Searching for gaps: Bottom-up approach for Slovakia

M. Chudý[a,*], R. Gábik[b], J. Bukovina[a], L. Šrámková[a]

[a]*Institute for Financial Policy, Ministry of Finance, Štefanovičova 5, 811 05, Bratislava*
[b]*Directorate of Financial Administration, Mierová 202, 821 05, Bratislava*

Abstract

More than half of the corporate income tax (CIT) revenues is associated with the largest multinational corporations, although they represent only 1% of the corporate universe in Slovakia. The rest of CIT revenues consists of small and medium enterprises (SMEs). Facing only a tiny risk of tax audit, some firms routinely "adjust" their actual tax base. We conjecture that such noncompliance, while individually negligible, casts a severe threat upon the scale and the dynamics of the Slovak CIT gap. We take the first step towards disentangling and monitoring this threat by building parametric and semi-parametric selection-bias-corrected regression models from firm-level data and capturing the main descriptive attributes of the gap. Our regional, sectoral, and country-wide estimates are based on individual characteristics of firms. Apart from the gap estimates, we provide a robust assessment of the quality of tax audit data, urging the Slovak tax authority to improve audits' utility.

Keywords: corporate income tax, risk based audit, semi-parametric sample selection model, censoring, prediction, transformation bias correction

*Corresponding author
Email addresses: marek.chudy@mfsr.sk (M. Chudý), rastislav.gabik@financnasprava.sk (R. Gábik), jaroslav.bukovina@mfsr.sk (J. Bukovina), lucia.sramkova@mfsr.sk (L. Šrámková)

1. Introduction

The share of the CIT[1] on the total tax revenues in Slovakia is about 10%. However, the actual tax revenues (collected each year) are typically several percents below the economy's potential. This difference between actual and theoretical revenues is known as the tax gap. It is commonly used as a measure of the country-wide level of noncompliance of taxpayers (see OECD, 2017). More generally, the tax gap has two primary sources, which we call the "policy gap" and the "non-compliance gap". The former one captures all revenues lost due to tax reliefs and allowances. Its quantification is usually straightforward and belongs to the standard analytical output of the Ministry of Finance. The noncompliance gap stems from taxpayers' deliberate or non-deliberate failure to comply with the current tax legislation. Therefore, one of the major concerns of Tax Administrations is the optimal allocation of risk-based audits to minimize the tax noncompliance gap. A complementary interpretation thus sees the tax gap as a measure of tax enforcement. Estimation and timely monitoring of the gap provides crucial input for decisions concerning the administration of taxes. At the same time, it opens a challenging problem concerning the heterogeneity of taxpayers and the lack of relevant information on their tax discipline. This paper presents the first micro-data based CIT gap estimator in Slovakia, which is generally known as the "bottom-up approach".

Conventional noncompliance gap estimates for the EU countries are based on national accounts data, known as the "top-down approach". Despite several challenges[2], a top-down estimator of the CIT gap is generally straight-forward and timely. It directly follows the concept of national accounts and requires data from tax returns and macroeconomic statistics, such as GDP. On the other hand, it does not assess the information about individuals who fail to comply with tax legislation. Therefore, the CIT top-down estimate can only be vaguely related to the individual characteristics of firms that evade CIT. By contrast, bottom-up tax gap estimates are effectively built from this information. They deliver tailor-made inputs for the design of effective tax policies and identify sectors or businesses prone to tax avoidance.

[1] While the CIT might be considered a harmful tax for the economic growth (Johansson et al., 2008), the share of the CIT revenues across countries is stable. The reduction in corporate rates across countries since the 1980s was compensated by base-broadening measures (Brys, 2011).

[2] One has to adjust for several conceptual differences to define the theoretical CIT base from macroeconomic data. The IMF's methodology (Ueda, 2018) adopted in Slovakia gives (besides Italy) the only publicly disclosed CIT gap top-down estimates available (see EC, 2018) to date.

The time-delay and constrained validity concerning the population of all firms limit the benefits of the bottom-up approach. It is, therefore, vital to have a complementary estimator available. The two approaches, top-down and bottom-up, are in a sense complementary (Table A in the Appendix gives general comparison), as each one can assess some information, which is invisible for the other method[3]. Focusing on the intersection of their scope, we cover the vast majority of the corporate universe in Slovakia. However, the largest multinational corporations (MNCs)[4] are excluded from the bottom-up target. While MNCs require specific bottom-up treatment, excluding them from the top-down approach for the sake of comparison would require non-trivial methodological adjustments. Therefore, it is currently beyond our scope to have the same target population for the two approaches. Instead, the bottom-up approach, presented here, explores only a part of the noncompliance monitored by the top-down approach, i.e., the SMEs' noncompliance. Any differences between the estimates, including their dynamics, must, therefore, be interpreted with caution. Moreover, being the first of their kind in Slovakia and one of very few in the EU, the numerical result must be interpreted with caution due to the (currently still) limited audit data quality.

Taking all that into account, disentangling the top-down estimate with our bottom-up approach for SMEs (see Figure 1) during the tax period of 2015 shows that the noncompliance gap was almost equally divided between SMEs (EUR 389 million) and MNCs (EUR 453 million). The dynamics of the bottom-up gap for SMEs generally accorded with the decline of top-down estimates in 2015. In 2016, however, the gap for SMEs increased, contrasting with the further decline in top-down estimates. While this discrepancy probably stem from the constrained target of our bottom-up approach and a limited number of audits available for the tax period of 2016, it also pointed to an increased risk of noncompliance among the SMEs.

Closing the major part (98%) of the SMEs' gap would have required more audits targeted on the micro-firms (see Figure 1). However, given the large number

[3]Top-down approach can be easily modified to take into account the "shadow economy".

[4]Slovak tax legislation defines the MNCs as banks, insurance companies and other financial or non-financial corporations that deliver at least during two subsequent tax periods revenues above EUR 40 million. Many of them are following the IFRS accounting standards compared to SMEs subject to domestic accounting standards. Their size and usually international exposure within corporate groups enables them some specific noncompliance opportunities (e.g., transfer pricing). Hence, the selection and audit procedures are different from the rest of the population. While they represent only 1% of the entire corporate universe, they still pay more than half of the actual tax revenues.

of micro-firms and limited capacities of the Financial Administration, potential revenues from these audits would have been very low. Striking a balance between the high revenues from auditing larger firms and the necessity to prevent micro-firms from evasion requires more sophisticated tools, foremost, better audit data (see Section 2). Having a tool to close the gap for all micro-firms, Slovakia would have raised additional EUR 380 million (0.4% of GDP) for the tax period of 2015 (about 10% of the CIT revenues in 2015).

Finally, the Tax Administration conducted about 1 500 audits for the tax period of 2015 and shrank the SMEs' gap by almost EUR 30 million. Based on our bottom-up model, the potential revenues from auditing 1 500 top evaders (among the SMEs) of 2015 were more than four times larger. Over EUR 100 million revenues were lost due to the selection of inactive or fully compliant firms for audits. This miss-selection reduced the number of effective audits by 75%.

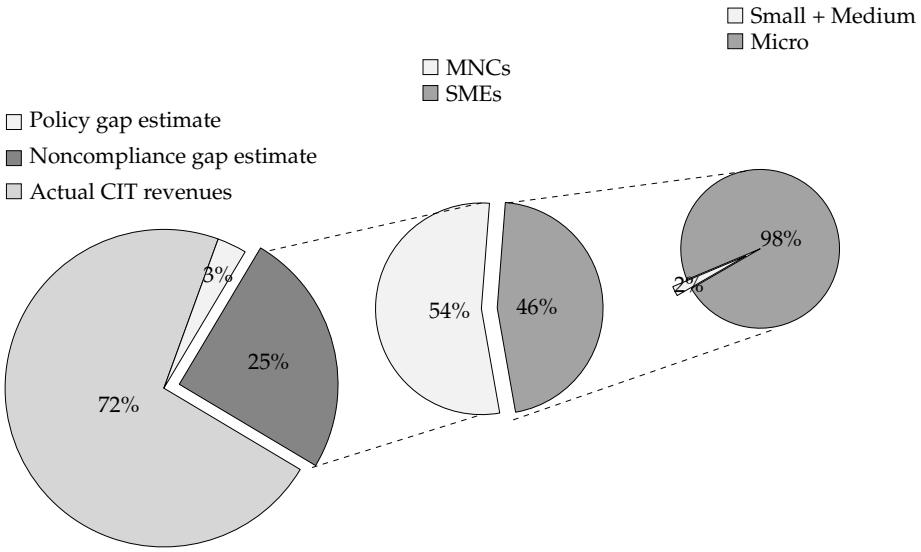

Figure 1: Potential CIT revenues for the tax period of 2015. In a most general sense, CIT gap (28%) consists of the policy gap (3%) and the noncompliance gap (25%), with the latter estimated by the top-down approach. Thanks to the novel bottom-up approach introduced in this paper, the noncompliance gap can be further disentangled into the gap of small and medium enterprises (SMEs, 46% of the noncompliance gap) and large multinational corporations (MNCs, 54% of the noncompliance gap). The Ministry of Finance estimates the policy gap and noncompliance gap on regular basis.

The paper continues as follows: Section 2 discusses risk-based audits, Section 3 explains the econometrics behind the analysis, Section 4 provides details about the data used and gives summary statistics, Section 5 follows the estimation step-by-step, and Section 6 summarizes the CIT gap estimates and compares different alternative approaches. Section 7 concludes.

2. Risk-based audits: curse or blessing?

The primary input for the bottom-up estimation of the CIT gap is a sample of CIT audits. They predetermine the choice of econometric methods, the scope of the analysis, and the results' stability and reliability. Monitoring bottom-up gap estimates seem to be particularly sensitive to the quality of the data assessed during a particular tax period. In the paper, we focus on two major problems of CIT audit data, i.e., the representatives (bias induced by the risk-based selection) and quality of individual tax deficiency detected. Both problems can be addressed with the appropriate choice of the model and the estimator.

Optimally, we would start with a stratified, representative random sample of audits, such as in Kleven et al. (2011). Instead, we have a non-random sample of firms selected for audit based on possibly dynamic and unobservable risk-criteria[5]. This incurs a sample selection bias, i.e., a phenomenon well known in labor economics (Mroz, 1987). On the other hand, observing only the result but not (directly) the causes for audits does not preclude reliable inference. Sophisticated econometric methods may turn the curse of targeted audits into a blessing. However, even the most sophisticated methods would benefit from a larger, more representative sample, and trackable criteria for audit selection. A basic illustration of the quality of the available data is obtained using simple, model-free scaling methods for bottom-up gap estimation (see Section 6, Table 7). While naive, these methods would provide robust and reliable benchmark estimates, if the set of audited firms were sufficiently representative. As we state below, this is not the case with the audit data at hand.

Working through our data carefully, we identify the most severe issues and suggest how to improve the data quality[6]. Foremost, almost 50% of audited firms are either economically inactive or "at least" do not respond to the auditor, who, in

[5]It takes much time to accomplish a sufficient amount of tax audits for a particular tax period. This causes a lag of several years between the targeted tax period and the point when the final tax gap estimate becomes available. Therefore, it is essential to keep records of all selection criteria.

[6]This part of our research complements the assessment of FASR's performance by The Tax administration Diagnostic Assessment Tool conducted in April 2018. See http://www.tadat.org.

turn, cannot determine the exact amount of deficiency. Moreover, from the active subset of audited firms, 43% are found to be fully compliant, i.e., auditor detects no tax deficiency. However, the rules for the selection are (still as of today) non-centralized and not recorded systematically. It is, therefore, impossible to identify any mistakes in the selection mechanism. Second, detected deficiencies are insufficient for compliance risk management. They do not provide enough data about the reason for selecting a firm nor the source of deficiency. These findings imply an urgent need for gathering detailed information about tax audit results in a standardized form, as the minimum requirement for a stable and reliable monitoring mechanism of the CIT noncompliance.

Nonetheless, all improvements in the audits selection and reporting mechanism will affect tax gap estimates of the next tax years only. Given the challenges we just described, in order to estimate the gap for the years of interest, we choose possibly robust econometric tools.

For our tax period of interest (2014-2016), we find that the naive scaling approaches, i.e., imputation based on stratified population, do not sufficiently correct the bias selection. They lead to extremely high predictions of the CIT gap (close to 80%), even higher than the gaps computed on the observed set of audits (close to 60%), which by definition, should provide an upper bound if the audit selection mechanism is not entirely spurious.

Instead, we identify a predictive regression model for the size of the deficiency using firms' characteristics. This task seems as common in the bottom-up tax gap estimation literature. Using non-random audit data, Hanlon et al. (2007) employ a censored regression model (Tobit) (see Tobin, 1958; Powell, 1984) later generalized by Heckman and others (see Heckman, 1974, 1979; Lancaster and Imbens, 1996). In particular, Heckman allows for endogenous sample selection. Alternative approaches include post-stratification (Rosenbaum and Rubin, 1983; Nicolay, 2013) based on the statistical matching of characteristics between audited and non-audited firms. Advanced binary selection model for estimation of noncompliance was developed by Feinstein (2001); Erard and Feinstein (2007); Feinstein and Erard (2010), who use vast US audit data collected under the National research program. They propose a so-called "detection controlled estimation" model, which can account for the bias induced by varying experience among auditors. Alternative approaches include Erard and Ho (2001), optimizing a firm's tax expenditures in the face of potential audits, and Warner et al. (2015), developing algorithms replicating evasion schemes among company owners based on their asset portfolio. Useful cross-country meta-studies and technical reports include Erard (1997) and EC (2018).

3. Econometric models

3.1. Literature overview

The starting model for this paper is the Heckman's model estimable by two-step OLS or maximum likelihood. The major drawback of these estimators is their inconsistency under non-Gaussian noise. The course of econometric literature therefore turned to semi-parametric methods, avoiding these assumptions. Seminal works on this subject include Cosslett (1987); Robinson (1988). Further extension in the direction of non-parametric estimation includes Ahn and Powell (1993) and Das et al. (2003). Most recent advances in both semi- and non- parametric selection bias robust estimators include Chen and Zhou (2010); Escanciano and Zhu (2015); Chen et al. (2018); Honoré and Hu (2018) with real-data applications in Newey et al. (1990); Schafgans (1998); Mora (2008); Huber and Melly (2015). As to our knowledge, we are the first to apply a semi-parametric sample selection model to estimate tax gaps.

As the next challenge for our models, 43% from the completed audits detect no deficiency. Such a high proportion of unsuccessful risk-based controls is suspicious. However, auditors focus on firms with (almost 7 times) higher revenues than what is typical for the population[7] based on the official data. Hypothetically, this can cause that some small amounts of unpaid taxes (up to few euros per firm) will be passed by unnoticed. In such a case, we deal with an additional censoring of the dependent variable on top of the sample selection rule discussed before. Taking this as an assumption, we develop a handy extension of the classical Heckman's model, tailor-made for the data at hand. We call this extension "Censored Heckman".

