0.059)	0.122** (0.053)	0.030 (0.063)
housing type: one-family	-0.046 (0.028)	-0.083** (0.037)	-0.058** (0.032)
housing type: multi-family	-0.053 (0.040)	-0.103*** (0.035)	-0.046 (0.044)
housing type: other	-0.302*** (0.126)	-0.238 (0.245)	-0.295** (0.133)
household size = 2	0.254*** (0.042)		0.131*** (0.046)
household size = 3	0.394*** (0.048)	0.414*** (0.181)	0.252*** (0.054)
household size = 4	0.511*** (0.058)	0.527*** (0.220)	0.498*** (0.065)
household size = 5	0.749*** (0.065)	0.369 (0.462)	0.582*** (0.085)
household size = 6	0.849*** (0.086)	0.967* (0.594)	0.934*** (0.127)
household size = 7	0.962*** (0.135)		1.207*** (0.163)
household size = 8	0.858*** (0.262)		1.135*** (0.254)
household size = 9	1.048*** (0.320)		1.189*** (0.321)
household size = 10	1.146*** (0.357)		1.493*** (0.392)
household size = 11	1.634*** (0.559)		1.928*** (0.873)
Constant	1.655*** (0.159)	6.295*** (0.187)	7.695*** (0.173)
Underreporting estimates			
IG private sector (point)	0.260*** (0.075)	0.306*** (0.077)	0.282*** (0.094)
IG private sector (upper)	0.302*** (0.070)	0.366*** (0.071)	0.323*** (0.088)
IG private sector (lower)	0.215** (0.079)	0.248** (0.083)	0.237** (0.099)
IG self-employed (point)	0.507*** (0.098)	0.618*** (0.088)	0.627*** (0.097)
IG self-employed (upper)	0.635*** (0.072)	0.787*** (0.066)	0.715*** (0.074)
IG self-employed (lower)	0.335** (0.132)	0.481*** (0.117)	0.512*** (0.127)
Variances of residuals from income regression			
Public sector employees	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-values)			
Weak instruments test	0.0000	0.0000	0.0000
Wu-Hausman endogeneity test	0.0000	0.0000	0.0000
Sargan overidentification test	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
Second-stage diagnostics			
Observations	4,955	2,566	4,089
R^2	0.244	0.189	0.232
Adj. R^2	0.239	0.166	0.225
Residual std. error	10.897	11.043	12.676

Continued on next page

	Dependent variable: log(HBS_expenses_food)		
	(1) All house- holds	(2) Adults = 2	(3) Adults ≥ 2
F-statistic	66.071***	17.063***	33.425***

Notes: Standard errors in parentheses. Parameters used for calculation of underreporting were bolded for convenience. Standard errors for IG parameters are obtained using 10,000 bootstrap iterations. Asterisks denote statistical significance: *p<0.1, **p<0.05, ***p<0.01. Survey weights applied.

4.2 Estimated scale of underreporting

The estimates indicate substantial underreporting of labour income among private-sector employee and self-employed households. Households classified as private-sector employee households underreport, on average, between 21.5% (lower PW bound) and 30.2% (upper PW bound) of their net labour income, with a point estimate of 26.0%.

$$\begin{aligned} \bar{k}_0 &= \exp\left(\frac{0.048}{0.159}\right) = 1.352, & \overline{IG}_0 &= \frac{1.352 - 1}{1.352} = 0.260, \\ \bar{k}_u &= \exp\left(\frac{0.048}{0.159} + \frac{1}{2}(0.396 - 0.279)\right) = 1.433, & \overline{IG}_u &= \frac{1.433 - 1}{1.433} = 0.302, \\ \bar{k}_l &= \exp\left(\frac{0.048}{0.159} - \frac{1}{2}(0.396 - 0.279)\right) = 1.275, & \overline{IG}_l &= \frac{1.275 - 1}{1.275} = 0.215. \end{aligned}$$

