Shortfall in Tax Revenue: Evaluating the Social Security Contribution Fraud[*]

Denisa Banulescu-Radu[†] Sylvain Benoit[‡] Christophe Hurlin[§]

November 30, 2023

Preliminary version

Abstract

The loss in tax revenue is defined as the potential sum of the tax adjustments that could have been imposed on firms that have committed fraud, if they had been effectively controlled by an inspection authority, whereas they were not in reality. We introduce a new framework to deal with such a social contribution fraud in order to estimate accurately the shortfall in tax revenue. First, we define the tax shortfall from a theoretical point of view and we highlight the tractability of the theoretical approach via Monte Carlo simulations. Second, we validate empirically the proposed methodology for estimating the shortfall in tax revenue (STaX) by using data provided by the *Mutualité Sociale Agricole* (MSA).

Keywords: Social Contribution Fraud; Tax Shortfall; Econometric Methods; Monte Carlo Simulations

JEL classification: C01, C13, C35, H26.

[*] We would like to warmly thank the Department of Statistics, Studies and Funds of the French agricultural social security agency (MSA) for providing us the data and for their expert judgement. We are grateful to seminar participants at the 2022 Methodological and Statistical days of the INSEE for their comments. This research benefits from the financial support of MLEforRisk ANR-21-CE26-0007, of "Data Science and Insurance Fraud Detection" research initiative under the aegis of the Europlace Institut of Finance, the "Fraud detection and Anti-Money Laundering" research project (REDFLAG APR-IA AE 2019-1850) funded by Centre-Val de Loire Region, and the ANR program CaliBank (ANR-19-CE26-0002-02). The results are performed using Matlab 2022a.

[†] University of Orléans, LEO, Rue de Blois, 45067 Orléans, France. Email: denisa.banulescu-radu@univ-orleans.fr

[‡] Université Paris-Dauphine, PSL University, UMR CNRS 8007, LEDa-SDFi, 75016 Paris, France. E-mail: sylvain.benoit@dauphine.psl.eu

[§] University of Orléans, LEO, Rue de Blois, 45067 Orléans, France. Email: christophe.hurlin@univ-orleans.fr

1 Introduction

Controlling the risks of social and fiscal fraud and combating illegal work are important problems for social justice and economic efficiency, which aims to reaffirm the balance of rights and duties and to ensure the sustainability of the social protection system. This paper proposes to address social contribution fraud by estimating accurately the shortfall in tax revenue. This tax shortfall is defined as the potential sum of the tax adjustments that could have been imposed on firms having defrauded (or made erroneous social declarations), if they had been effectively controlled by an inspection authority, whereas they were not in reality.

Two types of agents are hence involved in the study: the firms (taxpayers) and a control authority. Each of them takes decisions: each firm decides whether to fraud or not to fraud, and the control authority decides which firms to monitor.[1] It is important to note that these two decisions are neither sequential nor conditional. The fact of being audited does not condition in any way the fact of defrauding and vice versa. However, these decisions are linked: if the inspection authority fulfills its mission effectively, one should observe a greater probability of auditing for the firms having the highest probability of fraud. Additionally, two types of biases should be considered: a selection bias (related to the fact that the choice of controlled companies is not random), and a detection bias (since the illegitimate actions are not all detected). The objective of the paper is twofold: firstly, we define the shortfall in tax revenue (hereafter, $STax$) from a statistical point of view, and secondly, we validate the empirical methods used to estimate such an amount. To highlight the tractability of our approach, we use Monte Carlo simulations before relying on real data.

We start by proposing a theoretical definition of the moments of the conditional distribution of $STax$. This definition is obtained as part of a model or data generating process (DGP) specifying several behavioral equations related to the decision of control made by a social security entity, the fraud decision of firms, and the rule of setting the amount of tax adjustment (which is assumed to be equal to the true amount of fraud). Under normality hypotheses, this model is akin to a Tobit model with a double censored mechanism. In this DGP, the $STax$ is a random variable whose realizations are by nature unobservable. However, conditionally to the DGP, it is possible to characterize the first two moments (expectation and

[1] The actions of controllers in the field of social security contributions consists of an in-depth examination of the elements declared by the establishments, in particular with respect to the employment of their employees.

variance) of its distribution, which will be used to predict the $STax$ for each firm and build confidence intervals. Due to the dependence between the control and fraud decisions, we show that the conditional moments of the individual $STax$ do not depend on simple Mills ratios, but on expectations and variances of multivariate normal laws with double truncation. There is a rich and complex literature which deals with moments of the normal distribution under various truncation conditions (unilateral, bilateral) and the number of variables (univariate, bivariate, multivariate). For this study, we focus on the theoretical propositions in Manjunath and Wilhelm (2012). The validity of the theoretical formulas is assessed through Monte Carlo simulations. We compare the simulated mean and empirical variance of the tax adjustment for the uncontrolled firms corresponding to the first two theoretical moments (expectation and variance) and observe that their distributions are very close.

The model proposed to estimate the individual $STax$ consists on three equations taking into account (i) the control decision, (ii) the fraud decision, and (iii) the amount of the tax adjustment. Its structure is similar to a model with a censorship on the tax adjustment, similar to a Type II Tobit model introduced by Amemiya (1984). The censorship mechanism is represented by a bi-Probit model itself censored, or equivalently by a nested Probit with dependencies (built around the two decisions of control and fraud/detection). The full model is estimated via Maximum Likelihood (ML). Monte Carlo simulations are used also in order to assess the validity of the proposed log-likelihood function.

Finally, an empirical application is performed on real data issued from controls carried out by the MSA (*Mutualité Sociale Agricole*) on firms in the French agricultural system.[2] We will further refer to as the control entity or authority. Because responsibility and solidarity are fundamental values of the MSA, controlling the risk of fraud and the fight against illegal work are at the heart of its concerns. Abuses and fraudulent behaviors, which harm all of its beneficiaries, engage the responsibility of the MSA with regard to the funds it manages. To combat illegal activities, the MSA collects data systematically from their beneficiaries and organizes regular controls on a subsample of their taxpayers. Therefore, estimating $STax$ makes it possible to quantify the financial impact of the illegal activity for the entire agricul-

[2] The MSA provides social cover for the entire agricultural population and beneficiaries in France: farmers, employees (of farms, companies, cooperatives and professional agricultural organisations), employers of labor work. The MSA was officially recognized as a professional organization by the Ministry of Agriculture in 1940. Since then, its mission is to manage all the social risks of agricultural policyholders. With 27.4 billion euros in benefits paid to 5.4 million beneficiaries, it is the second largest social protection scheme in France.

tural system (fraud, abuse, intentional optimization or error, etc.). To this end, we design an approach based on both statistical indicators and expert knowledge to identify the best specifications of the econometric model, and finally calculate the shortfall in tax revenue for each of the configurations retained.

To the best of our knowledge, this paper is the first to produce a tractable parametric model for estimating shortfall in tax revenue. This model can be easily applied by any control authority which is unable to audit all firms under its jurisdiction in order to evaluate the fraud of the non-controlled firms. Although fraud detection is a research domain with a wide variety of different applications, including credit card fraud, insurance fraud, telecommunication fraud, social fraud (see Banulescu-Radu and Yankol-Schalck, 2021; West and Bhattacharya, 2016; Baesens, Van Vlasselaer, and Verbeke, 2015, and the references therein), the academic literature on social security contribution fraud is very scarce. As a reference, we can cite Van Vlasselaer, Eliassi-Rad, Akoglu, Snoeck, and Baesens (2017) who study the impact of network information for social security fraud detection. Their analysis focuses on the identification of the companies that intentionally go bankrupt in order to avoid contributing their taxes.