3.2. The prediction problem

For a fixed tax period, the level of noncompliance (individual deficiency) y_i for each active agent in the population $i = 1, \ldots, 200\,000$ is the difference between agent's (true) tax due and the actual amount paid. We would ideally observe y_i for each i and conclude the distribution $\mathcal{L}(Y)$ and, more importantly, about $\mathcal{L}(Y|\boldsymbol{X})$, with $\boldsymbol{X} \in R^k$ a set of firm's characteristics, using the entire population. In such a case, our results would only be polluted by measurement errors. In a feasible case, we would observe the individual deficiencies for a random sample of agents $i = 1, \ldots, n$. To obtain deficiency estimates for the population, we need

[7] Given the limited financial and personal resources of Financial administration and assuming that the tax deficiency grows with the revenues, auditors focus on those firms with high revenues.

to predict those y_i, which are missing. Assuming that Y depends on $\boldsymbol{X}$, one can estimate a linear model using standard estimators and use the observations $\boldsymbol{x}_i$ to predict the missing y_i's. In case of operational audits, the observations $(y_i, \boldsymbol{x}_i)$, are not drawn randomly. Moreover, it is natural to assume that the selection of firms, for which we observe the deficiencies, has a certain impact on the mean of the observed deficiencies (given a set of firm's characteristics), i.e., that $\mathrm{E}(Y|\boldsymbol{X}) \neq \mathrm{E}(Y|\boldsymbol{X}, \text{selected for audit})$.

3.3. Sample selection: Heckman's Gaussian model

Heckman's approach uses a conditionally normal latent variable Y_S^* to quantify a firm's propensity for noncompliance. A firm is selected for audit on the event $Y_S^* > 0|\boldsymbol{X}_S = \boldsymbol{x}_s$, when conditioned on a set of firm's characteristics $\boldsymbol{X}_S$. We observe Y_S^* indirectly through the binary random variable Y_S. Under the Probit model, $\mathbb{P}(Y_S = 1|\boldsymbol{X}_S) = \Phi(\boldsymbol{\beta}\boldsymbol{X}_S)$, where Φ is the distribution function of the standard normal distribution. The actual deficiencies (outcomes) are assumed to follow a linear model on a set of firms individual characteristics $\boldsymbol{X}_O$, which in general should be different from the first set $\boldsymbol{X}_S$. Heckman approach is based on the following model specification:
Assume that the two real random variables Y_S^*, Y_O^* satisfy

$$Y_S^* = \boldsymbol{\beta}_S' \boldsymbol{X}_S + \varepsilon_S, \tag{3.1}$$

$$Y_O^* = \boldsymbol{\beta}_O' \boldsymbol{X}_O + \varepsilon_O, \quad \text{where} \tag{3.2}$$

$$\begin{bmatrix} \varepsilon_S \\ \varepsilon_O \end{bmatrix} \overset{|\boldsymbol{X}_{S,O}}{\sim} \mathcal{N}_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{bmatrix} \right). \tag{3.3}$$

For $j = 1, \ldots, N$, we observe realizations of $\left(Y_{Sj}, Y_{Oj}, \boldsymbol{X}_{Sj}', \boldsymbol{X}_{Oj}'\right)$, where

$$Y_S \quad \begin{cases} 0 & Y_S^* \leq 0, \\ 1 & \text{else}, \end{cases} \tag{3.4}$$

$$Y_O \quad \begin{cases} \mathrm{NA} & Y_S = 0, \\ Y_O^* & Y_S = 1. \end{cases} \tag{3.5}$$

Heckman (1974) handles the sample selection bias as a problem of omitted significant predictor and proposes a two-step OLS estimator for the model. The outcome equation (3.2) under Heckman's Gaussian parametric model is OLS-estimable when augmented by a bias correction term

$$\mathrm{E}(\varepsilon_O|\varepsilon_S > -\boldsymbol{\beta}_S' \boldsymbol{X}_S) = \rho\sigma\lambda(\boldsymbol{\beta}_S, \boldsymbol{X}_S) \tag{3.6}$$

where

$$\lambda(\boldsymbol{\beta}_S, \boldsymbol{X}_S) = \frac{\phi(\boldsymbol{\beta}_S' \boldsymbol{X}_S)}{\Phi(\boldsymbol{\beta}_S' \boldsymbol{X}_S)}, \tag{3.7}$$

is the inverse Mill's ratio. I.e., the outcome equation is

$$Y_O = \boldsymbol{\beta}_O' \boldsymbol{X}_O + \rho\sigma\lambda(\boldsymbol{\beta}_S, \boldsymbol{X}_S) + u, \quad \text{where} \quad \mathrm{E}(u|\boldsymbol{X}_S) = 0. \tag{3.8}$$

Both $\hat{\boldsymbol{\beta}}_O^{OLS}$ and a more efficient Gaussian $\hat{\boldsymbol{\beta}}_O^{ML}$ estimators (implemented in R by Toomet and Henningsen (2008)), turn to be inconsistent for non-Gaussian joint noise $(\varepsilon_S, \varepsilon_O)$. However, Paarsch (1984) showed, in a large Monte-Carlo experiment that in finite samples, efficiency gains of the ML overcompensate the inconsistency. Still, even wen correctly specified, the likelihood function is non-convex; therefore, good starting values are needed to avoid local extrema (see Chen and Zhou, 2010, for a discussion).

3.4. Sample selection: Robust semi-parametric two-step model

If the joint distribution of the errors is miss-specified, both OLS and ML yield inconsistent estimators. A consistent semi-parametric estimator can be obtained by assuming that (3.1) and (3.2) hold, while $\mathbb{P}(Y_S = 1|\boldsymbol{X}_S) = F(\boldsymbol{\beta}\boldsymbol{X}_S)$, where F is some general link function (not necessarily Gaussian) for the binary selection model (3.4).

The key input for the semi-parametric two-step estimator of $\boldsymbol{\beta}_O$ is the estimator of function F, the estimator of coefficients $\boldsymbol{\beta}_S$, and the correction term $\lambda(\cdot)$. For the consistent estimation of F and $\boldsymbol{\beta}_S$, several approaches are available. We follow Newey et al. (1990); Newey (2009), and use the efficient quasi-maximum likelihood approach of Klein and Spady (1993). The estimator of each $F_j = F(\boldsymbol{X}_{S,j}\boldsymbol{\beta}_S), j = 1, \ldots, N$ is obtained as $\hat{F}_j = (\boldsymbol{X}_{S,j}\boldsymbol{\beta}_S) \sum_{i=-[h/2]}^{[h/2]} d_i K(\frac{(\boldsymbol{X}_{S,j} - \boldsymbol{X}_{S,i})\boldsymbol{\beta}_S}{h})$, where $K(\cdot)$ is a kernel function. The bandwidth h is a nuisance parameter which controls the smoothness of $\hat{F}$. As advocated by Klein and Spady (1993), it can be obtained using a generalized cross validation[8] as proposed by Craven and Wahba (1978).

Next, the bias correction term $\lambda(\cdot)$ in (3.8) should be estimated. Following Cosslett (1987); Newey et al. (1990) and Newey (2009), we use a series approximation, i.e., $\lambda(\boldsymbol{\beta}_S, \boldsymbol{X}_S) \approx \sum_{k=1}^{K} a_k p_k(\tau(\boldsymbol{\beta}' \boldsymbol{X}_S))$, where $p_k(\cdot), a_k, k = 1, \ldots, K$ are the

[8]The estimator of Klein and Spady (1993) is implemented in the R-package *np*.

basis function and the unknown coefficients respectively. The basis functions are chosen as power series, i.e., $p_k = p^k$. The number of these functions can be also chosen by the same generalized cross validation criteria. The transformation $\tau(\cdot)$ can be also chosen in different ways (see, e.g. Newey, 2009). We use $\tau = \frac{\phi(\cdot)}{\Phi(\cdot)}$, because it preserves the shape of λ from the parametric model. Hence in our case, the leading term in the power series approximation is the inverse Mill's ratio.

Finally, the estimator of $\boldsymbol{\beta}_O$ can be obtained as the classical IV estimator. Denote, X_O, Y_O and $\hat{P} = (\hat{p}_1, \ldots, \hat{p}_K)$ the respective data matrices for audited firms only with n rows. Moreover, denote $\hat{Q} = \hat{P}(\hat{P}'\hat{P})\hat{P}'$ and $\hat{M} = X_O'(I - \hat{Q})X_O/n$. Then the unbiased SP estimator for $\boldsymbol{\beta}_O$ is

$$\hat{\boldsymbol{\beta}}_O^{SP} = \hat{M}^{-1}X_O'(I - \hat{Q})Y_O/n. \tag{3.9}$$

3.5. Sample selection: Censored Heckman

As suggested in the introduction, our data on observed levels of noncompliance have a mixed distribution with a large proportion of (seemingly) compliant firms. I.e., in about 43% of cases $y_i = 0$ (see Figure A in the Appendix). It is a priori not clear, whether these amounts are truly 0's, or whether the auditors in their search for large deficiencies, could have ignored some tiny amounts (e.g., less than few dozens euros per firm). If the 0's are true, this phenomenon points to a false risk-based firm selection mechanism, and the audited sample could be relatively similar to a random sample. By contrast, if the second hypothesis is correct and the auditors overlook deficiencies below certain threshold Y_T, then the outcome equation (3.2) requires a corresponding - and to our knowledge, new - extension, driven by this conjecture. We call this new specification, Censored Heckman (CH).

The model specification under the twice censored dependent variable is same as in (3.1) and similar as in (3.2) but in addition[9] we have

$$Y_S \quad \begin{cases} 0 & Y_S^* \leq 0, \\ 1 & \text{else}, \end{cases} \tag{3.10}$$

$$Y_O \quad \begin{cases} \text{NA} & Y_S = 0, \\ 0 & Y_S = 1, \text{ and } Y_O^* \leq Y_T, \\ Y_O^* & Y_S = 1, \text{ and } Y_O^* > Y_T, \end{cases} \tag{3.11}$$

[9]The model can also be seen as a model for twice-censored data, i.e., a Tobit-2 with an embedded Tobit-1 model using the classification from Amemiya (1985, page 360).

and where $0 \leq Y_T < \infty$ is the known threshold. We can estimate the unknown parameters $\theta = \left(\boldsymbol{\beta}_S', \boldsymbol{\beta}_O', \rho, \sigma\right)$ using maximum likelihood which has the form (for one observation)

$$L(\theta|y_o, y_s, \boldsymbol{x}_S, \boldsymbol{x}_O, Y_T) = \tag{3.12}$$

$$\begin{aligned} &= \mathbb{P}\left(Y_O = \text{NA}\right)^{1-y_S} \mathbb{P}\left(Y_O = 0, Y_S = 1\right)^{y_S I\left\{y_O^* \leq Y_T\right\}} f_{Y_O, Y_S}(\theta, y_O, y_S)^{y_S I\left\{y_O^* > Y_T\right\}}, \\ &= \left(1 - \Phi\left(\boldsymbol{\beta}_S' \boldsymbol{x}_S\right)\right)^{1-y_S} \Phi_2\left(\frac{Y_T - \boldsymbol{\beta}_O' \boldsymbol{x}_O}{\sigma}, \boldsymbol{\beta}_S' \boldsymbol{x}_S, -\rho\right)^{y_S I\left\{y_O^* > Y_T\right\}} * \ldots \\ &\quad \ldots * \left(\frac{1}{\sigma}\phi\left(\frac{y_O - \boldsymbol{\beta}_O' \boldsymbol{x}_O}{\sigma}\right) \Phi\left(\frac{\rho/\sigma\left(y_O - \boldsymbol{\beta}_O' \boldsymbol{x}_O\right) + \boldsymbol{\beta}_S' \boldsymbol{x}_S}{\sqrt{1-\rho^2}}\right)\right)^{y_S I\left\{y_O^* > Y_T\right\}}, \end{aligned}$$

where f is the joint density of Y_O, Y_S, and $\phi, \Phi,$ and Φ_2 are the pdf, cdf and joint cdf of the standard normal random variables with correlation coefficient ρ.

4. Data

The estimation of the CIT gap is based on various individual characteristics of firms subjected to tax legislation in Slovakia. Therefore Institute for Financial Policy collected a large panel of these characteristics into the "firm database" (FD). This section summarizes the content of FD and provides basic descriptive statistics for variables that we eventually use for estimation of the CIT gap.

4.1. Scope, sources, and limitations

The complexity of the firm-level data is threefold: *Cross-sectional* scope is exhaustive, opened to the whole population of approximately 200 000 firms active in Slovakia each year. FD contains only legal entities that are liable to CIT. Unincorporated entities such as sole proprietorship liable to personal income tax (PIT) are not present. *Serial* scope is limited, and the most reliable data are available for the period 2014-2018. *Individual* information includes both non-financial (firm's legal, sectoral, geographical, and social profile) and financial (mostly fields from the financial statement as well as CIT returns and audits) characteristics.

FD relies on five key sources: Registry of financial statements - RFS, Statistical Office of the Slovak Republic - SOSR, Social insurance agency - SIA, FinStat s.r.o.

and Financial administration of SR - FASR. The general rule is that one source supplements the other if the former is richer and more precise[10].

Concerning the limitations of FD, we point out that there is only imprecise information about the actual number of active firms in a particular year. This is because FASR's registry of taxpayers contains many firms, which are economically inactive yet did not deregister, thus are considered active.

Furthermore, special attention is paid to the set of audited firms, which are the key input for the bottom-up CIT gap estimation. It is only for this set of firms that we observe the amount of noncompliance. Several aspects of the audits need to be taken into consideration.

First, the results of CIT audits are available with a delay of at least a few months. The average length of a tax audit is about nine months. However, in some cases, because of legal appeals, the complete results may also be delayed by several years. Therefore, currently, we have collected data on finalized audits up to the tax year 2016 (including). Also, due to changes in the information systems of FASR at the end of 2014, only information about tax audits finalized over the period from April 2015 to June 2018 is available.

Second, results of an audit are bound to the audited tax period; therefore, when looking for common patterns among firms, their characteristics must match with this period (not the period during which audit took place).

Third, the number of handy audits is tiny compared to the population size (see Section 4.2). This has objective reasons since FASR works with a limited budget, and thorough audits are expensive. However, we have identified several problems that further decrease the amount of usable data, and we list these in the Appendix. To correct for the low number of usable audits available each year, we merge all audits across the entire range of tax periods (2014-2016)[11], striking a compromise between quality and the quantity of the historical data.

4.2. Data-cleaning procedure

The raw panel for the tax period 2014 - 2016 (with the most reliable data available) contains approximately 300 000 unique firm-year entries per year.

[10]The primary source of non-financial information is SOSR. However, the quality of the data is the least precise one. Much of SOSR's information, such as legal form, NACE code for the sector or size category, is self-reported by firms and might be wrong or missing. Moreover, SOSR does not verify nor update this data unless the firm has over 19 employees, which, as we will see, most companies do not have.

[11]Tax audits conducted and finalized in the period from March 2015 to December 2018.

First, some of the individual characteristics providing similar or complementary information are merged. For instance, we augment the number of employees by the number of the executive (a statutory body). This helps to reduce the missing values occurring mostly with micro-firms where the executive person may also be an employee but has no obligation to report it.