Households classified as self-employed underreport, on average, between 33.5% (lower PW bound) and 63.5% (upper PW bound) of their net labour income, with a point estimate of 50.7%. As before, these figures do *not* imply that the same proportion of income is unreported at the aggregate level of the whole economy.

$$\begin{aligned} \bar{k}_0 &= \exp\left(\frac{0.113}{0.159}\right) = 2.035, & \overline{IG}_0 &= \frac{2.035 - 1}{2.035} = 0.507, \\ \bar{k}_u &= \exp\left(\frac{0.113}{0.159} + \frac{1}{2}(0.878 - 0.279)\right) = 2.744, & \overline{IG}_u &= \frac{2.744 - 1}{2.744} = 0.635, \\ \bar{k}_l &= \exp\left(\frac{0.113}{0.159} - \frac{1}{2}(0.878 - 0.279)\right) = 1.505, & \overline{IG}_l &= \frac{1.505 - 1}{1.505} = 0.335. \end{aligned}$$

The substantially larger income gap among the self-employed is consistent with economic intuition and previous empirical findings, as self-employment provides more opportunities to conceal revenues (e.g., through non-recorded cash transactions). Underreporting among private-sector employees is likely to stem from "envelope wages", where part of the salary is paid informally in cash, as well as from other sources of unregistered income earned outside the main job (e.g., informal tutoring, household work, childcare, and minor repairs).

Relative to the existing literature, our estimated income gaps are substantially larger than those reported in the only earlier PW-type analysis for Bulgaria (Kukk et al., 2020), which used the EU-HBS 2010 survey. That study found statistically insignificant differences in underreporting for the self-employed, with bounds between 7.4% and 9.8%. This discrepancy is unsurprising given methodological differences: the earlier study relied on

self-reported survey income, which is subject to downward bias and used all employees rather than public-sector employees as the reference group, thereby masking underreporting among private-sector employees. A study employing a framework more closely aligned with ours, i.e. based on matched tax-register income data and a reference group restricted to public-sector employees, was conducted for Estonia (Paulus, 2015). Household classification in that case relied on the sector of the household head. That analysis produced income-gap bounds of 23.2-34.3% for private-sector employee households and 56.1-78.4% for self-employed households, which are even higher than our corresponding estimates. However, the Estonian sample was restricted to two-adult households and when compared with our Column 2 results, the findings are very similar.

Restricting our sample to households with two adults (accompanied by a reduced set of controls due to the limited number of self-employed observations) or to households with at least two adults does not change the qualitative conclusion that substantial non-compliance is present, though it affects the magnitude of the estimated underreporting. The point estimates of the mean income gap among private-sector employees are 26.0%, 30.6%, and 28.2% across the three specifications, while for self-employed households the corresponding estimates are 50.7%, 61.0%, and 62.7%. A similar pattern emerges when alternative expenditure variables are used.

The selection of the preferred model reflects two considerations. First, using the full sample maximizes the number of observations and ensures that the distribution of household types closely resembles the national population. Second, expenditures on food consumed at home are used in accordance with the original PW framework and its subsequent applications. It should be emphasized that model choice has a non-trivial impact on the resulting national-level estimates of total unreported income and the associated revenue losses from personal income tax and social-security contributions. Alternative specifications may therefore be considered in future applications of the PW methodology.

5 Aggregation to Macro-Level Estimates

The Pissarides-Weber (PW) model operates at the household level, requiring translation of micro-level estimates to economy-wide aggregates. As the literature provides limited guidance on this aggregation procedure, we develop our own approach. Our point estimates indicate that households classified as private sector employees underreport 26.0% of true income, while self-employed households underreport 50.7%. We consider three important factors that could result in potential bias of the macro-level estimates.

First, the PW model does not identify variation in underreporting across the income distribution. If lower-income households exhibit higher relative underreporting rates, e.g. due to stronger incentives to reduce tax and social security obligations, applying average income gaps would overstate aggregate unreported income. Conversely, tax avoidance strategies employed by higher-income households may fall outside our measurement framework, operating through mechanisms distinct from direct income underreporting.