The rest of the article is organized as follows. Section 2 sets the theoretical framework with a focus on the control and fraud decisions, the formal definition of the tax shortfall and the main assumptions and propositions used to construct the data generating process and then the parametric econometric model. Section 3 presents the design and the results of the Monte Carlo simulations used to evaluate the validity of the theoretical formulas. Section 4 describes and validates the approach used to estimate the complete model. Section 5 presents and discusses the empirical results whereas Section 6 provides some conclusive remarks.

2 Theoretical framework

2.1 Social security fraud detection background: control and fraud decisions

Two types of agents are considered: (1) a sample of n firms, and (2) an inspection authority. The two agents can take different decisions:

- each firm decides whether or not to commit fraud,
- the inspection authority decides whether or not to control a given firm.

It is important to note that these two decisions are neither sequential nor conditional[3]. The fact of being audited is not linked to the fact of fraud and vice versa. On the other hand, these decisions are linked: if the inspection authority fulfills its mission accurately, we should observe a greater likelihood of inspection for companies with the highest probability of fraud. This decision process, which will serve as the basis for the first part of our data generating process (DGP), is shown in Figure 1.

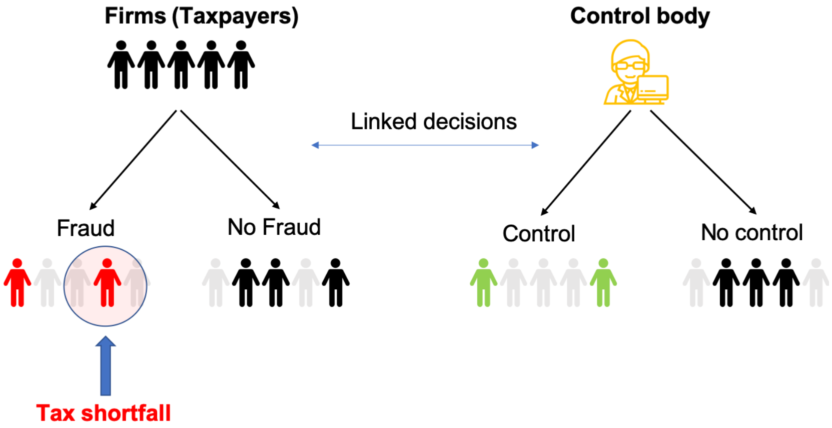

Figure 1: Decision process and theoretical definition of the tax shortfall

An important point at this stage concerns the observability of these two decisions. The control decision is observable: ex-post, it can be determined which firms have been audited and which have not. On the other hand, the fraudulent or non-fraudulent status of each firm is not observable in practice, independently of the control. This is why we introduce the notion of detection in the estimation (see Section 4): a firm is detected as fraudulent if and only if (i) it commits fraud, and (ii) it was controlled.

Therefore, the detected/not detected status is what we considered as observable in the estimation. However, at a first time we place ourselves from the point of view of an omniscient modeler who observes the type (fraud/non-fraud) of all firms, and the fraud decision is therefore assumed to be observable in the DGP. Formally, we introduce two dummy variables associated with these two decisions.

Let C_i be the dummy control variable indicating whether the firm i is controlled ($C_i = 1$)

[3] In our theoretical model, we neglect the temporal dimension and thus the decisions of fraud and control are supposed to be concomitant. On real data, it will be necessary to question the fact of dating or not the decisions. Indeed, a firm can start a fraud after having been audited by playing on the low probability of repeating an audit when the first audit is negative

or not ($C_i = 0$). The likelihood of an inspection relies on a set of k_c factors $\mathbf{X}_{c,i}$ based on the following mechanism:

$$C_i = \begin{cases} 1 & \text{if } \; C_i^* = \mathbf{X}_{c,i}\beta_c + \varepsilon_{c,i} > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \forall i = 1, \ldots, n \tag{1}$$

where C_i^* is a latent variable, β_c a vector with k_c parameters and $\varepsilon_{c,i}$ an i.i.d. error term such that $\mathbb{E}\left(\varepsilon_{c,i}\right) = 0$ et $\mathbb{V}\left(\varepsilon_{c,i}\right) = \sigma_c^2$.

Let $\widetilde{D}_i$ be the dummy detection variable indicating whether the firm i commits fraud ($\widetilde{D}_i = 1$) or not ($\widetilde{D}_i = 0$).[4] The likelihood of fraud relies on a set of k_d factors $\mathbf{X}_{d,i}$ and can be partly similar to the $\mathbf{X}_{c,i}$ factors. The choice to commit fraud is described by the following mechanism:

$$\widetilde{D}_i = \begin{cases} 1 & \text{if } \; D_i^* = \mathbf{X}_{d,i}\beta_d + \varepsilon_{d,i} > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \forall i = 1, \ldots, n \tag{2}$$

where D_i^* is a latent variable, β_d a vector with k_d parameters and $\varepsilon_{d,i}$ an i.i.d. error term such that $\mathbb{E}\left(\varepsilon_{d,i}\right) = 0$ and $\mathbb{V}\left(\varepsilon_{d,i}\right) = \sigma_d^2$.[5]

Note that the control and fraud decisions, C_i and $\widetilde{D}_i$, are linked in two ways. First, some variables may explain both the fraud and the control decisions (appearing at the same time in $X_{d,i,v}$ and $X_{c,i,v}$), i.e. common factors might lead both the firm to commit fraud and the inspection authority to control. Second, the error terms $\varepsilon_{c,i}$ and $\varepsilon_{d,i}$ may be correlated, meaning that we have a link between the omitted factors in the two equations. A positive correlation coefficient ρ_{cd} suggests a higher probability of control for firms with the largest probabilities of fraud.

2.2 Shortfall in tax revenue

We denote by $STax_i$, with $STax_i \in \mathbb{R}^+$, the individual tax shortfall associated with firm $i \in 1, \ldots, n$, where n is total number of firms, and $STax = \sum_{i=1}^{n} STax_i$ the global tax shortfall ($STax$). The individual $STax$ can be equal to zero when the firm does not commit fraud. To theoretically define $STax$, it is now necessary to focus on the amount of fraud, or equivalently on the amount of the tax adjustment.

[4] In practice, the decision to fraud of the firm is unobservable, this is why we note it $\widetilde{D}_i$ in opposition to the dummy detection variable D_i which is observable and introduced in Section 4.

[5] In Section 4, for estimating this model, we introduce an observed variable D_i equal to $\widetilde{D}_i$ for the firms controlled by the inspection authority ($C_i = 1$) and to 0 for the firms that have not been controlled.

Let M_i^* be the latent variable indicating the potential amount in euros of the tax adjustment for firm $i \in 1, \ldots, n$. We assume the existence of a linear relation between the potential amount M_i^* and a set of k_m explanatory variables $\mathbf{X}_{m,i}$, such that:

$$M_i^* = \begin{cases} \mathbf{X}_{m,i}\beta_m + \varepsilon_{m,i} & \text{if } \widetilde{D}_i = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \forall i = 1, \ldots, n, \tag{3}$$

where β_m is a vector of k_m parameters, and the error term $\varepsilon_{m,i}$ satisfies $\mathbb{E}(\varepsilon_{m,i}) = 0$ and $\mathbb{V}(\varepsilon_{m,i}) = \sigma_m^2$.[6] It is important to note that this potential adjustment amount is positive for all firms that have committed fraud, whether they were controlled or not.[7]

The potential amount of fraud is only observable for firms which have been (i) audited by the inspection authority, and (ii) reassessed following the discovery of fraud. We note M_i the amount of the tax adjustment actually observed such that:

$$M_i = \begin{cases} M_i^* & \text{if } C_i = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \forall i = 1, \ldots, n, \tag{4}$$

or equivalently

$$M_i = \begin{cases} \mathbf{X}_{m,i}\beta_m + \varepsilon_{m,i} & \text{if } C_i = 1 \text{ and } \widetilde{D}_i = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \forall i = 1, \ldots, n. \tag{5}$$

From the potential tax adjustment M_i^*, we can deduce the $STax_i$ for each firm in the population. $STax_i$ is positive for firms that have defrauded and not been audited, and zero for all others, as shown in Figure 1. Formally, if we denote $STax_i$ the shortfall associated with firm i, it verifies:

$$STax_i = M_i^* \times \mathbf{1}_{(C_i=0)} \times \mathbf{1}_{(\widetilde{D}_i=1)}, \quad \forall i = 1, \ldots, n, \tag{6}$$

Let us focus now on the aggregate $STax$, noted $STax = \sum_{i=1}^n STax_i$. It is defined by the sum of the potential adjustments that could have been imposed on fraudulent firms if they had been effectively audited, when in fact they were not.