Next, variables with too many categories are simplified. For instance, we simplify the ownership and NACE sector classification in order to distinguish domestic from foreign firms, and 12 main business sectors (see further details below).

Finally, from the 117 individual and financial characteristics, we select the 9 most populated variables[12] including most of the non-financial characteristics and the key high-level aggregates from financial statements and tax returns. The complete list of all variables available is part of the supplementary material.

Given that many entries in the panel contain missing or invalid values or need to be excluded from the analysis for the sake of comparison with the top-down estimate[13], we apply further cleaning steps, which are described in the Appendix. Table 1) below shows the number of remaining entries after each of these steps.

4.3. Descriptive analysis

We present a summary of the selected characteristics of all firms (after cleaning). A detailed summary for all characteristics[14] used for stratification of firms can be found in the Appendix (see Tables B and C). These characteristics show a stable distribution across the three periods under consideration.

First, we are aware that in the cleaning step (iii), we omit, among others, the large multinational corporations (i.e., 1% of the entire population) from the active population. This reduces the total revenues, and the total declared tax revenues by over 50%. Nonetheless, we argue that the omitted firms have different characteristics and behavior due to their size and by far more intense tax controls, than the rest 99% of the population. Our analysis, therefore, focuses exclusively on

[12]These include the Number of employees, NACE Sector, Ownership (domestic/foreign), Administrative region, Revenues, Profit/Loss before tax, Costs, Net assets, Value added.

[13]Details on why we exclude individual firms from the top-down estimate are in Ueda (2018).

[14]These characteristics include 8 administrative Regions: *Bratislava, Trnava, Trenčín, Nitra, Žilina, Banská Bystrica, Prešov and Košice*, 12 sectors: *Agriculture, Industry, Construction, Wholesale & Retail, Transport, Accommodation, Information, Finance, Real Estate, Specialized services (advisory, architects, surveys...), Supporting services (employment agencies, travel agencies...) and Others (dental, ambulance and social care, retail on street markets, supporting services for performing arts and sport events).*, Ownership type: *domestic and foreign* and 4 size categories by a number of employees and total revenues (see Table 2).

Number of firms after	Population			Audits
	2014	2015	2016	2005-16
(0) raw data	308 363	309 111	307 768	4 776
(i) omit missing tax ID	236 481	237 521	233 446	4 776
(ii) omit firms with missing data	186 970	187 868	198 005	4 776
(iii) omit non-profit, MNCs, etc.	184 964	185 962	196 572	4 749
(iv) filter only audits conducted for tax-period 2014-2016				3 251
(v) omit audits where firm did not respond				1 574
(vi) omit audits targeted at "the minimum income tax"				1 432
(vii) omit repetitive firms over consecutive years				1 336
(viii) omit audits resulting in negative tax adjustment				1 335
(ix) omit audits resulting in reduction of the declared tax loss				1 153
(x) omit audits with missing characteristics				1 126

Table 1: List of all cleaning steps applied to the raw panel of individual characteristics for the entire population of firms active during the years 2014-2016. Note that the audits considered here were conducted for the tax periods 2005-2016, but the actual audits took place between 2015-2018. The steps are explained in more detail in the Appendix.

Size category	Classification rules		Number of firms	
	Number of empl.	Total revenues	Population	Audits
Micro	≤ 9	≤ 2 Mill. €	162 002	505
Small	≤ 49	≤ 10 Mill. €	19 763	453
Medium	≤ 249	≤ 50 Mill. €	3 793	151
Large (SMEs)	≥ 250	≥ 50 Mill. €	404	17

Table 2: After excluding about 1% of firms from the population (e.g., MNCs) in cleaning step (iii) (see Table 1), we further divide the targeted population into 4 size categories based on the classification of Small and medium-sized enterprises in EU recommendation 2003/361. The last category is for firms which do not fulfill both conditions for medium-sized firms simultaneously. Usually, these Large (SMEs) become MNCs over time if they reach revenues over 40 million at least during two subsequent tax periods. If such conditions has not been full-filed, EU classification applies.

small end medium enterprises further divided into 4 size categories, as explained above. The vast majority (88%) of active firms are micro-firms. Moreover, at least one third of them has only (up to) one employee. The rest of the population in-

cludes mainly small businesses with revenues up to EUR 10 million per year. By contrast, the number of medium and large (SMEs) firms is negligible. The representation of each category in the audits is more balanced compared to the population. For instance, the share of micro-firms is reduced to half, while the share of other categories is several times higher than in the population. Table 3 shows summary statistics of the observed deficiencies for each size category. Among those firms selected for an audit based on risk criteria, small and medium firms both evade (in total) twice as much as micro-firms, which seems intuitive, yet, we have to keep in mind that the proportion of micro-firms in the population is almost twice as high. Among the micro-firms, the smallest ones evade, typically[15], almost 10 times more than the rest. Finally, medium firms have the highest total and typical evasion.

The distribution of audits among the 12 sectors of the economy is relatively close to the population. The leading sector in terms of the relative frequency is Wholesale & Retail (25% population vs. 34% audits) followed by Specialized services (16% population vs. 7% audits) resp. by Construction (11% population vs. 21% audits). The vast majority of firms are domestic (85% population vs. 82% audits). The evasion is concentrated in the largest Wholesale & Retail sector (47%), while the typical level of evasion is the highest in Agriculture. The highest evasion per firm is in Supporting sector. Domestic firms have higher total and typical evasion, foreign firms have higher evasion per firm.

Concerning the financial characteristics such as labour costs, revenues, profit/loss before tax, our audits are genuinely focused on larger SMEs (both in terms of capital and employees). In particular, the median yearly revenues of an audited firm are 30 times larger than in the population. By contrast, the declared tax base (relative to total revenues) among audited firms was typically smaller than in the population. Finally, the effective tax rate based on the declared tax base and tax were close to the nominal level (22%) both in the population and among audited firms.

[15]The empirical distribution of deficiencies has a log-normal shape. However, even after log-transformation, skew to the right is present, which is why we prefer to use rank-statistics (e.g., with "typically", we refer to the median) when possible. For modeling, all financial variables are sign-log-transformed, i.e., the absolute values are transformed to logarithms and then multiplied by respective sign.

	Min €	Mean €	Median €	Max Mill. €	Total Mill. €
Size					
Micro	0	16 132	240	2.07	10.28
	-	33%	40%	55%	19%
thereof ≤ 1 *employee*	0	34 709	2 114	0.86	3.99
	-	71%	352%	23%	7%
Small	0	43 436	727	1.47	19.68
	-	89%	121%	39%	36%
Medium	0	153 956	1 009	3.76	23.25
	-	314%	168%	100%	42%
Large	0	120 888	173	1.10	2.06
	-	246%	29%	29%	4%
Region					
Bratislava	0	30 825	176	2.00	7.71
	-	62%	29%	53%	13%
Trnava	0	100 499	-	3.76	12.86
	-	204%	0%	100%	23%
Trenčín	0	22 979	1 803	0.64	3.03
	-	46%	300%	16%	5%
Nitra	0	104 823	3 353	3.25	11.21
	-	214%	558%	87%	20%
Žilina	0	72 442	2 711	3.31	11.66
	-	147%	451%	88%	21%
B. Bystrica	0	32 301	695	0.86	4.46
	-	65%	115%	22%	8%
Prešov	0	16 415	1 293	0.33	2.00
	-	33%	215%	8%	3%
Košice	0	26 352	-	1.27	2.32
	-	53%	0%	33%	4%

	Min €	Mean €	Median €	Max Mill. €	Total Mill. €
Sector					
Accommmodation	0	10 828	319	0.15	0.49
	-	22%	53%	3%	1%
Agriculture	0	54 211	6 307	0.53	0.87
	-	110%	1049%	14%	1%
Construction	0	40 980	2 387	3.25	9.71
	-	83%	397%	86%	17%
Industry	0	35 250	542	1.20	4.79
	-	71%	90%	32%	9%
Information	0	91 569	337	1.74	2.93
	-	186%	56%	46%	5%
Others	0	8 286	64	0.14	0.23
	-	17%	11%	4%	0%
Real Estate	0	36 138	0	1.47	1.62
	-	73%	0%	39%	3%
Specialized	0	42 734	528	1.27	3.33
	-	87%	88%	34%	6%
Supporting	0	72 212	2 608	1.10	4.04
	-	147%	433%	29%	7%
Transport	0	16 928	665	0.15	1.08
	-	34%	111%	4%	2%
Wholesale& Retail	0	67 368	132	3.76	26.14
	-	137%	22%	100%	47%
Ownership					
Domestic	0	44 062	845	3.76	40.76
	-	89%	141%	100%	74%
Foreign	0	72 162	0	2.52	14.50
	-	147%	0%	67%	26%

Table 3: Nominal and relative summary statistics of deficiencies in each category of selected characteristics detected in the set of all firms (after cleaning) audited for the tax years 2014-2016. The relative values in % are computed with respect to the respective statistics based on all 1 126 audits.

5. Bottom-up CIT gap: estimation step-by-step

In this section, we describe the implementation of the methods mentioned in Section 3 and discuss several crucial steps in more detail.

5.1. Data preparation

Problems which we tackled by cleaning steps described in the previous section lead to a dramatic decrease of already modest audit set. To circumvent this issue and to obtain the best possible estimate of the CIT gap for the entire population and each of the tax-periods 2014, 2015, and 2016, we decided to merge the audits over these 3 years. For the sake of robustness, our results presented in the next section are based on both approaches, i.e., with and without merging.

We start by cleaning and merging the audit data for the tax period 2014-2016 following the steps described in Section 4. The resulting merged set of audits contains 1126 different firms and is used as such to estimate model parameters for each tax period. However, as will be clear from the next paragraph, the estimated parameters of all models differ from one tax period to the next, because the non-audited firms change. For the sake of space, we provide a detailed description of the estimated models only for the tax period 2015[16].

Next, for a given tax period, say 2015, we first apply the same cleaning to all firms. From the population, we draw a random sample of 5000 (2.7% of the population) non-audited firms[17], i.e., excluding all firms present in the audited set. The balanced sample contains 5000 randomly selected non-audited companies together with the 1126 audited companies.

The balanced sample is augmented by the 9 explanatory variables and their interactions (e.g., Revenues per Number of Employees). The non-negative variables are transformed to logarithms and screened for outliers[18].

[16]The results across the years are similar because we use the common set of audited firms. Possible differences arise only due to the set of non-audited firms, which enter into the binary selection model. These results are available from the authors upon request.

[17]The size of the choice-based sample of non-audited firms is supposed to provide a trade-off between the information loss due to sampling and the frequency imbalance between non-audited and audited firms. Note that estimating the probit model on the entire population of firms would make the model too parsimonious to choose any firm for audit. Our robustness checks include alternative random samples of size 2000 and 8000 and can be obtained from authors on request.

[18]We replace those observations whose distance from the sample median is more than 5 times the inter-quartile range by the respective lower or upper bound of the outlier filter.

5.2. Estimation of models and prediction

All our CIT Gap estimates from the three different Sample selection model specifications (i.e., Gaussian Heckman, Two-step semi-parametric, and Censored Heckman) introduced in Section 3 are obtained from the same balanced sample of firms. There are 4 key steps starting with the choice of variables and ending with the application of the models:

(i) selection of the best subset of predictors X_S and X_O for (3.1) and (3.2),
(ii) estimation of the binary selection model,
(iii) estimation of the outcome equation (3.2) resp. (3.8),
(iv) prediction from the estimated model.

Model selection. When there are p un-nested predictors in the initial set, there are always 2^p possible models to be estimated. Each model will give a different prediction of the gap. Our goal is to select the best subset of predictors. All variables from the initial set can enter either into the selection equation (3.1) via X_S or the outcome equation (3.2) via X_O or both. We utilize the LASSO estimator, a popular model-selection and estimation tool when the focus is on predictive performance rather than the in-sample fit, to decide which variables to include or not. Tibshirani and Knight (1999) proposed the LASSO estimator

$$\hat{\boldsymbol{\beta}}^{\text{LASSO}} = \underset{\boldsymbol{\beta}\in\mathcal{R}^K}{\operatorname{argmin}} \left(-\log L(\boldsymbol{\beta}|y_o, y_s) + \lambda \sum_{j=1}^{K} |\beta_j| \right), \tag{5.1}$$

with the tuning parameter $\lambda > 0$. The first part of LASSO's objective function is the standard log-likelihood of binary model resp. the sum of squared errors for the outcome regression (3.2). The strength of regularization depends on the tuning parameter λ which we select by 10 fold cross-validation with respect to the chosen objective. For binary selection model, the objective is to correctly classify firms, therefore, we select the λ which minimizes the miss-classification error, i.e.,

$$\lambda = \operatorname{argmin} \frac{1}{6124} \sum_{j=1}^{6124} I(\hat{Y}_{S,j}(\lambda)), \tag{5.2}$$

where $I(\hat{Y}_{S,j}) = 1$ whenever the firm was falsely classified, i.e., $\hat{Y}_{S,j} \neq y_{S,j}$ and 0 otherwise. For the outcome equation (3.2) λ is obtained by minimizing the mean absolute error. The best subset of predictors for equation (3.1) resp. (3.2) is the subset $M \subseteq \{1, \ldots, p\}, s.t., \forall i \in M, \hat{\boldsymbol{\beta}}_i^{\text{LASSO}} \neq 0$. For the tax period 2015, LASSO

selected 21 common variables for X_S and X_O, while 11 variables were selected only for the binary model but not for the outcome equation[19].

Estimation of the binary selection model. Sample selection models use a single index score obtained from the respective binary selection model (Probit model or a semi-parametric model) based on individual characteristics. This binary selection is a proxy for FASR's internal audit selection. A large discrepancy in the frequency of the audited and non-audited firms requires additional up-weighting of the audited firms in the balanced sample of 6126 firms. Each firm is weighted to reflect the presumed proportion of each category (audited or non-audited) in the entire population. The optimal weights according to Manski and Lerman (1977) are

$$w_i = \frac{\text{presumed proportion of } i \text{ the population}}{\text{proportion of } i \text{ in the choice-based sample}}, \quad \text{for } i \in \{\text{audited}, \text{not}\}. \tag{5.3}$$

The practical problem is that due to the limited resources of the tax authority, many firms in the population face no risk of being selected for audit. We do not know how many firms would have been subjected to audit if FASR's resources were unlimited. Therefore, we take the nominator in (5.3) as a hyper-parameter whose value is inferred from data in order to maximize the selection performance of our binary model. We require the sensitivity[20] on the entire population to be at least 50%. We achieve this by employing a constant $c \geq 1$, whose role is to control for the parsimony of the binary model. By taking larger c, we increase the tendency of the binary model to select firms for audit. The resulting weights are:

$$w_i^c = \begin{cases} \dfrac{1126c/n}{1126/6126}, & i = \text{audited}, \\ \dfrac{(n-1126c)/n}{5000/6126}, & i = \text{not}. \end{cases} \tag{5.4}$$

[19]Technically, we fulfill the exclusion restriction. However, there is hardly any intuitive explanation for this particular allocation of predictors between the binary selection and the outcome equation. Still, we do not deem this to be of any limitation of our results since we are primarily interested in predictions rather than in causal implications.