Second, the logarithmic specification of our econometric model necessitates excluding households with zero net labor income. Workers concealing all income enter our sample only if other household members report positive income. This exclusion introduces downward bias in aggregate estimates of unreported income.

Third, we assess the representativeness of our Household Budget Survey (HBS) sample. The weighted average net labor income in our sample falls below the economy-wide

average based on National Revenue Agency (NRA) administrative data. Given likely underrepresentation of high-income households in survey data, we avoid assuming that top earners exhibit underreporting patterns identical to sampled households. Instead, we scale our estimates using survey weights, which represent the population of similar households. This approach effectively imputes underreporting rates from well-represented households to underrepresented high-income households, introducing conservative bias by applying lower observed income levels to the full population.

Our estimation sample excludes 24.1% of HBS households where net labor income falls below other regular income (see Table 1). We assume these excluded low-income households exhibit underreporting rates equal to our estimated parameters, which has minimal impact on aggregate results given their small contribution to total labor income.

5.1 Estimation of Aggregate Unreported Income

We apply mean income gap estimates, 26.0% for private sector employees and 50.7% for self-employed, to calculate annual unreported income.

For each year t in our sample, we compute total unreported net income as:

$$ULI_{t,PRIVATE} = \bar{Y}_{t,PRIVATE} \times \frac{0.260}{1 - 0.260} \times W_{t,PRIVATE} \quad (33)$$

$$ULI_{t,SELF} = \bar{Y}_{t,SELF} \times \frac{0.507}{1 - 0.507} \times W_{t,SELF} \quad (34)$$

where $ULI_{t,j}$ denotes the Unregistered Labour Income for group j , $\bar{Y}_{t,j}$ denotes the weighted average net labor income for group j and $W_{t,j}$ represents the sum of survey weights for group j .

To mitigate the influence of year-specific sampling variation in the HBS – such as an anomalous decline in self-employed income in 2021 not reflected in administrative data – we calculate unreported income relative to GDP as a four-year average:

$$\frac{ULI}{GDP} = \frac{\sum_{t \in \{2017, 2018, 2019, 2021\}} (ULI_{t,PRIVATE} + ULI_{t,SELF})}{\sum_{t \in \{2017, 2018, 2019, 2021\}} GDP_t} \quad (35)$$

5.2 Tax Revenue Losses

We estimate personal income tax (PIT) and social security contribution (SSC) losses under the assumption that unreported net income would be subject to standard tax treatment if declared. Effective tax rates are calculated from our matched sample, which includes both net and gross labor income along with actual tax payments.

Annual revenue losses are computed as:

$$\text{Lost PIT}_{t,j} = ULI_{t,j} \times \tau_{t,j}^{PIT} \quad (36)$$

$$\text{Lost SSC}_{t,PRIVATE} = ULI_{t,PRIVATE} \times \tau_{t,PRIVATE}^{SSC} \times 2.41 \quad (37)$$

$$\text{Lost SSC}_{t,SELF} = ULI_{t,SELF} \times \tau_{t,SELF}^{SSC} \quad (38)$$

where $\tau_{t,j}^{PIT}$ and $\tau_{t,j}^{SSC}$ denote effective tax rates for group j . The factor 2.41 in the private sector SSC calculation accounts for employer contributions, which averaged 141% of employee contributions during the sample period based on NRA administrative data.