Definition 1 *Ex-post, the aggregate $STax$ is defined by:*

$$STax = \sum_{i:(C_i=0)\cap(\widetilde{D}_i=1)} M_i^* = \sum_{i:C_i=0} M_i^* \times \mathbf{1}_{(\widetilde{D}_i=1)} = \sum_{i=1}^n M_i^* \times \mathbf{1}_{(C_i=0)} \times \mathbf{1}_{(\widetilde{D}_i=1)} \tag{7}$$

[6] In this specification, the notional amount of the adjustment can theoretically be negative. In order to avoid this, a solution consists in modeling the logarithm of the potential adjustment $\ln(M_i^*)$. In the case $\widetilde{D}_i = 1$, the logarithm guarantees the positivity of the adjusted amount, independently of the parameters values β_m and the distribution of $\varepsilon_{m,i}$, since $M_i^* = \exp(\mathbf{X}_{m,i}\beta_m + \varepsilon_{m,i}) > 0$. On the other hand, when the company had taken the decision not to defraud, the amount of the potential recovery M_i^* remains null.

[7] Based on a simulation exercise, we can therefore observe a potential adjustment amount M_i^* for all firms, and not just for firms that have been controlled.

where $\mathbf{1}_{(.)}$ is the dummy variable taking the value 1 when the condition is observed and 0 elsewhere.

We assume that the variable $STax_i$ is a random variable, and the aggregate $STax$ as well. In this case, our purpose is to determine the first two theoretical moments of its distribution, i.e. its expectation $\mathbb{E}\left(STax_i\right)$ and its variance $\mathbb{V}\left(STax_i\right)$.[8]

2.3 Assumptions and Propositions

In the current framework, the usual empirical counterparts for estimating the first two moments cannot be used since the realizations of the random variable $STax_i$ are latent. One solution is to assume a parametric distribution on the variable $STax_i$ for inferring a closed-form expression of the first two theoretical moments with respect to a vector of parameters β associated with this distribution such that $\mathbb{E}\left(STax_i\right) = f\left(\beta\right)$ and $\mathbb{V}\left(STax_i\right) = g\left(\beta\right)$.[9] Thus, when we have a consistent estimator $\widehat{\beta}$ of the true parameters of a distribution, we can directly compute the first two theoretical moments by using their explicit formulas $f(\widehat{\beta})$ and $g(\widehat{\beta})$.

More precisely, we focus on theoretical moments of the conditional distribution of the aggregate $STax$, and propose a three-step approach. First, we design a model for setting the conditional distribution of $STax_i$. Second, from this model, we compute closed-form formulas for the first two theoretical moments, i.e. the conditional expectation $\mathbb{E}\left(STax_i|\mathbf{X}_i = \mathbf{x}_i\right) = f\left(\mathbf{x}_i;\beta\right)$ and the conditional variance $\mathbb{V}\left(STax_i|\mathbf{X}_i = \mathbf{x}_i\right) = g\left(\mathbf{x}_i;\beta\right)$, where $\mathbf{X}_i$ is a set of explanatory variables. Third, we estimate the true model parameters β (see Section 4) to estimate the two conditional moments.

To be valid, this approach assumes that:

1. The postulated model on the conditional distribution of $STax$ must be well-specified;

2. The formulas for the conditional moments must be correct;

3. The method for estimating the model parameters must be valid and leads to convergent estimations.

[8] From these two moments, we can get a estimation of $STax$ along with its confidence interval. One alternative approach could be to fully describe the distribution of $STax$ or at least its fractiles with some Value-at-Risk (VaR). For example, we can report a VaR at 99% to determine a threshold such that we have a 1% probability of observing a value above this threshold.

[9] For example, if we assume that the variable $STax_i$ follows a normal distribution law $\mathcal{N}\left(\mu,\sigma^2\right)$, the first two theoretical moments are given by $\mathbb{E}\left(STax_i\right) = f\left(\beta\right) = \mu$ and $\mathbb{V}\left(STax_i\right) = g\left(\beta\right) = \sigma^2$ with $\beta = \left(\mu,\sigma^2\right)'$.

To go further, two technical assumptions about the distribution of the error terms and identifiability should be added for satisfying the previous elements.

Assumption A1 ***(normality):*** *We assume that the error terms $\varepsilon_{c,i}$ and $\varepsilon_{d,i}$ follow a bivariate normal distribution with ρ_{cd} as correlation coefficient.*

Assumption A2 ***(identification):*** *There is at least one explanatory variable for the control decision $X_{c,i,u} \in X_{c,i}$ and one explanatory variable for the fraud decision $X_{d,i,v} \in X_{d,i}$ such that $\mathbb{C}\text{ov}(X_{c,i,u}, X_{d,i,v}) = 0$, $\mathbb{C}\text{ov}(X_{c,i,u}, \varepsilon_{d,i}) = 0$, and $\mathbb{C}\text{ov}(X_{d,i,v}, \varepsilon_{c,i}) = 0$.*[10]

Assumption A1 implies that the two dummy variables $(C_i, \widetilde{D}_i)$ can be modeled by a usual bivariate Probit model. Assumption A2 imposes that at least one explanatory variable of the fraud decision is not linked to the set of explanatory variables defining the control decision, nor to its error term. Such a condition guarantees the strong identification of both control and fraud probabilities. In other words, without this assumption, we are unable to disentangle the two decisions, except if we want a weak identification based on the residuals of Equations 1 and 2.

Two additional technical assumptions should be added as well:

Assumption A3 ***(normality):*** *We assume that the error term $\varepsilon_{m,i}$ follows a normal distribution, and can be linked to the error term $\varepsilon_{c,i}$ and $\varepsilon_{d,i}$. Their corresponding correlation coefficient is denoted by ρ_{cm} and ρ_{dm}, respectively.*

Assumption A4 ***(identification):*** *There is at least one explanatory variable for the tax adjustment $X_{m,i,u} \in X_{m,i}$ and one explanatory variable for the control decision $X_{c,i,v} \in X_{c,i}$ such that $\mathbb{C}\text{ov}(X_{m,i,u}, X_{c,i,v}) = 0$, $\mathbb{C}\text{ov}(X_{m,i,u}, \varepsilon_{c,i}) = 0$, and $\mathbb{C}\text{ov}(X_{c,i,v}, \varepsilon_{m,i}) = 0$.*

The normality assumption A3 allows the actual amount M_i (defined in Equation 4) to be represented by a Tobit model. The factors of the amount of fraud $\mathbf{X}_{m,i}$ and of the fraud decision $\mathbf{X}_{d,i}$ can be identical or different depending on the censoring mechanism.[11] On the

[10] Another possibility for writing this assumption is to separate within the two equations the common and specific explanatory variables, which may note this as follows: $C_i^* = \mathbf{X}_{c,i}\beta_c + \mathbf{W}_{c,i}\delta_c + \varepsilon_{c,i}$ and $D_i^* = \mathbf{X}_{d,i}\beta_d + \varepsilon_{d,i}$, with $\mathbb{C}\text{ov}\left(\mathbf{W}_{c,i}, \mathbf{X}_{d,i}\right) = 0$ and $\mathbb{C}\text{ov}\left(\mathbf{W}_{c,i}, \varepsilon_{d,i}\right) = 0$.