[20] Given that we have very few audits available, the out-of-sample sensitivity of the binary model is of very high importance. By sensitivity, we mean the proportion of firms selected by the binary model in the set of all audited firms. Alternatively, we could measure the performance of the binary model using the specificity and the "area under the receiver operating characteristic curve" - AUC, which measures the probability that for given random pair of firms, each one from a different category, the model assigns a higher score to the audited firm.

Note that if $c = 1$, we implicitly assume that the authority picked all the suspicious firms from the entire population. However, such weights turn out to be impractical, since they imply that the binary again does not select any firm for audit. We, therefore, look for the smallest $c > 1$ for which the sensitivity of the binary model achieves at least 50%. Search on the grid of whole numbers (over all three tax periods) found this value to be $c = 50$, and we keep it fixed for all tax periods.

Table 4 compares these characteristics for binary model estimated without weights and with weights (5.4) including binary models with $c = 1$ and $c = 50$. The higher choice of c increases the AUC by 6pp, and guarantees the sensitivity above 50%. The cost of having higher sensitivity is higher miss-classification error (5.2) on the entire population. However, given the value of the audited firms, the preference is given to the least parsimonious model.

	no weights			weights with $c = 1$			weights with $c = 50$		
	2014	2015	2016	2014	2015	2016	2014	2015	2016
Weight values									
audited	-	-	-	0.03	0.03	0.03	1.64	1.63	1.54
non-audited	-	-	-	1.22	1.22	1.22	0.85	0.86	0.88
Probit fit summary									
sensitivity (%)	44	41	44	0	0	0	68	63	64
specificity (%)	95	95	95	100	100	100	90	90	91
AUC (%)	89	89	89	84	82	83	90	89	89
miss-classification (%)	5	5	5	0.60	0.60	0.57	10	10	9

Table 4: Comparison of 3 alternative weighted Probit models. Predictions for the entire population of firms. Characteristics include the area under the ROC curve (AUC). For each tax period, 5000 randomly chosen non-audited and 1126 audited firms are used for model estimation. Weights used in the middle and the right block are of Manski and Lerman (1977). These weights are computed either using the actual proportion of audited firms in the population ($c = 1$) or with $c = 50$, which guarantees the sensitivity being $> 50\%$.

Estimation of the outcome equation. With predictors in X_S and X_O selected, and with the starting values for the binary model from the previous paragraph, we estimate the Heckman-type models by maximizing the Gaussian log-likelihood[21] and the

[21]Note that log-likelihood maximization for Heckman-type models often suffers from strong sensitivity towards the starting values. We, therefore, compared the results from models where

semi-parametric two-step approach as described in Section 3.

The semi-parametric (SP) two-step approach offers a robust tool for the estimation of sample selection models. It relaxes the restrictive distributional assumptions required by the parametric ML approach. The price for the robustness is a lower precision in cases when the parametric model is admissible, i.e., when ML and semi-parametric approaches yield similar models. In this case, one should undoubtedly prefer the parametric approach. Newey et al. (1990) used a Hausman-type test to verify such a hypothesis. Comparing the two model specifications (ML and SP) means testing for

$$\chi_K = B' \text{AVar}^{-1} B, \tag{5.5}$$

where $B = \hat{\boldsymbol{\beta}}_O^{SP} - \hat{\boldsymbol{\beta}}_O^{ML}$, is the difference between the two estimators, and AVar^{-1} is the inverse of the asymptotic covariance matrix of B. The asymptotic distribution of (5.5) is χ^2 with K degrees of freedom[22] . The asymptotic covariance matrix of B can be approximated using the difference of the respective asymptotic covariance matrices[23] of $\hat{\boldsymbol{\beta}}_O^{SP}$ and $\hat{\boldsymbol{\beta}}_O^{ML}$ (see Proposition 3 in Li and Stengos, 1992). The results of this test are in Table 5 and indicate that the two model specifications result in statistically similar outcome estimates. This is a justification for proceeding with the parametric Gaussian ML estimators, since they provide more efficient estimates.

Further evidence of sample selection and censoring. If (5.5) is used on the difference of $\hat{\boldsymbol{\beta}}_O$ estimated from a (weighted) OLS linear model against ML or SP from sample selection model, the rejection can be interpreted as evidence of sample selection bias in the underlying balanced sample. Table 5 gives strong evidence of the presence of sample selection bias in the data, i.e., for using sample selection models over the classical OLS or ML linear models. More evidence for accounting for selection bias and censoring is in Table 6. It gives the values of the log-likelihood function for the ML estimator. Notably, censoring without sample selection improves the fit of a linear model. The log-likelihood values for sample selection models are smaller than for the other two models, but this is because these models are estimated from 6 126 observations, which is more than 1126 used in the other models. Finally, the presence of sample selection bias is supported by the

we used starting values obtained by LASSO with those obtained by original Heckman's two-step OLS, concluding that the latter provided more robust results.

[22] Degrees of freedom equal to the difference in the actual number of predictors in (3.2) and (3.8)

[23] The asymptotic covariance matrix for the SP estimator can be found in Newey (2009).

significant values of the correlation parameter ρ from (3.3), which measures the "magnitude" of selection bias.

Summary statistics of the estimated weighted OLS, Gaussian (ML) Heckman, and the Censored Heckman models for the tax year 2015 can be found in the Appendix (Tables D).

Models estimated on			all 1126 audits			year-specific audits		
Hypotheses	DF	CV	2014	2015	2016	2014	2015	2016
(1) vs. (2)	1	3.84	85.82	74.38	86.29	480.68	122.85	210.92
(1) vs. (3)	4	9.49	75.26	63.81	82.83	124.02	163.92	78.6
(2) vs. (3)	4	9.49	1.40	1.88	2.73	0.79	1.25	2.98

Table 5: Hausman-type test of model specification (1) OLS regression without sample selection, (2) Gaussian max-likelihood sample selection, (3) semi-parametric sample selection. The values in columns 4-9 are the test statistics (5.5) which follows a χ^2 distribution with DF degrees of freedom. CV is the respective 95% critical value. The null hypothesis is that the two model specifications provide statistically similar estimates of $\boldsymbol{\beta}_O$ from (3.2).

	Free parameters			Log-likelihood (thous.)			$\hat{\rho}$		
	2014	2015	2016	2014	2015	2016	2014	2015	2016
Linear reg.	23	24	24	-3.31	-3.31	-3.31	-	-	-
Censored reg.	25	25	25	-2.47	-2.47	-2.47	-	-	-
Selection reg.	58	59	59	-7.03	-7.35	-7.11	0.47***	0.51***	0.53***
Censored sel. reg.	59	60	60	-4.77	-4.76	-4.66	0.32	0.36	0.25
Signif. codes:	. p<0.05; *p<0.01; **p<0.001; ***p<0.00								

Table 6: Summary of regression models estimated by maximum likelihood using all 1 226 audited firms. The middle panel gives the attained maxima of log-likelihood by the Nelder-Mead maximization algorithm implemented in R (Henningsen and Toomet, 2011). The right panel gives the estimated parameter ρ from sample selection model (3.3), which measures the "magnitude" of selection bias.

Prediction of the outcome. With our estimated models, we predict the deficiency for the entire population of firms active in a particular tax period. These deficiencies are then summed up as $\text{TD} = \sum_{i=1}^{N} \hat{y}_i$ to get bottom-up CIT gap estimates defined as

$$\text{CIT}_{\text{gap}} = \frac{\text{TD}}{\text{TD} + \text{TR}}, \tag{5.6}$$

where TD is the *estimated* total deficiency, and TR is the total of *observed* individual CIT revenues for the targeted population.

As already mentioned, all of our models use the transformed dependent variable, i.e., y_i is, in fact, the log-deficiency. Assuming that the outcome equation (3.2) is correctly specified, we have the best linear unbiased predictor for the log-deficiencies of firm $i \in \{1, \ldots, N\}$

$$\hat{Y}_{O,i} = X_{O,i}\hat{\boldsymbol{\beta}}_O. \tag{5.7}$$

However, the back-transformed predictor is not unbiased for the original (non-logarithmic) deficiencies, because $\mathrm{E}e^{\varepsilon_O} \neq 1$ in general. The biased predictor $e^{\hat{Y}_{O,i}}$ could be easily bias-corrected using a multiplicative factor[24] $\mathrm{E}e^{\varepsilon_O}$. This yields the transformation-bias corrected predictor of the total deficiency (TD) for the population of firms:

$$\mathrm{TD}_{\text{unbiased}} = \sum_{i=1}^{N} e^{\hat{Y}_{\mathrm{O,i}}} * \hat{\mathrm{m}}_n, \tag{5.8}$$

where $\hat{\mathrm{m}}_n$ is sample median of $e^{\hat{\varepsilon}_i}, i = 1, \ldots, n$. Ignoring the bias induced by the log-transformation is yet another possible approach. Ignoring bias means shrinking the multiplicative bias correction factor $\hat{\mathrm{m}}_n$ to 1 (which is equivalent to shrinking of an additive parameter to 0).

$$\mathrm{TD}_{\text{shrinked}} = \sum_{i=1}^{N} e^{\hat{Y}_{\mathrm{O,i}}} * 1. \tag{5.9}$$

It is a priori not clear, whether correcting for transformation bias leads to better out-of-sample predictions[25]. Empirical support for the opposite exists (Bårdsen and Lütkepohl, 2011). We compare the predictive accuracy of the two alternative predictors (5.8) and (5.9) using simulation, which closely follows our real data. However, for the sake of simplicity, we presume that the population consists of noncompliant firms only and disregards the sample selection and censoring issues in our simulation setup. The simulation results suggest that both predictors perform similarly in terms of prediction error, and therefore, we decide to employ the shrinkage predictor (5.9) as our final predictor. The alternative CIT Gap estimates of the unbiased predictor (5.8) can be found in the Appendix together with a detailed description of the simulation setup and R code.

[24] However, this factor needs to be estimated from the in-sample residuals of the model. Note that the distribution of e^{ε_O} is heavy-tailed and right-skewed hence using the sample median of the transformed residuals instead of sample average to estimate the mean of e^{ε_O} is preferable.

[25] Due to the estimation and specification uncertainty, shrinkage estimators have often better finite sample properties than the theoretically "optimal" unbiased estimators.

		CIT Gap (%)			TD (Mill. €)			mean D (€)			median D (€)			max D (Mill. €)		
		2014	2015	2016	2014	2015	2016	2014	2015	2016	2014	2015	2016	2014	2015	2016
Data	*Observations*															
	All 1126 audits	64	53	50	54	35	27	48 913	37 385	27 784	564	626	331	3.76	3.76	3.76
	Year-specific audits	61	68	18	23	14	0.23	56 535	35 467	4 486	613	874	0	3.76	1.47	0.10
Imputation (biased)	*Scaling*															
	Naive	90	88	89	9 078	9 127	9 647	48 913	37 385	27 784	564	626	331	3.76	3.76	3.76
	Stratification	82	79	80	4 442	4 422	4 573	76 486	764 86	76 486	18 628	18 628	18 628	0.86	0.86	0.86
	Propensity matching															
	Gaussian log-lik	82	88	89	4 574	8 437	9 426	25 141	46 310	48 372	0	0	0	3.76	3.76	3.76
	Semi-parametric	82	80	87	4 346	4 854	7 526	23 887	26 643	38 624	0	0	0	3.76	3.76	3.76
Computation by regression	*Linear regression*															
	Least squares	20	19	44	240	276	907	1 318	1 516	4 656	204	210	214	15.58	16.60	17.44
	Least absolute dev.	35	31	34	536	547	582	2 944	3 004	2 986	332	345	347	0.54	0.63	0.54
	Censored regression															
	Gaussian log-lik	14	19	40	153	283	763	839	1 554	3 914	64	52	68	11.64	9.17	36.85
	Semi-parametric	36	34	34	555	625	600	3 052	3 429	3 078	374	511	380	0.41	0.83	0.60
	Sample selection															
	Gaussian log-lik	14	15	37	157	206	688	861	1 133	3 532	112	96	89	11.82	17.87	11.96
	Semi-parametric	25	23	34	322	362	598	1770	2 009	3 063	703	810	862	2.07	1.38	1.74
	Censored Sample sel.															
	Gaussian log-lik	26	24	34	339	388	588	1 861	2 133	3 018	302	124	314	1.03	4.73	1.98
	Gaussian log-lik*	26	24	27	346	389	421	1 904	2 133	2 162	1	600	372	1.65	0.40	2.09

Table 7: Summary of CIT gap, individual firm-level deficiency (D), and the total deficiency (TD), predicted for tax periods 2014, 2015, and 2016. The CIT gap is % of potential CIT revenues (TD + declared tax). The results for the simple methods including the up-scaling, simple stratification (216 strata) according to the 12 NACE sectors, 8 + 1 regions, and 2 ownership types, and Propensity score matching based on nearest neighbor method, are severely biased. We report them only to illustrate the scale of the sample selection problem from a model-free perspective. Unlike all previous rows, the last row contains estimates from the Censored Heckman model estimated on year-specific audits as sensitivity checks for using the common set of audits.

6. Results

In this section, we present the summary statistics for population-wide estimates of the individual deficiency $\mathrm{D} = e^y$, total deficiency $\mathrm{TD} = \sum_{i=1}^{N} D_i$, and the CIT Gap (5.6) defined as $\mathrm{TD}/(\mathrm{TD} + \mathrm{TR})$ in the Table 7.

6.1. Alternative bottom-up estimates

Gaussian sample selection models became a standard econometric tool of applied economic research, especially for estimation of wage gaps, tax gaps, the impact of training programs, and others. Sophisticated semi- and non-parametric sample selection models, although more robust, are less frequently used, probably because their implementation in the leading econometric software packages is still limited.

Empirical studies rarely provide any comparison with alternative predictive approaches. However, we augment our main CIT Gap predictions from regression models with several more or less naive alternative methods presented in the first part of this section. We start with the most naive, model-free methods such as scaling. Given the lack representative firms in the audits, these methods will suffer from severe biases. We report them in Table 7 for two reasons:. First, to illustrate the scale of sample selection problem at hand from a model-free perspective which in this particular case highlights the benefits of more sophisticated methods. Second, because their transparency streamlines better understanding of the results presented in the second part of the section.

Scaling: naive and stratification. The most straightforward bottom-up CIT gap estimate can be computed by up-scaling of the mean deficiency observed among audited firms. This approach is naive as it neither takes into account the heterogeneity of firms recorded in their characteristics nor the fact that the audited sample is biased. Additionally, using the sample average estimator of the population mean of deficiency is sub-optimal, because the distribution of deficiencies is heavy-tailed. Consequently, the estimates of the CIT Gap become too high to be realistic. An alternative stratified scaling uses the same principle, but first, we stratify the population of firms (for each tax period) according to the Ownership, Region, and Size creating 90 strata. We compute the average deficiency within each stratum from the observed deficiencies. If no audited firm falls in a stratum, the stratum has 0 deficiency. In our case, this concerns 20% of all strata. Such simple stratification based on 3 factors reduces the CIT Gap estimates by 50%; however, the nominal values are still unrealistically high. Nevertheless, the strat-

ification illustrates one possible way how to reduce the sample selection bias in a model-free way.