We define tax gaps as the ratio of lost revenues to theoretical revenues (actual plus lost):

$$\text{PIT Gap} = \frac{\sum_t (\text{Lost PIT}_{t,\text{PRIVATE}} + \text{Lost PIT}_{t,\text{SELF}})}{\sum_t (\text{PIT Revenue}_t + \text{Lost PIT}_{t,\text{PRIVATE}} + \text{Lost PIT}_{t,\text{SELF}})} \quad (39)$$

$$\text{SSC Gap} = \frac{\sum_t (\text{Lost SSC}_{t,\text{PRIVATE}} + \text{Lost SSC}_{t,\text{SELF}})}{\sum_t (\text{SSC Revenue}_t + \text{Lost SSC}_{t,\text{PRIVATE}} + \text{Lost SSC}_{t,\text{SELF}})} \quad (40)$$

5.3 Main Results

Table 3 presents our macro-level estimates averaged across 2017, 2018, 2019, and 2021. Unreported labor income averages 6.37% of GDP, with private sector employees contributing 5.36 percentage points despite lower underreporting rates, reflecting their larger population share relative to the self-employed (1.01 percentage points).

Table 3: Estimated Unreported Labor Income and Tax Revenue Losses

Measure	Average 2017–2021
Unreported labor income (% of GDP)	6.37
Private sector employees	5.36
Self-employed	1.01
Lost PIT revenues (% of GDP)	0.54
PIT gap (%)	13.8
Lost SSC revenues (% of GDP)	1.71
SSC gap (%)	16.5

Note: Estimates represent averages for 2017, 2018, 2019, and 2021 (2020 excluded).

Source: Authors' calculations based on HBS data matched with NRA administrative records; GDP from NSI; PIT revenues from Ministry of Finance; SSC revenues from NRA.

The PIT gap – defined as lost revenues relative to theoretical revenues – averages 13.8%, while the SSC gap reaches 16.5%. The larger SSC gap reflects the tax structure whereby social contributions reduce the PIT base. In absolute terms, lost PIT and SSC revenues equal 0.54% and 1.71% of GDP, respectively.

6 Impact of socioeconomic factors on PIT compliance

To examine variation in underreporting across socio-economic groups, we extend the baseline PW model by incorporating interaction terms between the self-employed and private sector employee status variables and household characteristics (i.e. allowing for changes in estimated γ parameters across households). For a characteristic such as gender, the second-stage regression takes the form:

$$\begin{aligned} \log(C_i) = & \alpha_0 + \beta \log(\hat{Y}_i^{\text{registered}}) + \gamma_{\text{SE,Male}} \cdot \text{SE}_i \cdot \text{Male}_i + \gamma_{\text{SE,Female}} \cdot \text{SE}_i \cdot \text{Female}_i \\ & + \gamma_{\text{PE,Male}} \cdot \text{PE}_i \cdot \text{Male}_i + \gamma_{\text{PE,Female}} \cdot \text{PE}_i \cdot \text{Female}_i + \alpha_1 \text{Male}_i + \alpha_2 \text{Female}_i + \alpha \mathbf{X}_i + \eta_i \end{aligned} \quad (41)$$

from which the scaling factor and income gap for self-employed households with male heads are computed as:

$$\bar{k}_{SE, Male} = \exp\left(\frac{\gamma_{SE, Male}}{\beta}\right) \quad (42)$$

$$\overline{IG}_{SE, Male} = 1 - \frac{1}{\bar{k}_{SE, Male}} \quad (43)$$

Including multiple simultaneous interactions often yields subgroups with insufficient observations for reliable inference. We therefore estimate separate models, each containing a single interaction term, allowing us to examine specific relationships while maintaining adequate sample sizes. This approach provides insights into the direction and magnitude of effects, though it does not capture the full complexity of multi-way interactions. Therefore, the interaction coefficients from different models cannot be additively combined, as interaction effects may be nonlinear and context-dependent.

For each interaction model, we calculate point estimates of income gaps for analyzed subgroups, with public sector employee households serving as the reference category. Table 4 presents income gap estimates across socio-demographic characteristics. Interpretation requires first assessing statistical significance to determine whether estimates differ from zero, then examining magnitudes. We use the following notation: “=” indicates estimates within ± 5 percentage points (pp) of baseline for private employees (i.e. 26%) or ± 10 pp for self-employed baseline (i.e. 50.7%); “+” denotes significantly higher estimates beyond these thresholds; “-” indicates non-significant or substantially lower estimates. These differences may reflect compositional effects rather than causal relationships and should be interpreted with appropriate caution.