[11] This model is a type I Tobit if the explanatory variables for the adjustment equation $\mathbf{X}_{m,i}$ and for the fraud equation $\mathbf{X}_{d,i}$ are identical and the error terms are identical $\varepsilon_{d,i} = \varepsilon_{m,i}$. Otherwise, it is a type II Tobit representation in the sense of Amemiya (1984).

other hand, just as for the fraud decision, for identification reasons, some determinants of the amount of fraud must be different from those of the control decision. Finally, the error terms can be linked. For instance, a positive ρ_{cm} correlation indicates a greater probability of control for firms with the highest potential amounts of fraud. In the end, we assume that the three error terms $\varepsilon_{c,i}$, $\varepsilon_{d,i}$, and $\varepsilon_{m,i}$ verify the following properties:

$$\begin{pmatrix} \varepsilon_{c,i} \\ \varepsilon_{d,i} \\ \varepsilon_{m,i} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}\right) \quad \text{with} \quad \mathbf{\Sigma} = \mathbf{DRD} \tag{8}$$

$$\mathbf{D} = \begin{pmatrix} \sigma_c & 0 & 0 \\ 0 & \sigma_d & 0 \\ 0 & 0 & \sigma_m \end{pmatrix} \qquad \mathbf{R} = \begin{pmatrix} 1 & \rho_{cd} & \rho_{cm} \\ \rho_{cd} & 1 & \rho_{dm} \\ \rho_{cm} & \rho_{dm} & 1 \end{pmatrix} \tag{9}$$

One last assumption must be done regarding the information structure on control and fraud decisions, since to compute the expectation and variance of $STax$, it is necessary to understand what is random and what is not random in the three equations of Definition 7.

Table 1 summarizes the link between the information structure and the nature of the variables entering into the $STax$ definition.

Table 1: Assumptions on the information structure and STax definition.

Note: The term r.v. means random variables.

Assumptions	Aggregate STax definition and variables status
C_i and $\widetilde{D}_i$ realized	$STax = \sum_{i:(C_i=0)\cap(\widetilde{D}_i=1)} \underbrace{M_i^*}_{\text{r.v.}}$ $= \sum_{i:C_i=0} \underbrace{M_i^*}_{\text{r.v.}} \times \mathbf{1}_{(\widetilde{D}_i=1)} = \sum_{i=1}^{n} \underbrace{M_i^*}_{\text{r.v.}} \times \mathbf{1}_{(C_i=0)} \times \mathbf{1}_{(\widetilde{D}_i=1)}$
C_i realized	$STax = \sum_{i:C_i=0} \underbrace{M_i^*}_{\text{r.v.}} \times \underbrace{\mathbf{1}_{(\widetilde{D}_i=1)}}_{\text{r.v.}} = \sum_{i=1}^{n} \underbrace{M_i^*}_{\text{r.v.}} \times \mathbf{1}_{(C_i=0)} \times \underbrace{\mathbf{1}_{(\widetilde{D}_i=1)}}_{\text{r.v.}}$
Ex-ante	$STax = \sum_{i:C_i=0} \underbrace{M_i^*}_{\text{r.v.}} \times \underbrace{\mathbf{1}_{(\widetilde{D}_i=1)}}_{\text{r.v.}} = \sum_{i=1}^{n} \underbrace{M_i^*}_{\text{r.v.}} \times \underbrace{\mathbf{1}_{(C_i=0)}}_{\text{r.v.}} \times \underbrace{\mathbf{1}_{(\widetilde{D}_i=1)}}_{\text{r.v.}}$

We identify three cases:

1. If we compute the moments of the $STax$ once the control and fraud decisions (C_i and $\widetilde{D}_i$, respectively) are observed for all firms, the three $STAx$ definition are perfectly

identical. In this case, the dummy variables $\mathbf{1}_{(C_i=0)}$ and $\mathbf{1}_{(\widetilde{D}_i=1)}$ are not random, and the single unknown component of $STax$ comes from the amount of tax adjustment M_i^*.

2. If we compute the moments of $STax$ once the control decision (C_i) is observed for all firms, but the fraud decision is unobserved, then the dummy variable $\mathbf{1}_{(\widetilde{D}_i=1)}$ is random and the amount of tax adjustment, M_i^*, as well.

3. If we compute the moments of $STax$ ex-ante, i.e. without observing control and fraud decisions (C_i and $\widetilde{D}_i$, respectively), then the dummy variables $\mathbf{1}_{(\widetilde{D}_i=1)}$ and $\mathbf{1}_{(C_i=0)}$, and the amount of tax adjustment, M_i^*, are random variables.

Depending on the hypothesis retained on the information structure, the definition of the moments is not the same. In Appendix A, we detail the expressions of the first two moments of $STax$ for the three informational hypotheses. To ease the presentation, we retain the most relevant informational hypothesis and present the results only for this one.

Assumption A5 ***(information structure)*** *We assume that we evaluate the moments of the aggregate $STax$ while the control decision has been made by the inspection authority, but the firm's fraud decision is not observable.*

Proposition 1 *Under the Assumptions A1-A5 and the data generator process described by Equations 1-9, the first two conditional moments of the aggregate $STax$ satisfy:*

$$\mathbb{E}_X\left(STax\right) = \sum_{i:C_i=0} \mathbb{E}_X\left(M_i^* | \left(C_i=0\right) \cap \left(\widetilde{D}_i=1\right)\right) \times \Pr\left(\widetilde{D}_i=1|C_i=0\right) \tag{10}$$

$$\mathbb{V}_X\left(STax\right) = \sum_{i:C_i=0} \mathbb{V}_X\left(M_i^* | \left(C_i=0\right) \cap \left(\widetilde{D}_i=1\right)\right) \times \Pr\left(\widetilde{D}_i=1|C_i=0\right) \tag{11}$$

where $\mathbf{X} = (\mathbf{X}_c : \mathbf{X}_d : \mathbf{X}_m)$ is the set of explanatory variables of the model, and $\mathbb{E}_X(.) \equiv \mathbb{E}(.|\mathbf{X}=\mathbf{x})$ and $\mathbb{V}_X(.) \equiv \mathbb{V}(.|\mathbf{X}=\mathbf{x})$ are the conditional expectation and variance with respect to $\mathbf{X}$.

Note that under the previous Assumptions A1-A5, the $STax$ of two firms i and j are conditionally independent, i.e. $\mathbb{C}ov\left(STax_i, STax_j|\mathbf{X}=\mathbf{x}\right)$ with $i \neq j$ which explains that the variance of the aggregate $STax$ is defined as the sum of the variances of individual $STax$ of the two firms.