Propensity score matching. A more sophisticated, model-based, but still relatively simple estimator can be obtained with statistical matching techniques. Each firm i in the audited set of $n = 1126$ firms is a potential donor of the value D_i, to a corresponding recipient in the population of N firms. The matching between donors and recipients is based on their similarity. The similarity is defined as the "propensity" of noncompliance, or in our case, the propensity of being selected for an audit, which is not necessarily the same. A binary selection model computes the propensities (e.g., Probit), and the matching algorithm can be, e.g., the nearest neighbor. This approach pushes the previously discussed stratification bias correction to the limit when each audited firm is a stratum itself[26]. However, with the data at hand, propensity score matching does not seem to work well. The Gaussian, as well as the robust (distribution-free) weighted binary selection models estimated from a balanced sample of 5 000 + 1 226 firms (as explained in the previous section), do not provide an appropriate basis for the nearest neighbor matching. The empirical densities of the predicted deficiencies copy the shape of the histogram of observed deficiencies, as shown in Figure A (in Appendix). No shift of the density location to the left, which would signalize some bias correction, is present. Therefore it is not surprising that the gap estimates are only slightly more realistic than those of the naive scaling approach.

Linear regression. The classical (weighted) OLS linear model is our first step towards the actual model-based prediction of the individual noncompliance. Compared to the propensity score matching, the individual deficiencies are computed, and not just imputed from the observed data. However, since unrestricted, they can become too big. Even one such excessive prediction can distort the estimates of TD and compromise the comparison of the CIT Gap levels over the years. To make the approach robust, one can use the least absolute deviation (LAD) instead of OLS. The LAD is less affected by outliers in the training sample by predicting the conditional median, not the mean. The LAD prediction of CIT Gap is more stable than OLS and close to 30% over the entire period. OLS predictions doubled

[26]The assumption is that the dependence between audit status (selected or not) and the level of noncompliance is only due to the characteristics X_S and disappears inside each of the groups. One needs to verify the "common support condition", i.e., whether the audited and non-audited firms have sufficient overlap in the predicted propensity score and whether these groups of firms have similar characteristics.

between 2015 and 2016. On the other hand, the robust predictions do not provide a good fit for the upper tail of deficiency distribution. The maximal LAD-predicted deficiencies are 4 times smaller than those observed in the audited sample. While OLS and LAD provide more realistic CIT gap values than previous methods, they differ a lot. This raises a question about their reliability.

Censored regression. As already discussed in Sections 1 and 3, there are good reasons to believe that the individual noncompliance D is not observed entirely. When $D < 4$ Eur[27] it may be disregarded and reported merely as 0. In this case, censored regression (also known as Tobit regression) is more appropriate than a linear model. A consistent estimator for Tobit is the Gaussian likelihood or a more robust semi-parametric estimator called censored LAD (CLAD). Intuitively, the CIT gap under such censoring will be higher than under a linear or sample selection model. Hence, censored regression gives a conservative upper-bound prediction of the CIT gap. In Table 7, we see that the CLAD predictions follow the same pattern as LAD, just 1 pp higher. Also, CLAD and LAD share the limitation concerning the upper tail of the deficiency distribution, where the discrepancy between predicted and observed D is enormous.

6.2. Main bottom-up estimates

Simple sample selection model. Table 7 shows results for both Gaussian and two-step semi-parametric selection regression models. According to both methods, the CIT gap has increased from 2014 to 2016 by more than 10 pp, while in 2015, there was either a tiny drop or increase in the gap. The total level of noncompliance during this period has doubled, also but not only because the population has grown in size.

To see how sample bias correction works, we compare the predicted distribution of D from the OLS and Gaussian sample selection model. Based on the mean, median, we see a clear shift of location to the left. A more complex picture can be drawn from the empirical densities shown in the Appendix. From the bottom-left plot, which is based on prediction for the entire population using all 1 126 audits, we see that the two densities look very similar, and the bias correction affected almost exclusively the location, not the shape.

Compared to all previous models, the differences in the predicted D, TD, and CIT Gap from Gaussian and semi-parametric models are smaller, which increases the reliability of the sample selection approach and validity of the gap estimates.

[27] The threshold 4 is the smallest non-zero deficiency observed among audited firms.

As expected, the main estimates are smaller than from OLS, LAD, or CLAD, but the shift in levels is not dramatic. By contrast, the previously mentioned lack of fit of the semi-parametric estimates in the upper tail of deficiency has disappeared.

Censored selection model. Our main workhorse is the combination of sample selection and censoring, as defined in Section 3. The main advantage of this model over the simple sample selection is the fit that it gives for the observed noncompliance. In particular, the CIT Gap among audited firms for the tax year 2015 was 53%, and the Censored Heckman model in-sample prediction is 52% (other in-sample results are available from the authors upon request). This is by far the best prediction provided by any model considered in this paper. For illustration, the second-best prediction was 30%, i.e., much less than the observed value.

The predictions computed for the entire population are almost identical to the semi-parametric two-step sample selection, only a bit higher. The descriptive statistics show a similar mean and upper tail, while the significant difference is in the median level of deficiency.

As a robustness check, we also estimated the Censored Heckman using only audits targeted on specific tax year (without merging). The results are shown in the last two rows of Table 7 look very similar, which is another sign of stability of the Censored Heckman approach towards data issues. According to this model, the CIT Gap has been close to one-quarter of the potential CIT revenues and increased to one third in 2016.

The density plots of predicted log-deficiencies exhibit a shift of the location towards zero when sample selection is taken into account. Censored Heckman, however, compensates for that shift by lifting the upper tail of the deficiency distribution. Still, the upper tail is flat enough to prevent excessive predictions from compromising the results.

6.3. The scale and the dynamics of the CIT Gap

Drivers of the CIT Gap growth in 2015-2016. In the face of the small decrease in the gap between 2014 and 2015, a 10pp increase predicted for 2016 looks surprising. Moreover, the top-down estimate and the observed deficiencies suggest the opposite (see Table 7). We attempt to explain this exciting result of our bottom-up approach by exploring the dynamics of firms in 2016. Running up the facts from the Table B, we could create an "Identikit"[28] of a noncompliant firm based on the observed relative counts. A typical suspect would be a domestic micro-firm

[28] In criminology, identikit refers to a set of characteristics used to identify a suspect.

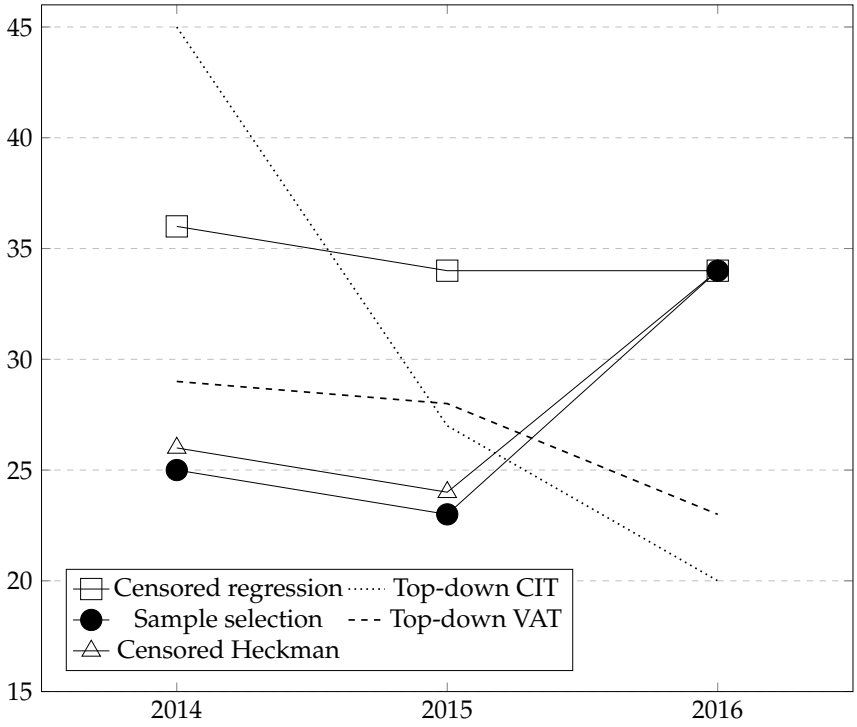

Figure 2: The predicted CIT gap = TD/(TD+TR) in %. For the bottom-up approaches, this includes all active firms except for financial, non-profit entities and the top largest firms (about 1% of all active firms). By contrast, the top-down approach includes the largest firms. Therefore the value of TR (total revenues) is twice as large as in the case of the bottom-up approach. The comparison of the two approaches in the plot does not take this fact into account because it requires non-trivial methodological adjustments.

registered in the capital with business in Wholesale and Retail. Interestingly, we would obtain the same set of characteristics when using the total predicted deficiency from 2015 instead of relative counts (see Table 8). The same Table also reveals the profile of the almost 20,000 (10 % of population[29]) new firms established just in 2016. These firms were from 99% domestic micro-firms, and about 30% of them were from Wholesale and Retail. To compare with the year before, only about 1,000 new firms appeared in 2015. Adjusting the population for the newly established corporations, the gap estimate in 2016 would reach EUR 364

[29]while 20,000 new firms were established, whole population increased by 10,000 on a year-on-year basis because other 10,000 companies ceased to exist

million and reversing the 10pp year-on-year gap increase. This CIT dynamics is in line with our broader results about the increased risk of noncompliance among the SMEs.

Top-down vs. bottom-up. The largest multinational corporations (1% of the entire population) by tax liability pay more than half of all tax revenues. We hypothesize that due to the mandatory audit of financial statements, a higher propensity to be audited by Financial administration, and due to higher reputation risk, MNCs do not avoid paying the taxes by under-reporting their income or overstating their costs but rather in a form of the profit shifting given their international exposure within corporate groups which is not a subject of this paper. Therefore, our bottom-up estimate should be considered as a subset of total gap delivered by top-down estimate covering a tax avoidance in a form of overstated costs or under-reported income likely among SMEs. Focusing the bottom-up estimation on the 99% of the active population, i.e., excluding the MNCs, reduces the value of TR (total revenues) in (5.6), which, given the small propensity of the MNCs to the audited types of tax evasion, leads to higher gap estimates in general.

6.4. Distribution of the CIT Gap

Despite some problems in the underlying data, we deem the bottom-up estimate not only an alternative for the top-down estimate but also a useful tool to analyze the gap distribution. Below, we give a summary for the year 2015 with details available in the Table 8.

(i) The micro-firms account for 98% of the CIT Gap in the entire population, but create only 19% of tax revenues from audits. Among them, firms with up to 1 employee have the highest typical noncompliance in the entire population.

(ii) The capital Bratislava is the region with the highest total noncompliance (25%). However, local firms have typically 5% lower level of evasion than the rest of the population. The highest typical evasion (as well as evasion per firm) is in Nitra (50% more than the respective population statistic).

(iii) Sectoral distribution of the gap is concentrated in Wholesale and Retail (33%), followed by Specialized (14%) and Construction (10%) sectors. The highest typical evasion (as well as evasion per firm) is in Agriculture.

(iv) While foreign firms typically evade two times more, domestic firms are responsible for 76% of the total evasion.

These results are based on our most reliable Censored Heckman model. The robust semi-parametric two-step selection model would lead to very similar conclusions, with one exception, which is the Construction sector. This is not surprising given that it was the second-largest among audited firms.

An additional cross-check can be done using the alternative stratified non-parametric estimator. As already mentioned, this estimate is biased due to the sample selection, while, on the other hand, it does not suffer from estimation uncertainty as much as the parametric approaches. The stratification suggests that micro-firms do not hold a 98% but rather "only" 72% share on the total noncompliance. However, the rest is in line with the conclusions presented above.

7. Conclusion

We provide the first set of *size-, sector-, and region*-specific corporate tax gap estimates for Slovakia as well as explicit, data-driven support for policies leading to more substantial audit capacity and improving the auditing process of all active firms (excluding financial, non-profit and the MNCs). Based on our analysis, the risk management and the selection process of audit cases should be centralized to ensure representativeness. The audit activity should be focused primarily on active companies to achieve cogent allocation of resources. In contrast, a still relatively high number of audits is focused on inactive firms. Tax audit results should describe both the most relevant characteristics and results of audits in standardised reports. In later stages of this project, a more in-depth analysis of audits should provide potential suggestions for the Ministry of Finance as a responsible body for an update of legislation and prevent potential tax avoidance[30].

In particular, we identify sectors and geographical regions with the highest total and average predicted deficiencies. These results, while first of their art in Slovakia and most EU countries, raise further discussion on how best to identify firms with the potentially highest level of noncompliance.

Our CIT gap estimates, while complementary to the currently used top-down results, require further refinement. In optimal circumstances, both approaches should provide comparable estimates, proportionate in volume, and following a similar trend. However, given the currently limited scope of the bottom-up approach, our CIT gap estimates cover only a tax avoidance of SMEs - a subset of total CIT gap delivered by top-down approach. Therefore, any direct comparison with top-down estimates is likely to be deficient for now. Extending the bottom-up estimates with the large multinational corporations should decrease the currently observed discrepancy from the top-down. However, both estimates have

[30] As an example can be considered research in the field of international tax avoidance. This research was a cornerstone for the introduction of various measures like thin-capitalization rules, regimes to avoid hybrid mismatches or controlled foreign company rules.

	Summary of D by parametric approach				Total D in Mill. €predicted by		
	Min €	Mean €	Median €	Max Mill. €	Para-metric	Semipara-metric	Nonpara-metric
Size							
Micro	1	1 390	69	5	380	344	3 784
	-	65 %	56 %	100 %	98 %	95 %	72 %
thereof ≤ 1	1	4 031	719	4	245	193	1 597
employee	-	188 %	585 %	93 %	63 %	53 %	30 %
Small	0	269	17	1	5	16	858
	-	13 %	14 %	23 %	1 %	4 %	16 %
Medium	0	76	6	0	0	2	584
	-	4 %	5 %	1 %	0 %	0 %	11 %
Large	0	6 219	3	2	3	0	49
	-	291 %	2 %	52 %	1 %	0 %	1 %
Region							
Bratislava	0	1 642	117	4	98	69	1 916
	-	77 %	95 %	94 %	25 %	19 %	22 %
Trnava	2	2 930	162	2	46	14	1 631
	-	137 %	132 %	49 %	12 %	4 %	19 %
Trenčín	1	1 986	102	2	28	31	337
	-	93 %	83 %	34 %	7 %	9 %	4 %
Nitra	1	3 258	186	2	65	61	2 118
	-	152 %	151 %	51 %	17 %	17 %	24 %
Žilina	1	2 212	126	1	41	60	1 386
	-	103 %	102 %	19 %	11 %	17 %	16 %
B. Bystrica	0	2 953	181	2	48	56	536
	-	138 %	147 %	40 %	12 %	15 %	6 %
Prešov	0	1 438	77	1	24	63	278
	-	67 %	63 %	17 %	6 %	17 %	3 %
Košice	0	1 970	93	5	39	8	528
	-	92 %	76 %	100 %	10 %	2 %	6 %

Table 8: Nominal and relative summary statistics of deficiencies (D) per category. Predictions for all firms (after cleaning) in the population for the tax year 2015. The summary statistics in the left panel are for the Censored Heckman model (param.). In the right panel, the totals are obtained by the semi-parametric sample selection model (semi-param.) or non-parametric up-scaling based on stratification. For instance, for the *Region*, the stratification is based on 8 strata.