Table 4

Our results indicate that households with children exhibit significantly higher underreporting, particularly among private sector employees (39.0% vs. 20.1%). This pattern might reflect increased financial pressures associated with child-rearing expenses, creating stronger incentives to minimize tax burdens. Another potential explanation is that the subgroup of respondents with children may be subject to non-random selection, generating systematic differences relative to childless respondents. Alternatively, the variable may function as a proxy that partially captures intra-group age heterogeneity within the 18–39 cohort—given that individuals with children tend to be closer to the upper bound of this range, which may, in turn, increase their perceived competence in navigating tax-related decisions. Married-couple households demonstrate higher underreporting rates across both employment categories, contrary to literature suggesting married taxpayers exhibit greater compliance. This finding may be confounded with the presence of children, as these household characteristics are correlated.

Age patterns differ substantially by employment type. Among private sector employees, underreporting declines monotonically with age, peaking at 31.2% for ages 18–39 and becoming statistically insignificant for those over 60. This comports with literature documenting age-related increases in tax compliance. Conversely, self-employed households exhibit maximum underreporting in middle age (40–59 years; 61.2%), with lower rates among younger workers and no significant underreporting among those over 60. These contrasting patterns suggest distinct mechanisms: private sector underreporting may concentrate among lower-income younger workers facing liquidity constraints, while self-employed underreporting peaks during prime earning years when income concealment opportunities and magnitudes are greatest.

Table 4: Income Gaps by Socio-Demographic Characteristics

Variable	\overline{IG}	SE	Interpretation
<i>Baseline model</i>			
Private sector employee	0.260***	0.074	
Self-employed	0.507***	0.098	
<i>Children in household</i>			
Private emp.: no children	0.201*	0.093	-
Private emp.: children	0.390***	0.112	+
Self-emp.: no children	0.466**	0.132	=
Self-emp.: children	0.599***	0.147	=
<i>Married couple in household</i>			
Private emp.: not married	0.216*	0.116	=
Private emp.: married	0.314***	0.095	+
Self-emp.: not married	0.306	0.227	-
Self-emp.: married	0.642***	0.098	+
<i>Age of primary earner (3 groups)</i>			
Private emp.: 18-39	0.312*	0.153	+
Private emp.: 40-59	0.291***	0.080	=
Private emp.: 60+	0.069	0.199	-
Self-emp.: 18-39	0.494*	0.221	=
Self-emp.: 40-59	0.612***	0.098	+
Self-emp.: 60+	0.056	0.399	-

Note: Standard errors estimated via bootstrap (5,000 iterations). Significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Interpretation symbols defined in text.

Source: Authors' calculations.

7 Conclusions

This study applies the Pissarides-Weber methodology to estimate income underreporting in Bulgaria using matched Household Budget Survey and administrative tax data for 2017-2021. We develop a novel aggregation framework to translate micro-level estimates into economy-wide measures of unreported income and associated tax revenue losses.

Our baseline estimates indicate that private sector employee households underreport 26.0% of true income, while self-employed households underreport 50.7%. At the aggregate level, unreported labor income averages 6.37% of GDP over the estimation period, comprising 5.36 percentage points from private sector employees and 1.01 percentage points from the self-employed. Despite lower underreporting rates among private sector employees, their substantially larger share in the labour market implies their larger role in the PIT and SSC gap.

These estimates translate into significant fiscal losses. The PIT gap reaches 13.8%, while the social security contributions gap amounts to 16.5%. In absolute terms, unreported income generates annual revenue losses equivalent to 0.54% of GDP for personal income tax and 1.71% of GDP for social security contributions. The larger SSC gap stems from the fact that social contributions reduce the taxable base for income tax.