To complete the definition of the theoretical moments of the aggregate $STax$, we must characterize the conditional moments $\mathbb{E}_X(M_i^*|\,(C_i = 0) \cap (\widetilde{D}_i = 1))$ and $\mathbb{V}_X(M_i^*|\,(C_i = 0) \cap (\widetilde{D}_i = 1))$, as well as the conditional probability $Pr(\widetilde{D}_i = 1|C_i = 0)$. Let's start with the latter, under Assumptions A1-A5 we have:

$$\begin{aligned}
\Pr(\widetilde{D}_i &= 1|C_i = 0) = \Pr\left(\varepsilon_{d,i} > -\mathbf{X}_{d,i}\beta_d|\varepsilon_{c,i} < -\mathbf{X}_{c,i}\beta_c\right) \qquad (12)\\
&= 1 - \Pr\left(\varepsilon_{d,i} < -\mathbf{X}_{d,i}\beta_d|\varepsilon_{c,i} < -\mathbf{X}_{c,i}\beta_c\right)\\
&= 1 - \frac{\Phi\left(-\mathbf{X}_{c,i}\beta_c, -\mathbf{X}_{d,i}\beta_d; \mathbf{\Sigma}_{cd}\right)}{\Phi\left(-\mathbf{X}_{c,i}\beta_c/\sigma_c\right)}
\end{aligned}$$

where $\mathbf{\Sigma}_{cd}$ corresponds to the covariance (VCV) matrix between $(\varepsilon_{c,i}, \varepsilon_{d,i})'$, $\Phi\left(.,.;\mathbf{\Sigma}_{cd}\right)$ is the cumulative distribution function (cdf) of a bivariate normal distribution with an expectation equal to 0 and a VCV matrix denoted by $\mathbf{\Sigma}_{cd}$, and $\Phi\left(.\right)$ is the cdf of the univariate standard normal distribution.

The conditional expectation of the individual $STax$ can be written as:[12]

$$\begin{aligned}
\mathbb{E}_X(M_i^*|\,(C_i = 0) \quad &\cap \quad (\widetilde{D}_i = 1)) = \mathbf{X}_{m,i}\beta_m + \mathbb{E}_X\left(\varepsilon_{m,i}|\left(\varepsilon_{c,i} < -\mathbf{X}_{c,i}\beta_c\right) \cap \left(\varepsilon_{d,i} > -\mathbf{X}_{d,i}\beta_d\right)\right)\\
&= \mathbf{X}_{m,i}\beta_m + \delta_c\mathbb{E}_X\left(\varepsilon_{c,i}|\left(\varepsilon_{c,i} < -\mathbf{X}_{c,i}\beta_c\right) \cap \left(\varepsilon_{d,i} > -\mathbf{X}_{d,i}\beta_d\right)\right)\\
&\quad +\delta_d\mathbb{E}_X\left(\varepsilon_{d,i}|\left(\varepsilon_{c,i} < -\mathbf{X}_{c,i}\beta_c\right) \cap \left(\varepsilon_{d,i} > -\mathbf{X}_{d,i}\beta_d\right)\right) \qquad (13)
\end{aligned}$$

where δ_c and δ_d are the estimated coefficients of $\varepsilon_{m,i}$ with respect to $\varepsilon_{c,i}$ and $\varepsilon_{d,i}$, respectively:

$$\delta_c = \frac{\sigma_{mc}\sigma_d^2 - \sigma_{md}\sigma_{cd}}{\sigma_c^2\sigma_d^2 - \sigma_{cd}^2} \qquad (14)$$

$$\delta_d = \frac{\sigma_{md}\sigma_c^2 - \sigma_{mc}\sigma_{cd}}{\sigma_c^2\sigma_d^2 - \sigma_{cd}^2} \qquad (15)$$

Accordingly, the conditional variance of individual $STax$ is defined by:

$$\begin{aligned}
\mathbb{V}_X(M_i^*|\,(C_i = 0) \quad &\cap \quad (\widetilde{D}_i = 1)) = \mathbb{V}_X\left(\varepsilon_{m,i}|\left(\varepsilon_{c,i} < -\mathbf{X}_{c,i}\beta_c\right) \cap \left(\varepsilon_{d,i} > -\mathbf{X}_{d,i}\beta_d\right)\right)\\
&= \delta_c^2\mathbb{V}_X\left(\varepsilon_{c,i}|\left(\varepsilon_{c,i} < -\mathbf{X}_{c,i}\beta_c\right) \cap \left(\varepsilon_{d,i} > -\mathbf{X}_{d,i}\beta_d\right)\right)\\
&\quad +\delta_d^2\mathbb{V}_X\left(\varepsilon_{d,i}|\left(\varepsilon_{c,i} < -\mathbf{X}_{c,i}\beta_c\right) \cap \left(\varepsilon_{d,i} > -\mathbf{X}_{d,i}\beta_d\right)\right)\\
&\quad +2\delta_c\delta_d\mathbb{C}ov_X\left(\varepsilon_{c,i}, \varepsilon_{d,i}|\left(\varepsilon_{c,i} < -\mathbf{X}_{c,i}\beta_c\right) \cap \left(\varepsilon_{d,i} > -\mathbf{X}_{d,i}\beta_d\right)\right) \qquad (16)
\end{aligned}$$

[12] In fact, the error term $\varepsilon_{m,i}$ can be defined as follows $\varepsilon_{m,i} = \delta_c\varepsilon_{c,i} + \delta_d\varepsilon_{d,i} + \mu_i$, with $\mathbb{C}ov\left(\mu_i, \varepsilon_{c,i}\right) = \mathbb{C}ov\left(\mu_i, \varepsilon_{d,i}\right) = 0$. The conditional expectation being a linear operator, we obtain:

$$\begin{aligned}
\mathbb{E}_X\left(\varepsilon_{m,i}|\left(\varepsilon_{c,i} < -\mathbf{X}_{c,i}\beta_c\right) \cap \left(\varepsilon_{d,i} > -\mathbf{X}_{d,i}\beta_d\right)\right) &= \delta_c\mathbb{E}_X\left(\varepsilon_{c,i}|\left(\varepsilon_{c,i} < -\mathbf{X}_{c,i}\beta_c\right) \cap \left(\varepsilon_{d,i} > -\mathbf{X}_{d,i}\beta_d\right)\right)\\
&\quad +\delta_d\mathbb{E}_X\left(\varepsilon_{d,i}|\left(\varepsilon_{c,i} < -\mathbf{X}_{c,i}\beta_c\right) \cap \left(\varepsilon_{d,i} > -\mathbf{X}_{d,i}\beta_d\right)\right)
\end{aligned}$$

since $\mathbb{E}_X\left(\mu_i|\left(\varepsilon_{c,i} < -\mathbf{X}_{c,i}\beta_c\right) \cap \left(\varepsilon_{d,i} > -\mathbf{X}_{d,i}\beta_d\right)\right) = \mathbb{E}_X\left(\mu_i\right) = 0$ by definition.

These different results can be summarized by the following two propositions.

Proposition 2 *Under the assumptions A1-A5 and the data generator process described by Equations 1-9, the conditional expectation of the aggregate $STax$ is defined by:*

$$\begin{aligned}\mathbb{E}_X\left(STax\right) \quad = \quad & \sum_{i:C_i=0} \mathbf{X}_{m,i}\beta_m \left(1-\frac{\Phi\left(b_{c,i},a_{d,i};\mathbf{\Sigma}_{cd}\right)}{\Phi\left(b_{c,i}/\sigma_c\right)}\right) \\ & +\delta_c \sum_{i:C_i=0} \mathbb{E}_X\left(\varepsilon_{c,i}\right|\left(\varepsilon_{c,i}<b_{c,i}\right)\cap\left(\varepsilon_{d,i}>a_{d,i}\right)) \times \left(1-\frac{\Phi\left(b_{c,i},a_{d,i};\mathbf{\Sigma}_{cd}\right)}{\Phi\left(b_{c,i}/\sigma_c\right)}\right) \\ & +\delta_d \sum_{i:C_i=0} \mathbb{E}_X\left(\varepsilon_{d,i}\right|\left(\varepsilon_{c,i}<b_{c,i}\right)\cap\left(\varepsilon_{d,i}>a_{d,i}\right)) \times \left(1-\frac{\Phi\left(b_{c,i},a_{d,i};\mathbf{\Sigma}_{cd}\right)}{\Phi\left(b_{c,i}/\sigma_c\right)}\right)\end{aligned} \tag{17}$$

where the truncation thresholds are defined by $b_{c,i}=-\mathbf{X}_{c,i}\beta_c$ and $a_{d,i}=-\mathbf{X}_{d,i}\beta_d$.