	Summary of D by parametric approach				Total D in Mill. €predicted by		
	Min €	Mean €	Median €	Max Mill. €	Para-metric	Semipara-metric	Nonpara-metric
Sector							
Accommmo-dation	0	1 105	74	0	8	7	77
	-	52 %	60 %	2 %	2 %	2 %	1 %
Agriculture	3	6 198	519	2	28	48	250
	-	290 %	422 %	40 %	7 %	13 %	3 %
Construction	0	1 999	130	2	38	73	801
	-	93 %	106 %	51 %	10 %	20 %	9 %
Finance	1	2 196	240	1	4	3	28
	-	103 %	195 %	12 %	1 %	1 %	0 %
Industry	0	1 543	43	2	24	18	564
	-	72 %	35 %	35 %	6 %	5 %	6 %
Information	0	2 065	171	1	23	25	1 028
	-	97 %	139 %	24 %	6 %	7 %	12 %
Others	0	1 370	48	2	21	19	127
	-	64 %	39 %	52 %	5 %	5 %	1 %
Real Estate	0	988	90	0	13	7	485
	-	46 %	73 %	5 %	3 %	2 %	6 %
Specialized	0	1 907	147	4	56	47	1 287
	-	89 %	120 %	94 %	14 %	13 %	15 %
Supporting	1	3 008	281	1	34	56	846
	-	141 %	228 %	29 %	9 %	15 %	10 %
Transport	0	1 389	60	0	10	15	131
	-	65 %	49 %	9 %	3 %	4 %	1 %
Wholesale& Retail	0	2 851	194	5	130	44	3 170
	-	133 %	158 %	100 %	33 %	12 %	36 %
Ownership							
Domestic	0	1 915	117	4	295	315	6 980
	-	90 %	95 %	93 %	76 %	87 %	78 %
Foreign	0	3 496	211	5	94	46	1 987
	-	163 %	172 %	100 %	24 %	13 %	22 %

Table 8: (Continued 2/2) Summary statistics of deficiencies predicted for the tax period 2015 for the entire population of firms using the proposed modification of Heckman's approach.

their limits and should not be judged by their (lack of) ability to coincide.

References

Ahn, H. and J. L. Powell (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics 58*(1), 3 – 29.

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Bårdsen, G. and H. Lütkepohl (2011). Forecasting levels of log variables in vector autoregressions. *International Journal of Forecasting 27*(4), 1108 – 1115.

Brys, B. (2011). Making fundamental tax reform happen. *OECD Economics Department Working Papers* (3).

Chen, S. and Y. Zhou (2010). Semiparametric and nonparametric estimation of sample selection models under symmetry. *Journal of Econometrics 157*(1), 143 – 150. Nonlinear and Nonparametric Methods in Econometrics.

Chen, S., Y. Zhou, and Y. Ji (2018). Nonparametric identification and estimation of sample selection models under symmetry. *Journal of Econometrics 202*(2), 148 – 160.

Cosslett, S. (1987). Semiparametric estimation of a regression model with sample selectivity. In W. A. Barnett, J. Powell, and G. Tauchen (Eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, pp. 175–197. Cambridge University Press.

Craven, P. and G. Wahba (1978, Dec). Smoothing noisy data with spline functions. *Numerische Mathematik 31*(4), 377–403.

Das, M., W. K. Newey, and F. Vella (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies 70*(1), 33–58.

EC (2018). The concept of tax gaps. corporate income tax gap estimation methodologies. *Working paper No.73-2018*. Available at https://ec.europa.eu/taxation_customs/sites/taxation/files/taxation_papers_73_en.pdf.

Erard, B. (1997). A critical review of the empirical research on canadian tax compliance. *Technical Committee on Business Taxation Working Paper* (97-6).

Erard, B. and J. S. Feinstein (2007, December). Econometric Models for Multi-Stage Audit Processes: An Application to the IRS National Research Program. International Center for Public Policy Working Paper Series, at AYSPS, GSU paper0723, International Center for Public Policy, Andrew Young School of Policy Studies, Georgia State University.

Erard, B. and C.-C. Ho (2001). Searching for ghosts: who are the nonfilers and how much tax do they owe? *Journal of Public Economics 81*(1), 25 – 50.

Escanciano, J. C. and L. Zhu (2015). A simple data-driven estimator for the semiparametric sample selection model. *Econometric Reviews 34*(6-10), 734–762.

Feinstein, J. S. (2001, 12). Approaches for Estimating Noncompliance: Examples from Federal Taxation in the United States. *The Economic Journal 109*(456), 360–369.

Feinstein, J. S. and B. Erard (2010). *Econometric Models for Multi-Stage Audit Processes: An Application to the IRS National Research Program*, pp. 113–137. Routledge.

Hanlon, M., L. Mills, and J. Slemrod (2007). An empirical examination of corporate tax noncompliance. In A. J. Auerbach, J. R. Hines, Jr., and J. Slemrod (Eds.), *Taxing Corporate Income in the 21st Century*, pp. 171–210. Cambridge University Press.

Heckman, J. (1974). Shadow prices, market wages, and labor supply. *Econometrica 42*(4), 679–694.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–161.

Henningsen, A. and O. Toomet (2011). maxlik: A package for maximum likelihood estimation in R. *Computational Statistics 26*(3), 443–458.

Honoré, B. E. and L. Hu (2018). Selection without exclusion. *Working paper*. Available at https://scholar.princeton.edu/honore/publications.

Huber, M. and B. Melly (2015). A test of the conditional independence assumption in sample selection models. *Journal of Applied Econometrics 30*(7), 1144–1168.

Johansson, A., C. Heady, J. Arnold, B. Brys, and L. Vartia (2008). Taxation and economic growth. *OECD Economics Department Working Papers* (620).

Klein, R. W. and R. H. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica 61*(2), 387–421.

Kleven, H. J., M. B. Knudsen, C. T. Kreiner, S. Pedersen, and E. Saez (2011). Unwilling or unable to cheat? evidence from a tax audit experiment in denmark. *Econometrica 79*(3), 651–692.

Lancaster, T. and G. Imbens (1996). Case-control studies with contaminated controls. *Journal of Econometrics 71*(1), 145 – 160.

Li, Q. and T. Stengos (1992). A hausman specification test based on root-n-consistent semiparametric estimators. *Economics Letters 40*(2), 141 – 146.

Manski, C. F. and S. R. Lerman (1977). The estimation of choice probabilities from choice based samples. *Econometrica 45*(8), 1977–1988.

Mora, R. (2008). A nonparametric decomposition of the mexican american average wage gap. *Journal of Applied Econometrics 23*(4), 463–485.

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica 55*(4), 765–799.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal 12*(s1), S217–S229.

Newey, W. K., J. L. Powell, and J. R. Walker (1990). Semiparametric estimation of selection models: Some empirical results. *The American Economic Review 80*(2), 324–328.

Nicolay, K. (2013). Tax avoidance of german multinationals and implications for tax revenue. *working paper version: 9/2013*. Available at https://www.zew.de/en/team/kfi.

OECD (2017). The measurement of tax gaps. In *Tax Administration 2017: Comparative Information on OECD and Other Advanced and Emerging Economies*, pp. 181 – 188. Paris: OECD Publishing. Available at https://www.oecd-ilibrary.org/content/component/tax_admin-2017-19-en.

Paarsch, H. J. (1984). A monte carlo comparison of estimators for censored regression models. *Journal of Econometrics 24*(1), 197 – 213.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics 25*(3), 303 – 325.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Schafgans, M. M. A. (1998). Ethnic wage differences in malaysia: parametric and semiparametric estimation of the chinesemalay wage gap. *Journal of Applied Econometrics 13*(5), 481–504.

Tibshirani, R. and K. Knight (1999). The covariance inflation criterion for adaptive model selection. *J. R. Stat. Soc. Ser. B Stat. Methodol. 61*, 529–546.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica 26*(1), 24–36.

Toomet, O. and A. Henningsen (2008). Sample selection models in R: Package sampleSelection. *Journal of Statistical Software 27*(7).

Ueda, J. (2018). Estimating the corporate income tax gap : The ra-gap methodology. *IMF technical notes and manuals*. Available at https://www.imf.org/en/Publications/TNM/Issues/2018/09/12/.

Warner, G., S. Wijesinghe, U. Marques, O. Badar, J. Rosen, E. Hemberg, and U.-M. O'Reilly (2015). Modeling tax evasion with genetic algorithms. *Economics of Governance 16*(2), 165 – 178.

Appendix:

Top-down vs. bottom-up tax estimation

	Top-down	**Bottom-up** Random audits	**Bottom-up** Operational audits
Data required	1. Aggregated national accounts macroeconomic data for computation of potential tax base compiled independently of declared tax base and liability. 2. Understanding how national accounts data are constructed to include unobserved economic activities.	1. Firm's individual legal, sectoral, geographical and operational profile, financial profile including the sources of debt/funding and financial scoring, social profile including the number, type, and wages of the employees, VAT, CIT payed. 2. Records on all CIT/VAT audits, their scope and outcome.	1. and 2. Same as for random audits. 3. All potentially relevant characteristics for selection of firms for audit.
Used in	Most EU countries for VAT. Italy and Slovakia for CIT	Denmark, UK, Sweden, USA	Australia, Italy, UK, USA, Slovakia
Expected outcomes - fiscal impact of noncompliance	✓	✓	✓
- possible focus on specific evasion, e.g., transfer pricing		✓	✓
- identification of the source of the gap		✓	✓
- measure of impact of distinct types of noncompliance		✓	✓
- insights into taxpayers behaviors and risks		✓	✓
- insights into the structure (sector, region, ownership) of the gap		✓	✓
- implications for financial authority to improved audit targeting			✓

Table A: A comparison of the top-down and bottom-up approach for tax gap estimation concerning the data required and the benefits provided.

Data cleaning steps from Table 1 with additional explanations:

(i) Omit entries which cannot be found in FASR registries[31].
(ii) Omit entries that contain any missing values.
(iii) Omit non-profit organizations, financial institutions, insurance companies, EU-, central-, and local - government organizations for the sake of comparison with the top-down approach. Also, omit " large multinational corporations"(1) since they are not represented among the audited firms.

Our next steps include careful cleaning of the set of audits. For the sake of space, we use the established audit terminology, which is explained below.

(iv) Keep only audits finalized for tax period 2014-2016.
(v) Omit the "by-tools"(2) audits.
(vi) Omit the audits targeted as the "minimum income tax"(3).
(vii) Screen for duplicitous audits (e.g., audits targeted at the same firm over the same or subsequent years) and keep the most relevant entries only.
(viii) Omit the audits with negative deficiency.
(ix) Omit audits with reduced tax loss(4).
(x) Augment the audit data with the selected characteristics and omit missing values.

Notes

1. Slovak tax legislation defines the "large multinational corporations" as firms with total revenues above 40 000 000 Eur. These firms are subjected to audits under a special regime. Many of these firms are banks, insurance, and reinsurance companies and have different accounting standards (IFRS) than ordinary firms. Their size and possible relationship with the large multinational corporations give them some specific opportunities for noncompliance (e.g., transfer pricing). Given that the selection and audit procedures are different from the rest of the population, the CIT gap requires a different approach from the rest of the population.

2. Audits labeled as "by-tools" do not provide any useful information for the CIT-gap because the deficiency which they detect is simply the firm's revenues multiplied by the corresponding tax rate. Tax audit "by-tools", in general, means that the company does not cooperate with the tax auditor, does not provide bookkeeping for the auditor, or is not active anymore. In this case, tax auditors use available sources of information to assess tax liability. Since there is not any proof that costs stated in tax returns or profit and loss statements are valid, they cannot be taken into account for audit. The other example of "by-tools" audits an assessment of tax license for companies that

[31] Missing unique tax identifier either does not exist (which is indeed possible for some micro-firms), or this firm is not active but for any reason still appears in any of our other data sources, or the firm is active but does do not appear in FASR registry.

submitted a negative tax return. The proportion of such audits before cleaning is about 50%. After merging and augmentation with non-tax characteristics, we see that the proportion of "by-tools" audits decreases to only 10%.

3. The minimum income tax, so called "tax licence" was following 46b of Act no. 595/2003 Coll. on income tax as amended, the minimum taxable amount of the corporate income tax, after deduction of tax credits and after the offsetting of tax paid abroad, paid by the TA taxpayer for each taxable period for which he has declared: a tax liability lower than the tax licence, or zero tax liability, or tax loss. Firms were obliged to pay the minimum income tax over the tax-period 2014-2017 (including). It aimed to increase tax compliance of companies which do not report any profit for several years and make the in-active firms to de-register, thus helping to clean up the out-dated registries of FASR. The tax license was a fixed amount of 480, 960, or 2880 Eur, depending on specific conditions. Hence the audits targeted at the tax licenses are easy to identify from the detected deficiency. The chances that an audit targeted at the tax license would result in no detected deficiency are minimal. The nature of the noncompliance concerning tax license is, in most cases, non-deliberate, given that it is easily detectable. For this reason, these audits do not fit into our framework, which focuses on the deliberate and sophisticated tax-avoidance.

4. Reduction of tax loss means positive results of tax audits but is not automatically connected with assessed tax liability. The amount of tax loss is important for tax liability in the next four years. Following the Section 30 (1) of the Income Tax Act, it is possible to deduct the tax loss from the tax base of a legal person evenly during the four consecutive tax periods starting from the tax period immediately following the taxable period for which the tax loss has been recognized. Including the tax loss reduction in the estimated tax gap is non-trivial as it has to account for the delay with which the tax loss is effectively deducted from the tax base. If the audit result is a reduction of the firm's tax loss, there may have three reasons: 1. A firm's tax loss $X > 0$ was reduced to the loss of Y, where $X > Y > 0$. If the firm reaches a positive profit in the subsequent 4 years, it will not be able to claim the annual tax deduction by 1 / 4 of X but by 1 / 4 of Y. However, we can not predict whether or not it will be profitable, or it will still be in loss. So we do not know how much we have increased the CIT return. 2. A firm's tax loss $X > 0$ was reduced to 0. If it reaches profit in the following years, it will not be able to apply the annual tax deduction. We would have to know the level of tax bases in the years to come, in order to know how much we have earned by this tax reduction. 3. A firm's tax loss $X > 0$ was reduced to 0, and a profit of $Y > 0$ was detected. In this case, we know at least the momentum of the earned tax. We omit these audits to keep our sample homogeneous and to model them separately using different tools.