Extension of the baseline model through interaction analysis reveals substantial heterogeneity in underreporting behavior. Regarding household structure, households with children exhibit significantly higher underreporting (39.0% compared to 20.1% for private employees), likely reflecting increased financial pressures. Age patterns differ markedly by employment type. Private sector underreporting declines with age, reaching higher point at 31.2% for workers aged 18-39 and becoming statistically insignificant above age 60, consistent with life-cycle compliance patterns. Self-employed underreporting peaks in middle age (40-59 years at 61.2%).

Acknowledgments: The present article is based on the results obtained during the Project REFORM/SC2021/045 “Strengthening the Compliance Management by Assessing External Context and Taxpayers Behaviour in Bulgaria” funded by the European Commission (EC) through DG REFORM. The authors are grateful to the National Revenue Agency in Bulgaria, especially to Yordanka Chavorina-Kondareva and Nikolay Petkov, for their support. Moreover, the authors would also like to thank Marek Rozkrut for all his valuable comments and to all participants of the 9th International Conference The Shadow Economy, Tax Behaviour and Insights from Big Data in Warsaw, especially to Prof. Jason DeBacker who provided an excellent discussion.

References

- Alm, J., Schulze, W. D., von Bose, C., & Yan, J. (2019). Appeals to social norms and taxpayer compliance. *Southern Economic Journal*, 86(2), 638–666. <https://doi.org/10.1002/soej.12374>
- Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, 76(2), 169–217. Retrieved December 17, 2025, from <http://www.jstor.org/stable/1830482>
- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls†. *The Review of Economic Studies*, 81(2), 608–650. <https://doi.org/10.1093/restud/rdt044>

- Braithwaite, V. (2003). Dancing with tax authorities: Motivational postures and non-compliant actions. In {Valerie Braithwaite} (Ed.), *Taxing democracy: Understanding tax avoidance and evasion* (1st, pp. 15–39, Vol. 1). Ashgate Publishing Ltd.
- Buehn, A., Dell’Anno, R., & Schneider, F. (2018). Exploring the dark side of tax policy: An analysis of the interactions between fiscal illusion and the shadow economy. *Empirical Economics*, *54*(4), 1609–1630. <https://doi.org/10.1007/s00181-017-1277-7>
- 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. <https://doi.org/None>
- Campbell, J., & Mankiw, N. G. (1990). Permanent income, current income, and consumption. *Journal of Business & Economic Statistics*, *8*(3), 265–79. <https://EconPapers.repec.org/RePEc:bes:jnlbes:v:8:y:1990:i:3:p:265-79>
- Chernick, M. R., & LaBudde, R. A. (2011). *An Introduction to Bootstrap Methods with Applications to R*. Wiley.
- D’Hernoncourt, J., & Méon, P.-G. (2012). The not so dark side of trust: Does trust increase the size of the shadow economy? *Journal of Economic Behavior & Organization*, *81*(1), 97–121. <https://doi.org/https://doi.org/10.1016/j.jebo.2011.09.010>
- Dybka, P., Bazart, C., Hokamp, S., & Kirchler, E. (2024). Recent developments in the area of shadow economy and tax evasion research. *Journal of Economic Behavior & Organization*, *218*, 399–405. <https://doi.org/https://doi.org/10.1016/j.jebo.2023.12.018>
- Dybka, P., Kowalczyk, M., Olesiński, B., Torój, A., & Rozkrut, M. (2019). Currency demand and mimic models: Towards a structured hybrid method of measuring the shadow economy. *International Tax and Public Finance*, *26*(1), 4–40. <https://doi.org/10.1007/s10797-018-9504-5>
- Dybka, P., Olesiński, B., Rozkrut, M., & Torój, A. (2023). Measuring the model uncertainty of shadow economy estimates. *International Tax and Public Finance*, *30*(4), 1069–1106. <https://doi.org/10.1007/s10797-022-09737-x>
- Friedman, E., Johnson, S., Kaufmann, D., & Zoido-Lobaton, P. (2000). Dodging the grabbing hand: The determinants of unofficial activity in 69 countries. *Journal of Public Economics*, *76*(3), 459–493. [https://doi.org/10.1016/S0047-2727\(99\)00093-6](https://doi.org/10.1016/S0047-2727(99)00093-6)
- Gërzhani, K. (2004). The informal sector in developed and less developed countries: A literature survey. *Public Choice*, *120*, 267–300. <https://doi.org/10.1023/B:PUCH.0000044287.88147.5e>
- Goel, R. K., & Nelson, M. A. (2016). Shining a light on the shadows: Identifying robust determinants of the shadow economy. *Economic Modelling*, *58*, 351–364. <https://doi.org/https://doi.org/10.1016/j.econmod.2016.06.009>
- 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. <https://doi.org/10.1007/s10797-019-09562-9>
- Medina, L., & Schneider, F. (2018, January). *Shadow Economies Around the World: What Did We Learn Over the Last 20 Years?* (IMF Working Papers No. 2018/017). International Monetary Fund. <https://ideas.repec.org/p/imf/imfwpa/2018-017.html>
- Paulus, A. (2015, June). *Income underreporting based on income-expenditure gaps: Survey vs tax records* (ISER Working Paper Series No. 2015-15). Institute for Social and Economic Research. <https://doi.org/None>