Proposition 3 *Under the Assumptions A1-A5 and the data generator process described by Equations 1-9, the conditional variance of the aggregate $STax$ is defined by:*

$$\begin{aligned}\mathbb{V}_X\left(STax\right) \quad = \quad & \delta_c^2 \sum_{i:C_i=0} \mathbb{V}_X\left(\varepsilon_{c,i}\right|\left(\varepsilon_{c,i}<b_{c,i}\right)\cap\left(\varepsilon_{d,i}>a_{d,i}\right)) \times \left(1-\frac{\Phi\left(b_{c,i},a_{d,i};\mathbf{\Sigma}_{cd}\right)}{\Phi\left(b_{c,i}/\sigma_c\right)}\right) \\ & +\delta_d^2 \sum_{i:C_i=0} \mathbb{V}_X\left(\varepsilon_{d,i}\right|\left(\varepsilon_{c,i}<b_{c,i}\right)\cap\left(\varepsilon_{d,i}>a_{d,i}\right)) \times \left(1-\frac{\Phi\left(b_{c,i},a_{d,i};\mathbf{\Sigma}_{cd}\right)}{\Phi\left(b_{c,i}/\sigma_c\right)}\right) \\ & +2\delta_c\delta_d \sum_{i:C_i=0} \mathbb{C}ov_X\left(\varepsilon_{c,i},\varepsilon_{d,i}\right|\left(\varepsilon_{c,i}<b_{c,i}\right)\cap\left(\varepsilon_{d,i}>a_{d,i}\right)) \times \left(1-\frac{\Phi\left(b_{c,i},a_{d,i};\mathbf{\Sigma}_{cd}\right)}{\Phi\left(b_{c,i}/\sigma_c\right)}\right)\end{aligned} \tag{18}$$

where the truncation thresholds are defined by $b_{c,i}=-\mathbf{X}_{c,i}\beta_c$ and $a_{d,i}=-\mathbf{X}_{d,i}\beta_d$.

As soon as we know the value of the parameters β_c, β_d, β_m and $\mathbf{\Sigma}$, the results of the propositions (2) and (3) allow us to build a forecast of the aggregate $STax$ according to the characteristics $\mathbf{X}=(\mathbf{X}_c:\mathbf{X}_d:\mathbf{X}_m)$ of the uncontrolled firms $i:C_i=0$, as well as a confidence interval on this forecast.

The forecast of aggregate $STax$, denoted $\widehat{STax}$, and the $1-\alpha\%$ confidence interval associated with this forecast are defined by:

$$\widehat{STax}=\mathbb{E}_X\left(STax\right) \tag{19}$$

$$IC_{1-\alpha}=\left[\mathbb{E}_X\left(STax\right)\pm\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\mathbb{V}_X\left(STax\right)}\right] \tag{20}$$

These formulas are then used in Monte Carlo simulations and then on real data to estimate the aggregate $STax$.

2.4 Comments

When control and fraud decisions are not linked, more precisely if $\rho_{cd} = 0$, the formulas reported in Propositions 2 and 3 include the inverse of the Mills ratio. Indeed, if $\rho_{cd} = 0$, we have:

$$\begin{aligned}
\mathbb{E}_X\left(STax\right) &= \sum_{i:C_i=0} \mathbf{X}_{m,i}\beta_m \left(1-\Phi\left(a_{d,i}/\sigma_d\right)\right) \\
&\quad +\delta_c \sum_{i:C_i=0} \mathbb{E}_X\left(\varepsilon_{c,i}|\varepsilon_{c,i} < b_{c,i}\right) \times \left(1-\Phi\left(a_{d,i}/\sigma_d\right)\right) \\
&\quad +\delta_d \sum_{i:C_i=0} \mathbb{E}_X\left(\varepsilon_{d,i}|\varepsilon_{d,i} > a_{d,i}\right) \times \left(1-\Phi\left(a_{d,i}/\sigma_d\right)\right) \\
&= \sum_{i:C_i=0} \mathbf{X}_{m,i}\beta_m \left(1-\Phi\left(a_{d,i}/\sigma_d\right)\right) - \delta_c \sum_{i:C_i=0} \sigma_c \frac{\phi\left(b_{c,i}/\sigma_c\right)}{\Phi\left(b_{c,i}/\sigma_c\right)} \times \left(1-\Phi\left(a_{d,i}/\sigma_d\right)\right) \\
&\quad +\delta_d \sum_{i:C_i=0} \sigma_d \frac{\phi\left(a_{d,i}/\sigma_d\right)}{1-\Phi\left(a_{d,i}/\sigma_d\right)} \times \left(1-\Phi\left(a_{d,i}/\sigma_d\right)\right) \qquad (21)
\end{aligned}$$

In the general, when case $\rho_{cd} \neq 0$, the expression of the conditional moments of the tax adjustment M_i^* involves the moments of a bivariate normal law with double truncation. There is an extensive literature dealing with the moments of the normal distribution under different truncation conditions (unilateral, bilateral) and the number of variables (univariate, bivariate, multivariate). Rosenbaum (1961) provides a formula for the moments of a bivariate distribution with an upper truncation. Khatri and Jaiswal (1963) propose a recurrence relation to obtain all bivariate moments for the lower truncated case. For the doubly truncated case, Shah and Parikh (1964) and Dyer (1973) provide recurrence formulas for the bivariate moments. Begier and Hamdan (1971) give an explicit formula for the moments of doubly truncated bivariate normal variables with the same lower limit points. This result is extended by Muthén (1990) to different limits. Horrace (2005) provides different analytical results for multivariate distributions with single truncation. Using the moment generating function, Manjunath and Wilhelm (2012) extend these results in the case of a multivariate distribution with arbitrary double truncation and create also an R package tmvtnorm (Manjunath and Wilhelm, 2010). More recently, Kan and Robotti (2017) propose an alternative approach based on recurrence relations between integrals that involve the density of the multivariate normal. These recurrences allow in particular to reduce the computation time for high dimensional normal laws. The authors propose different codes to compute the moments for multivariate distributions with arbitrary double truncations.

The most important point is that these double-truncated moments can be very different from the inverses of Mills ratios which are typically used to address selection problems (Heckman, 1976, 1979). In Appendix B, we report the general expression of the moments of a multivariate normal distribution with double truncation (Manjunath and Wilhelm, 2012). Except for particular cases, there is no analytical form for these moments, this is why we illustrate this difference through a numerical illustration reported in Appendix C.

3 Monte Carlo Simulations

In this section, we evaluate the validity of our theoretical formulas through Monte Carlo simulations. First, we compare the simulated mean and empirical variance of the tax adjustment for the uncontrolled firms corresponding to the first two theoretical moments (expectation and variance) computed in Section 2.3 (see Equations 17 and 18).