Identified problems and some recommendations for Financial Administration

We present 4 major problems that we identified while analyzing audits conducted during 2015-2018. These issues complicate not only the estimation of tax gaps but also harm effective compliance risks management. For each problem, we provide an adequate solution based on the data provided by the FASR, previous findings of TADAT from 2018, and the discussions with the staff of the FASR:

(i) *Uncoordinated targeting of tax audits:*
The selection of entities for tax audit is uncoordinated among the organizational units of the FASR. Consequently, the audits do not cover different

types of taxpayers, and selection criteria are not transparent. This, in turn, precludes a thorough evaluation of these selection criteria.

(ii) *Insufficient use of information obtained during tax audit:*
The output of the tax audit does not provide detailed information about errors found during the tax audit. The prescribed structure of the tax audit protocol is not sufficiently rich for subsequent analytical processing.

(iii) *Ineffective allocation of resources:*
50% of tax controls proceed "by-tools", i.e., deal with inactive companies[32], whose undeclared taxes are unenforceable. This leads to a highly inefficient allocation of resources since the findings of such audits cost much effort, but bring nothing in return.

(iv) *Outdated register of taxpayers administrated by FASR:*
The register contains a large number of long-term inactive entities, who still have a valid registration on CIT. The use of the registry is burdensome for analytical purposes as there is no indicator, which subjects are active.

Therefore, we propose the following steps to be taken:

(i) Centralize the selection of firms for audits and keep the historical record of the criteria which lead to an audit.

(ii) Introduce a standardized, analyst-friendly protocol, including the reasons for selecting the firm through precisely defined indicators and, after completion, the details of the findings.

(iii) Increase the number of income tax audits focused exclusively on economically active entities; minimize the number of "by-tools" audits. This will require a change of the key performance indicators of local tax offices. The indicators should monitor and evaluate the amount of tax paid after the audit rather than the possible amount.

(iv) Clean up the taxpayer's register of active entities in cooperation with the Ministry of Justice. This requires some legislative changes.

[32] Companies which became inactive before or at the moment when an audit is triggered.

	Population			Audits			
	2014	2015	2016	2014	2015	2016	Total
				All			
before cleaning	308 363	309 111	307 768	1 432	1 527	319	3 278
				44%	47%	10%	100%
after cleaning	184 964	185 962	196 572	416	568	142	1 126
				37%	50%	13%	100%
thereof 0 tax adjustment				*170*	*252*	*62*	*484*
				15%	*22%*	*6%*	*43%*
				Size			
Micro	162 294	162 002	172 137	193	253	59	505
	88%	87%	88%	46%	45%	42%	45%
thereof ≤ 1 *employee*	*67 655*	*63 014*	*64 857*	*46*	*53*	*16*	*115*
	37%	*34%*	*33%*	*11%*	*9%*	*11%*	*10%*
Small	18 643	19 763	20 309	160	243	50	453
	10%	11%	10%	38%	43%	35%	40%
Medium	3 635	3 793	3 725	55	64	32	151
	2%	2%	2%	13%	11%	23%	13%
Large	392	404	401	8	8	1	17
	0%	0%	0%	2%	1%	1%	2%
				Region			
Bratislava	61 310	62 153	65 869	78	143	29	250
	33%	33%	34%	19%	25%	20%	22%
Trnava	16 105	16 227	17 292	44	67	17	128
	9%	9%	9%	11%	12%	12%	11%
Trenčín	14 762	14 686	15 565	53	56	23	132
	8%	8%	8%	13%	10%	16%	12%
Nitra	20 227	20 208	21 102	37	59	11	107
	11%	11%	11%	9%	10%	8%	10%
Žilina	19 101	19 136	20 285	56	88	17	161
	10%	10%	10%	13%	15%	12%	14%
B. Bystrica	16 591	16 581	17 587	66	61	11	138
	9%	9%	9%	16%	11%	8%	12%
Prešov	16 874	16 945	17 886	49	56	17	122
	9%	9%	9%	12%	10%	12%	11%
Košice	19 993	20 025	20 986	33	38	17	88
	11%	11%	11%	8%	7%	12%	8%

Table B: Nominal and relative counts of firms in each category of selected characteristics. Left panel: the entire (after cleaning) population of firms active in Slovakia in the respective year. Right panel: the set of all firms (after cleaning) audited for the respective tax year.

	Population			Audits			
	2014	2015	2016	2014	2015	2016	Total
				All			
before cleaning	308 363	309 111	307 768	1 432	1 527	319	3 278
				44%	47%	10%	100%
after cleaning	184 964	185 962	196 572	416	568	142	1 126
				37%	50%	13%	100%
thereof 0 tax adjustment				*170*	*252*	*62*	*484*
				15%	*22%*	*6%*	*43%*
				Sector			
Accommodation	7 077	7 125	7 619	16	26	3	45
	4%	4%	4%	4%	5%	2%	4%
Agriculture	4 518	4 604	4 837	8	5	3	16
	2%	2%	2%	2%	1%	2%	1%
Construction	19 442	19 541	20 871	64	120	53	237
	11%	11%	11%	15%	21%	37%	21%
Finance	1 971	2 070	2 388	-	1	-	1
	1%	1%	1%	0%	0%	0%	0%
Industry	15 984	15 998	16 754	62	58	16	136
	9%	9%	9%	15%	10%	11%	12%
Information	10 807	11 226	12 027	11	16	5	32
	6%	6%	6%	3%	3%	4%	3%
Others	14 942	15 363	16 356	14	11	3	28
	8%	8%	8%	3%	2%	2%	2%
Real Estate	13 396	13 429	14 106	17	21	7	45
	7%	7%	7%	4%	4%	5%	4%
Specialized	29 617	30 112	32 055	30	42	6	78
	16%	16%	16%	7%	7%	4%	7%
Supporting	11 578	11 716	12 820	24	26	6	56
	6%	6%	7%	6%	5%	4%	5%
Transport	7 695	7 730	7 848	29	26	9	64
	4%	4%	4%	7%	5%	6%	6%
Wholesale& Retail	47 937	47 048	48 891	141	216	31	388
	26%	25%	25%	34%	38%	22%	34%
				Ownership			
Domestic	156 746	158 420	168 636	337	468	120	925
	85%	85%	86%	81%	82%	85%	82%
Foreign	28 218	27 542	27 936	79	100	22	201
	15%	15%	14%	19%	18%	15%	18%

Table B: (Continued 2/2) Nominal and relative counts of firms in each category of selected characteristics.

	Population			Audits			
	2014	2015	2016	2014	2015	2016	Total
Firms (Ths.)	181.93	182.18	194.86	0.42	0.57	0.14	1.13
	90%	89%	92%	29%	37%	45%	34%
	Revenues						
missing	0	0	0	0	0	0	0
	-	-	-	-	-	-	-
total (Mld. €)	80.40	88.03	86.52	1.16	2.06	0.50	3.72
	47%	46%	46%	43%	73%	85%	61%
mean (Mill.. €)	0.44	0.48	0.44	2.79	3.62	3.53	3.30
	48%	48%	46%	69%	99%	104%	87%
med (Ths. €)	34.68	39.78	38.37	1 080.47	1 255.11	1 206.13	1 179.74
	102%	106%	99%	143%	116%	108%	123%
sd (Mill.. €)	2.16	2.43	2.57	4.98	8.10	6.45	6.90
	8%	9%	9%	28%	73%	91%	50%
min (Mill.. €)	-0.65	-2.58	-1.63	0.00	0.00	0.00	0.00
	5%	100%	100%	0%	-	-	0%
max (Mill.. €)	192.09	240.86	581.37	33.99	114.07	51.20	114.07
	3%	3%	8%	11%	59%	94%	36%
	Profit before tax						
missing	0	0	0	0	0	0	0
	-	-	-	-	-	-	-
total (Mld. €)	3.51	3.55	3.52	0.06	0.03	0.02	0.10
	38%	34%	36%	332%	82%	127%	152%
mean (Mill.. €)	0.02	0.02	0.02	0.13	0.06	0.11	0.09
	39%	36%	37%	537%	111%	156%	218%
med (Ths. €)	1.06	1.83	1.82	10.59	12.19	14.01	11.78
	102%	106%	99%	252%	140%	139%	188%
sd (Mill.. €)	2.05	0.69	0.43	1.03	4.32	0.39	3.13
	67%	25%	18%	84%	115%	106%	116%
min (Mill.. €)	-93.92	-79.50	-87.04	-2.91	-79.50	-0.44	-79.50
	73%	100%	61%	15%	100%	56%	100%
max (Mill.. €)	813.61	155.23	46.98	17.18	63.85	4.08	63.85
	100%	23%	9%	100%	100%	100%	100%

Table C: Summary statistics of selected individual financial characteristics. Left panel: the entire population of firms active in Slovakia after the cleaning step (iii) of Table 1. Right panel: the set of all firms (after cleaning step (x)) audited for the tax years 2014-2016. The percentages below each row in the table represent the proportion of the respective quantities obtained from the raw data on active firms (i.e., after the cleaning step (i)).

	Population			Audits			
	2014	2015	2016	2014	2015	2016	Total
Firms (Ths.)	181.93	182.18	194.86	0.42	0.57	0.14	1.13
	90%	89%	92%	29%	37%	45%	34%
	Corporate tax base declared						
missing	0	0	0	2	167	75	244
	-	-	-	0%	18%	32%	14%
total (Mld. €)	4.11	5.25	5.03	0.07	0.03	0.01	0.10
	42%	44%	44%	67%	52%	99%	63%
mean (Mill. €)	0.02	0.03	0.03	0.16	0.08	0.08	0.12
	47%	49%	48%	137%	79%	122%	109%
median (Ths. €)	1.02	2.09	2.01	12.15	15.13	14.96	13.52
	166%	123%	116%	582%	162%	133%	303%
sd (Mill. €)	0.20	0.23	0.21	1.00	0.26	0.20	0.71
	10%	11%	11%	121%	37%	110%	94%
min (Mill. €)	-0.01	-0.03	-0.06	0.00	0.00	0.00	0.00
	100%	51%	100%	-	-	-	-
max (Mill. €)	39.25	40.47	42.46	16.84	3.62	1.49	16.84
	8%	7%	8%	100%	29%	100%	100%
	Corporate tax declared						
missing	0	0	0	2	167	75	244
	-	-	-	0%	18%	32%	14%
total (Mld. €)	0.98	1.20	1.15	0.02	0.01	0.00	0.02
	45%	46%	46%	67%	53%	98%	63%
mean (Mill. €)	0.01	0.01	0.01	0.04	0.02	0.02	0.03
	50%	51%	50%	137%	79%	121%	109%
median (Ths. €)	0.96	0.96	0.96	2.88	3.09	3.03	2.90
	100%	100%	100%	123%	107%	105%	101%
sd (Mill. €)	0.04	0.05	0.05	0.22	0.06	0.04	0.16
	10%	11%	11%	121%	38%	110%	95%
min (Mill. €)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	-	-	-	-	-	-	-
max (Mill. €)	8.64	8.90	9.34	3.70	0.80	0.33	3.70
	8%	7%	8%	100%	29%	100%	100%

Table C: (Continued 2/3) Summary statistics of selected individual financial characteristics.

	Population			Audits			
	2014	2015	2016	2014	2015	2016	Total
Firms (Ths.)	181.93	182.18	194.86	0.42	0.57	0.14	1.13
	90%	89%	92%	29%	37%	45%	34%
	Value added						
missing	0	0	0	0	0	0	0
	-	-	-	-	-	-	-
total (Mld. €)	16.36	17.82	18.65	0.19	0.15	0.08	0.42
	49%	48%	49%	63%	69%	80%	68%
mean (Mill. €)	0.09	0.10	0.10	0.47	0.26	0.56	0.38
	51%	50%	49%	102%	93%	99%	97%
median (Ths. €)	5.11	7.10	7.12	83.88	71.73	155.48	79.57
	103%	109%	99%	201%	168%	157%	170%
sd (Mill. €)	0.59	0.61	0.61	1.46	3.39	0.92	2.59
	13%	12%	12%	75%	103%	56%	98%
min (Mill. €)	-74.85	-74.08	-6.99	-2.45	-74.08	-0.44	-74.08
	100%	84%	5%	39%	100%	51%	100%
max (Mill. €)	43.58	33.57	77.10	17.06	15.78	5.05	17.06
	6%	3%	9%	69%	49%	26%	53%
	Net assets						
missing	0	0	0	0	0	0	0
	-	-	-	-	-	-	-
total (Mld. €)	109.80	108.92	108.31	0.79	1.16	0.34	2.28
	52%	50%	49%	46%	76%	80%	62%
mean (Mill. €)	0.60	0.60	0.56	1.89	2.04	2.36	2.03
	54%	53%	50%	75%	103%	98%	90%
median (Ths. €)	35.83	40.96	39.88	532.51	554.14	780.26	561.85
	102%	106%	99%	116%	118%	105%	116%
sd (Mill. €)	19.75	5.99	4.79	4.01	7.70	4.01	6.15
	49%	16%	12%	40%	101%	80%	73%
min (Mill. €)	-1.94	-4.61	-7.21	-0.07	0.00	0.00	-0.07
	100%	100%	100%	66%	0%	-	1%
max (Mill. €)	8073.96	1550.19	949.96	28.03	139.18	22.41	139.18
	89%	17%	10%	17%	100%	48%	84%

Table C: (Continued 3/3) Summary statistics of selected individual financial characteristics.