- Pissarides, C. A., & Weber, G. (1989). An expenditure-based estimate of Britain's black economy. *Journal of Public Economics*, 39(1), 17–32. [https://doi.org/https://doi.org/10.1016/0047-2727\(89\)90052-2](https://doi.org/https://doi.org/10.1016/0047-2727(89)90052-2)
- Schneider, F., & Enste, D. H. (2000). Shadow economies: Size, causes, and consequences. *Journal of Economic Literature*, 38(1), 77–114. <https://doi.org/10.1257/jel.38.1.77>
- Torgler, B., Schneider, F., & Schaltegger, C. (2010). Local autonomy, tax morale, and the shadow economy. *Public Choice*, 144(1), 293–321. <https://doi.org/10.1007/s11127-009-9520-1>
- Verboon, P., & van Dijke, M. (2007). A self-interest analysis of justice and tax compliance: How distributive justice moderates the effect of outcome favorability. *Journal of Economic Psychology*, 28(6), 704–727. <https://doi.org/https://doi.org/10.1016/j.joep.2007.09.004>

Appendix A *Variables used in the model*

Table 5: Variables included in the final econometric model

Group of variables for our analysis	Name of the variable	Description and source
Expenditure variable in Pissarides-Weber model	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	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	log(hsh_NRA_net_income)	Natural logarithm of household net income reported to the NRA in constant 2021 prices. Source: HBS
Control variable	year	Ordinal variable for year: (1) 2017 (reference category), (2) 2018, (3) 2019, (4) 2021. Source: HBS/NRA
Control variable	hsh_primary_earner_sex	Ordinal variable for sex of household primary earner: (1) Female (reference category), (2) Male. Source: HBS
Control variable	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	children_0_6	Number of children aged 0-6. Source: HBS
Control variable	children_7_12	Number of children aged 7-12. Source: HBS
Control variable	children_13_18	Number of children aged 13-18. Source: HBS
Control variable	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	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	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	working_number_HBS	Number of household members who declared in the Household Budget Survey that they were working. Source: HBS

Continued on next page

Table 5 - *Continued from previous page*

Group of variables for our analysis	Name of the variable	Description and source
Control variable	working_number_- NRA	Number of household members who reported positive net income in their tax return. Source: NRA
Control variable	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	household_members_- 60plus_share	Share of household members who are at least 60 years old. Source: HBS
Control variable	hsh_primary_earner_- studying	Binary variable taking the value 1 in households in which the primary earner is currently studying. Source: HBS
Control variable	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	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	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	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