3.1 Parameters Setting

Based on the DGP described by the Equations (1)-(9), parameters are set as follows: we consider $k_c = 4$ explanatory variables for the equation of the latent variable C_i^*, $k_d = 2$ explanatory variables for the equation of the latent variable D_i^*, and $k_m = 1$ explanatory variables for the equation of the latent variable M_i^*. We assume that the explanatory variables $\mathbf{X}_{c,i}$, $\mathbf{X}_{d,i}$ and $\mathbf{X}_{m,i}$ are i.i.d. for all $i = 1, \ldots, n$, and satisfy $\mathbf{X} = (\mathbf{X}_{c,i} : \mathbf{X}_{d,i} : \mathbf{X}_{m,i}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{k_c+k_d+k_m})$. For each equation, we add a constant in each set of explanatory variables. The constant value in the equation of C_i^* is determined such that the marginal probability of control is equal to 5%, i.e. $\mathbb{E}(C_i) = 5\%$. The constant value in the equation of D_i^* is determined such that the marginal probability of fraud is equal to 10%, i.e. $\mathbb{E}(\widetilde{D}_i) = 10\%$. Finally, the constant value in the equation of M_i^* is determined such that the probability of observing a potential negative tax adjustment is lower than 0.01%. Parameters of other explanatory variables are defined by integers with an alternation of positive and negative signs as following:

$$\beta_c' = \begin{pmatrix} -9.27 & 1 & -2 & 3 & -4 \end{pmatrix}$$

$$\beta_d' = \begin{pmatrix} -10.13 & 5 & -6 \end{pmatrix}$$

$$\beta'_m = \begin{pmatrix} 33.75 & 7 \end{pmatrix}$$

Finally, we assume that the errors $\sigma_{c,i}$, $\sigma_{d,i}$ and $\sigma_{m,i}$ are correlated and that the variance of the error term in the tax adjustment equation has a higher variance than the variance of the error term in the fraud equation, which itself has a higher variance than the variance of the error term in the control equation. Formally, we assume $\sigma_m^2 > \sigma_d^2 > \sigma_c^2$ and set the values of the variance $\mathbf{D}$ and correlation $\mathbf{R}$ matrices to:

$$\mathbf{D} = \begin{pmatrix} \sqrt{2} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{5} \end{pmatrix} \qquad \mathbf{R} = \begin{pmatrix} 1 & 0.8 & 0.3 \\ 0.8 & 1 & 0.5 \\ 0.3 & 0.5 & 1 \end{pmatrix}$$

$$\mathbf{\Sigma} = \begin{pmatrix} 2.0000 & 1.6000 & 0.9487 \\ 1.6000 & 2.0000 & 1.5811 \\ 0.9487 & 1.5811 & 5.0000 \end{pmatrix}$$

Given this parametrization, the conditional probability of control for the fraudulent firms is equal to 5.30%, while the conditional probability of control for the non-fraudulent firms is of 4.61%. This configuration also implies that the average ratio of effective adjustments to the aggregate $STax$ is equal to 6%.

3.2 Results based on one simulation

We simulate the latent variables C_i^*, D_i^*, M_i^* as well as the variables C_i, $\widetilde{D}_i$ and M_i for all firms $i \in \{1, \ldots, n\}$ of the population, with $n = 10,000$. As a reminder, in this simulation exercise we assume that the fraud indicator $\widetilde{D}_i$ is observable, which is obviously not the case when we try to estimate the parameters of the model (see Section 4).

Contrary to what happens in reality, we can observe the realizations of the potential amount of tax adjustment M_i^* for all firms, including those that have not been audited. From these observations, we can infer a realization of $STax$ for all unaudited firms, which is positive for defrauding firms and zero for others. The realization of $STax$ for an uncontrolled firm that commits fraud is written as[13]:

$$stax_i = m_i^* \times \mathbf{1}_{(c_i=0)} \times \mathbf{1}_{(\widetilde{d}_i=1)} = \exp\left(\mathbf{x}_{m,i}\beta_m + \varepsilon_{m,i}\right) \times \mathbf{1}_{(c_i=0)} \times \mathbf{1}_{(\widetilde{d}_i=1)} \tag{22}$$

The simulated aggregate $STax$ is then equal to:

$$stax = \sum_{i=1}^{n} stax_i = \sum_{i=1}^{n} \exp\left(\mathbf{x}_{m,i}\beta_m + \varepsilon_{m,i}\right) \times \mathbf{1}_{(c_i=0)} \times \mathbf{1}_{(\widetilde{d}_i=1)} \tag{23}$$

[13] By convention, if we write the realizations of the random variables in lower case.

In Figure 2, we plot the amounts of notional tax adjustments m_i^* (in red) and the amounts of actual tax adjustments m_i (in blue) obtained for a particular simulation. For this simulation, $1,038$ firms actually defraud (the empirical frequency of fraud is equal to 10.38%) and 468 firms were audited (the empirical frequency of auditing is equal to 4.68%). Out of the $1,038$ fraudulent firms, 55 were controlled (5.30%) and 983 were not controlled (94.70%). In this simulation, the sum of the effective tax adjustments $\sum_{i=1}^{n} m_i$ is equal to $1,867$, while the realized aggregate $STax$ ($stax = \sum_{i=1}^{n} stax_i$) is equal to $33,796$, implying a ratio of 5.52%.

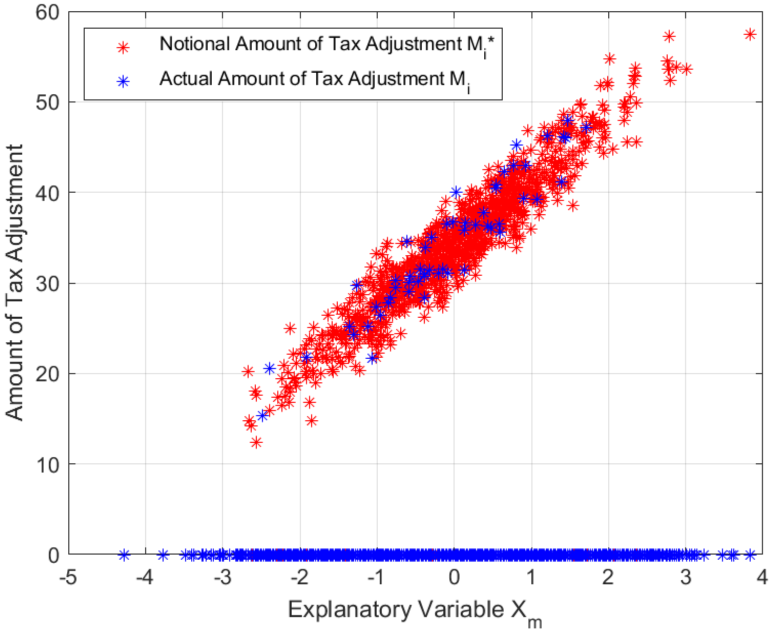

Figure 2: Amount of true and realized tax adjustment.

In the top panel of Figure 3, we represent the distribution of individual $STax$ for the $9,532$ firms that were not controlled. A very large majority of these firms did not defraud and have a zero $STax$. In the bottom panel, we represent the distribution of individual $STax$ for the 983 firms that actually defrauded among these firms that were not controlled. For these firms, the mean of the individual $STax$ is 34.38 and the variance is equal to 53.45.

In this simulation, we can check the validity of our forecast and confidence interval formulas on the aggregate $STax$. By applying the formulas from Equations 17 and 18, we have $\widehat{STax} = \mathbb{E}_X(STax) = 33,420$; $\mathbb{V}_X(STax) = 1,028$; and $IC_{95\%} = [33,357; 33,483]$. In this case, the realization of the aggregate $STax$, i.e. $stax = 33,796$, is relatively close to the realization of

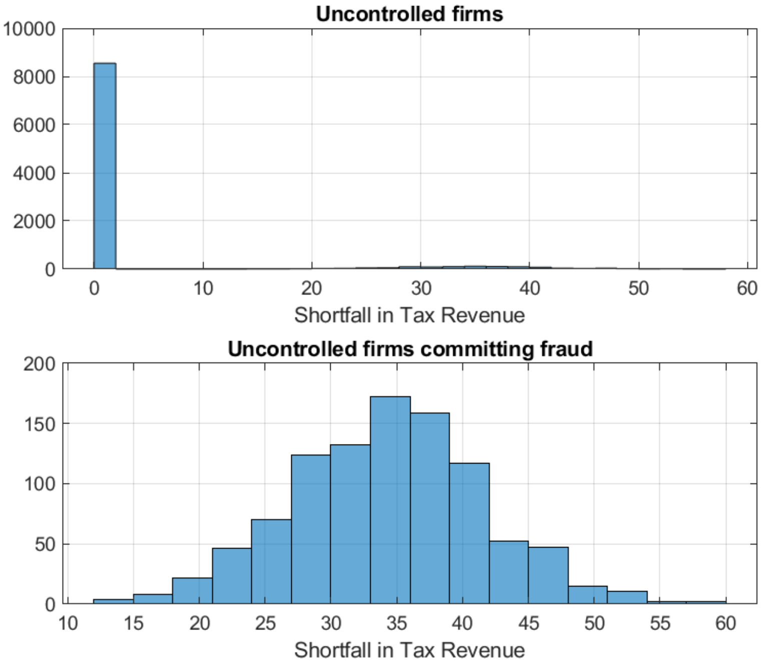

Figure 3: Distribution of realized individual STax.

the $STax$ expectation, even if this value does not belong to the 95% confidence interval.