	Type	OLS regression	Gaussian sample selection		Censored sample selection	
		–	Selection	Outcome	Selection	Outcome
Constant		9.50	−9.651***	9.327***	−6.551***	9.692***
		(0.00)	(0.248)	(0.480)	(0.210)	(0.851)
No. of employees	Num	0.025	4.622***	0.107	3.761***	0.105
		(0.080)	(0.223)	(0.070)	(0.227)	(0.144)
Region	Cat					
Bratislava		−1.33***	−0.516***	−1.463***	−0.248**	−0.899***
		(0.15)	(0.086)	(0.170)	(0.077)	(0.354)
Trnava		−2.59***		−2.568***		0.028
		(0.33)		(0.202)		(0.423)
Trencin		−0.243	0.354***	−0.177	0.320***	−0.159
		(0.25)	(0.098)	(0.200)	(0.090)	(0.411)
Nitra		–	−0.087	–	0.116	–
		–	(0.100)	–	(0.090)	–
Zilina		0.65***	0.344***	0.489***	0.326***	−0.120
		(0.13)	(0.095)	(0.188)	(0.087)	(0.385)
B. Bystrica		–	0.282**	–	0.228*	–
		–	(0.097)	–	(0.089)	–
Presov		0.12	0.084	0.200	0.214*	−0.682
		(0.13)	(0.100)	(0.205)	(0.091)	(0.422)
Kosice		−2.70***	−0.214*	−2.684***	−0.192*	−0.655
		(0.23)	(0.105)	(0.229)	(0.094)	(0.489)
Sector	Cat					
Industry		−1.71***	0.151	−1.836***	0.073	−0.214
		(0.237)	(0.142)	(0.382)	(0.115)	(0.782)
Accomodation		−1.78***	0.973***	−1.799***	0.543***	−0.338
		(0.352)	(0.162)	(0.440)	(0.135)	(0.902)
Wholesale & retail		−2.26***	0.698**	−2.337***	0.320**	0.029
		−0.25	(0.130)	(0.359)	(0.102)	(0.730)
RealEstate		−3.93***	0.267.	−3.943***	−0.019	0.929
		(0.271)	(0.159)	(0.444)	(0.132)	(0.444)
Agriculture		0.621**	−0.344	0.583	0.006	0.704
		(0.247)	(0.204)	(0.574)	(0.172)	(1.172)
Construction		−0.606**	1.048***	−0.703*	0.643***	0.749
		(0.239)	(0.135)	(0.369)	(0.108)	(0.369)
Information		−2.317***	0.120	−2.345***	0.196	0.982
		(0.374)	(0.178)	(0.477)	(0.138)	(0.477)
Specialized		−1.425***	0.079	−1.105***	0.197.	0.036
		(0.261)	(0.149)	(0.404)	(0.114)	(0.829)
Finance		–	−0.652	–	−0.014	–
		–	(0.550)	–	(0.346)	–
Transport		−0.991***	0.587***	−1.002**	0.342**	−0.014
		(0.316)	(0.155)	(0.415)	(0.128)	(0.851)
Supporting		0.235	1.628***	0.215	0.446***	0.866
		(0.321)	(0.151)	(0.424)	(0.129)	(0.424)
Ownership	Cat					
Domestic		–	0.228***	–	0.197***	–
Signif. codes:		. p<0.05; *p<0.01; **p<0.001; ***p<0.00				

Table D: Summary statistics for 3 estimated regression models for tax period 2015. The variables in bold were selected for the binary selection model but not for the outcome model. The values in brackets are standard errors.

	Type	OLS regression	Gaussian sample selection		Censored sample selection	
		–	Selection	Outcome	Selection	Outcome
Constant		9.50	−9.651***	9.327***	−6.551***	9.692***
		(0.00)	(0.248)	(0.480)	(0.210)	(0.851)
Value added	Num	0.259***		0.252***		−0.106***
		(0.032)		(0.014)		(0.031)
Compensations	Num	−0.404***	−0.080**	−0.416***	−0.099***	−0.322***
		(0.030)	(0.028)	(0.031)	(0.028)	(0.058)
Net assets	Num	–	−2.254***	–	−0.964***	–
		–	(0.206)	–	(0.220)	–
Total revenues	Num	–	−1.564***	–	−2.276***	–
		–	(0.293)	–	(0.307)	–
Value added / employee	Num	–	−0.052***	–	−0.029***	–
		–	(0.007)	–	(0.006)	–
Net assets / employee	Num	–	2.291***	–	0.869***	–
		–	(0.209)	–	(0.231)	–
Total revenues / employee	Num	–	2.181***	–	2.901***	–
		–	(0.301)	–	(0.32)	–
Value added / pers. costs	Num	−0.172***	−0.084***	−0.174***	−0.014	−0.032***
		(0.052)	(0.009)	(0.026)	(0.009)	(0.054)
Net assets / pers. costs	Num	−0.178***	0.242***	−0.237***	0.236***	−0.155***
		(0.019)	(0.049)	(0.029)	(0.053)	(0.054)
Total revenues / pers. costs	Num	–	−0.182***	–	−0.319***	–
		–	(0.052)	–	(0.056)	–
Period in loss	Cat					
zero		–	0.814***	–	0.549***	–
		–	(0.092)	–	(0.086)	–
one		–	0.297**	–	0.232*	–
		–	(0.107)	–	(0.101)	–
two		−0.578***	−0.183	−0.590	0.058	0.572
		(0.113)	(0.152)	(0.361)	(0.133)	(0.749)

Table D: (Continued 2/2) Summary statistics for 3 estimated regression models for tax period 2015. All variables were transformed into logarithms.

Distribution of the predicted and observed CIT noncompliance

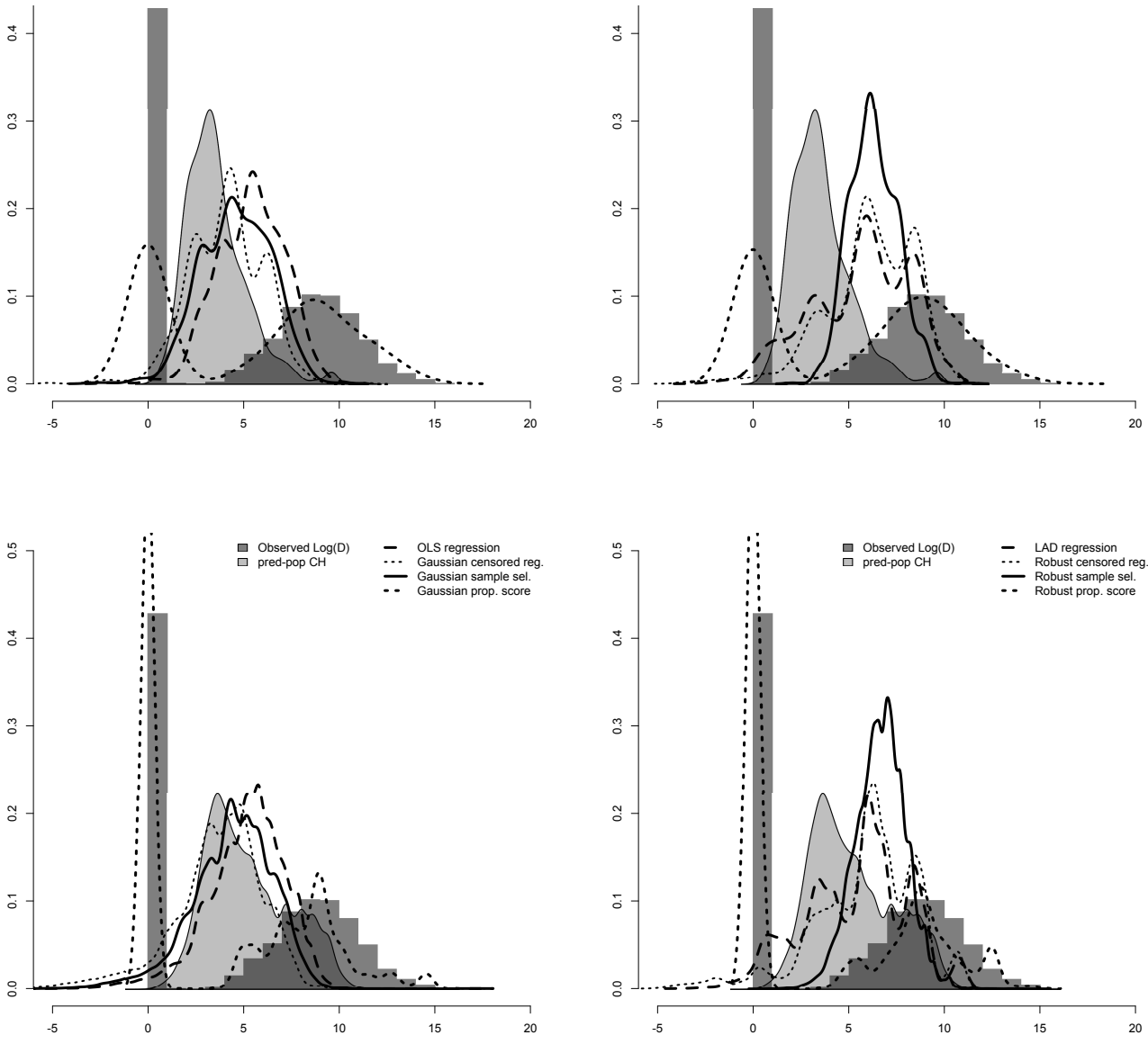

Figure A: The densities of the observed and predicted CIT log-deficiencies in the tax year 2015. The top row shows densities based on predictions for audited firms only. The bottom row includes all firms in the targeted population.

Simulation and empirical results based on the unbiased predictor

(i) First, we randomly select $np = 1100$ firms from the 1126 audited firms;
(ii) fit a weighted OLS predictive model using randomly selected $n = 1000$ firms with all 10 predictors selected by LASSO (see Table D);
(iii) we use the fitted model to predict the deficiency for the 1100 firms, i.e., both in-sample and out-of-sample and we compute the total deficiency using (5.8) and (5.9) as $\hat{\text{TD}} = \sum_{i=1}^{np} e^{\hat{y}}$;
(iv) estimation and prediction are repeated $nrep = 200\ 000$ times. This gives 200 000 predictions of TD for both the unbiased and the shrinked predictors;
(v) Next, we compute the root-mean-square-prediction-error as

$$\text{RMSPE} = \frac{1}{200\ 000}\sqrt{\sum_{i=1}^{200\ 000}\left(\text{TD}_i - \hat{\text{TD}}_i\right)^2}.$$

(vi) We compare the RMSPE of the unbiased and shrinked estimator to the naive transformation-bias corrected estimator, which uses sample average of the residuals as the estimator of the bias correction factor.

The results suggest that the relative RMSPE's of (5.8) is 1.53% and of (5.9) is 1.55%. Hence there is no strong evidence that correction for the transformation bias actually dominates in terms of mean square prediction error compared with the shrinkage.

The hypothesis $H_o : \text{E}e^{\varepsilon_O} \leq 1$ can be rejected on a 5% level (average t-test p-value based on the 200 000 runs is 0.019), but not on 1% level, which speaks in support of shrinking the parameter to 1.

As a alternative to the main results obtained with the shrinkage predictor in Table 7, we provide also results obtained with the unbiased predictor below in Table E.

R code for simulation results

```
## Large population scenario - based on observed data and actual linear model
# WARNING: before runnig you need to preload the y and x used in weighted ols
# model with 0 audits
##-----------------------------------------------------------------------------------
nrep=200000; np<-1100; n <- 1000; mz1 <- mz2 <- mz3  <-0;

pv_vec <- rep(0,nrep);
data<-as.data.frame(cbind(y,x[,NAME_SELVARS_Ogls_WITHzero]));
names(data)<-c('LOG_TAUFIN_FASR_NUM',names(data)[-1])
for (i in 1:nrep) {
  set.seed(i); pp<-data[sample(x=1:nrow(data),size = np, replace = TRUE),]
  my <- mean(pp[,1]);  s <- sd(pp[,1]); s2 <- s^2
  mz <- mean(exp(pp[,1])); sz <- sum(exp(pp[,1]))

  set.seed(i); s<-sample(1:np,n); y <- pp[,1]
  olsfit<-lm(as.formula(paste('LOG_TAUFIN_FASR_NUM~',
                              paste(NAME_SELVARS_Ogls_WITHzero,collapse = "+"))),data=pp[s,])
  studentt_res<-abs(residuals(olsfit)/sigma(olsfit)); studentt_res<-studentt_res^2
  olsfit<-lm(as.formula(paste('LOG_TAUFIN_FASR_NUM~',
                              paste(NAME_SELVARS_Ogls_WITHzero,collapse = "+"))),
             weights=1/studentt_res,data=pp[s,]); my1<-predict(olsfit, newdata=pp)
  f0<-exp(y-my1); f1<-mean(f0); f2<-median(f0); tmp<-t.test(x=f0,mu = 1,alternative = "greater");
  pv_vec[i]<-tmp$p.value;
  mz1 <- mz1 + (sz-sum(exp(my1)))^2; mz2 <- mz2 + (sz-sum(exp(my1))*f1)^2;
  mz3 <- mz3 + (sz-sum(exp(my1))*f2)^2; mz4 <- mz3 + (sz-np*exp(mean(my1))*f1)^2;

}
RRMSE<-sqrt(c(mz1,mz3)/mz2)
# [1] 0.01553108 0.01539296

##-----------------------------------------------------------------------------------
plot(cbind(seq_along(pv_vec),pv_vec), type="p",ylab = "p-val")
abline(h=0.05, col="red")
abline(h=0.01, col="green")
abline(h=mean(pv_vec), col="blue");mean(pv_vec)
# [1] 0.01884574
```

		CIT Gap (%)			TD (Mill. €)			mean D (€)			median D (€)			max D (Mill. €)			Bias correction factor		
		2014	2015	2016	2014	2015	2016	2014	2015	2016	2014	2015	2016	2014	2015	2016	2014	2015	2016
Data	*Observations*																		
	All 1126 audits	64	53	50	54	35	27	48 913	37 385	27 784	564	626	331	3.76	3.76	3.76	-	-	-
	Year-specific audits	61	68	18	23	14	0.23	56 535	35 467	4 486	613	874	0	3.76	1.47	0.10	-	-	-
Imputation (biased)	*Scaling*																		
	Naive	90	88	89	9 078	9 127	9 647	48 913	37 385	27 784	564	626	331	3.76	3.76	3.76	-	-	-
	Stratification	82	79	80	4 442	4 422	4 573	76 486	764 86	76 486	18 628	18 628	18 628	0.86	0.86	0.86	-	-	-
	Propensity matching																		
	Gaussian log-lik	82	88	89	4 574	8 437	9 426	25 141	46 310	48 372	0	0	0	3.76	3.76	3.76	-	-	-
	Semi-parametric	82	80	87	4 346	4 854	7 526	23 887	26 643	38 624	0	0	0	3.76	3.76	3.76	-	-	-
Computation by regression	*Linear regression*																		
	Least squares	26	25	53	340	393	1 286	1 868	2 155	6 597	289	298	303	22.07	23.60	24.72	1.42	1.42	1.42
	Least absolute dev.	35	31	34	536	547	582	2 944	3 004	2 986	332	345	347	0.54	0.63	0.54	1.00	1.00	1.00
	Censored regression																		
	Gaussian log-lik	47	61	74	862	1 886	3 347	4 740	10 354	17 178	359	348	300	65.76	61.10	161.72	5.65	6.67	4.39
	Semi-parametric	36	34	34	555	632	600	3 052	3 470	3 078	374	517	380	0.41	0.84	0.60	1.00	1.01	1.00
	Sample selection																		
	Gaussian log-lik	17	14	33	200	194	561	1 098	1 067	2 879	143	91	73	15.08	16.85	9.75	1.28	0.94	0.82
	Semi-parametric	25	20	24	317	307	356	1 741	1 687	1 826	692	688	514	2.04	1.17	1.04	0.98	0.85	0.60
	Censored Sample sel.																		
	Gaussian log-lik	26	34	28	342	610	444	1 881	3 346	2 277	305	194	237	1.04	7.42	1.49	1.01	1.57	0.75

Table E: All model-based predictions are bias-corrected (see Section 5) using the bias correction factor Ee^{ε} computed as the sample median of in-sample residuals. Note that using the sample average instead would lead to extremely volatile predictions since the distribution of e^{ε} is right-skewed and heavy-tailed. The rest of the Table is the same as in the Table 7