The top panel of Figure 4 displays the distribution of individual $STax$ for the $9,532$ firms that were not controlled, while the bottom panel represents the distribution for the 983 firms that did defraud among them. First, we observe that the distributions are very close, which confirms the validity of our predictions. However, there is a difference: in reality all the $9,532$ firms have a zero $STax$ because fraud is not observable for the uncontrolled entities. In the current DGP, the $STax$ expectation of uncontrolled fraudulent firms is different from zero since it is the product of a conditional expectation M_i by a probability of fraud, which even if correctly estimated, are never zero. Therefore, the values of the expectation for these firms are small but not zero. This explains the difference on the left of the two distributions.

Figure 5 shows observed individual $stax_i$ (y-axis) compared to the conditional expectations $\mathbb{E}_X\left(STax_i\right)$ (x-axis), for all uncontrolled firms. We observe that for many uncontrolled firms the effective $STax$ is zero because they did not commit fraud, whereas the model tends to assign them a non-zero $STax$ expectation due to their probability of fraud and the notional amount expectation. Despite this caveat, the model performs well for all firms that did

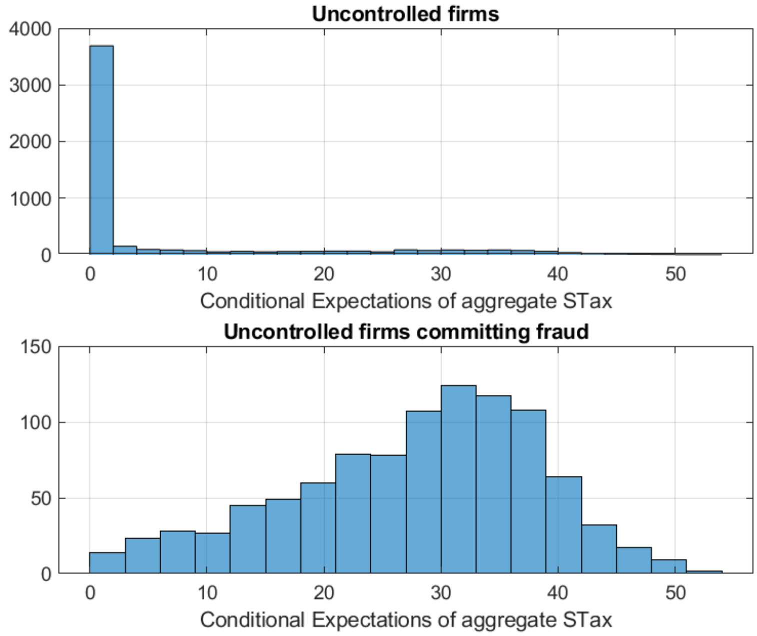

Figure 4: Histogram of realized and predicted individual STax.

defraud.

3.3 Results based on repeated simulations

We now repeat the previous experiment $1,000$ times. For each run, we keep the aggregate $STax$ ($stax = \sum_{i=1}^{n} stax_i$) and the forecast of the aggregate $STax$ defined by its conditional expectation $\widehat{STax} = \mathbb{E}_X\left(STax\right)$. Figure 6 represents the histograms of the realizations and the predictions obtained for the $1,000$ simulations. We observe that the distributions are very close.

Figures 7 and 8 display the empirical densities estimated by kernel estimators of the densities of the realized $STax$ and predicted $STax$. These distributions are also very close, which confirms the validity of the formulas for the $STax$ expectation (Equation 17).

Finally, Figure 9 displays the scatter plot of the realizations and the predictions of aggregate $STax$. This diagram confirms the goodness of fit captured by the $STax$ expectation (Equation 17).

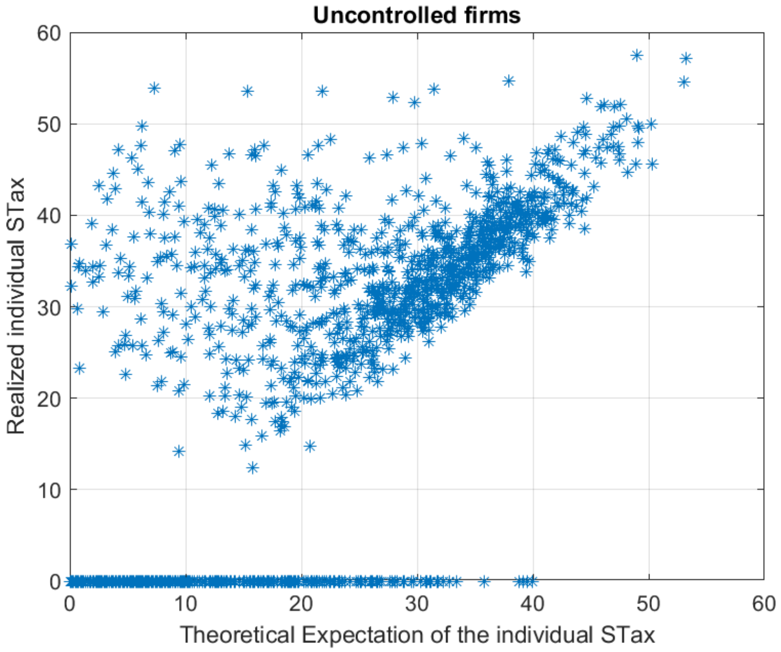

Figure 5: Scatter plot of individual STax and their forecasts.

4 Estimation

This section presents the approach used to estimate the $k_c + k_d + k_m + 4$ parameters denoted $\theta = (\beta_c' \; \beta_d' \; \beta_m' \; vech\,(\mathbf{\Sigma}))'$.

4.1 Empirical Design

The estimation is realized on the sample observed for the variables $\{C_i, D_i, M_i\}$. As previously mentioned, D_i is equal to $\widetilde{D}_i$ and represents the fraud decision observed *only* for controlled companies. This dichotomous variable D_i indicates whether the controlled firms $i = 1, 2, \ldots, n$ has been redressed ($D_i = \widetilde{D}_i = 1$) or not ($D_i = \widetilde{D}_i = 0$) following the control. For the rest of the paper, the variable D_i will be hence associated to the fraud detection. The estimated model is represented by the equations (24) to (28):

$$C_i = \begin{cases} 1 & \text{if } \; C_i^* = \mathbf{X}_{c,i}\beta_c + \varepsilon_{c,i} > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \forall i = 1, \ldots, n \tag{24}$$

$$\widetilde{D}_i = \begin{cases} 1 & \text{if } \; D_i^* = \mathbf{X}_{d,i}\beta_d + \varepsilon_{d,i} > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \forall i = 1, \ldots, n \tag{25}$$

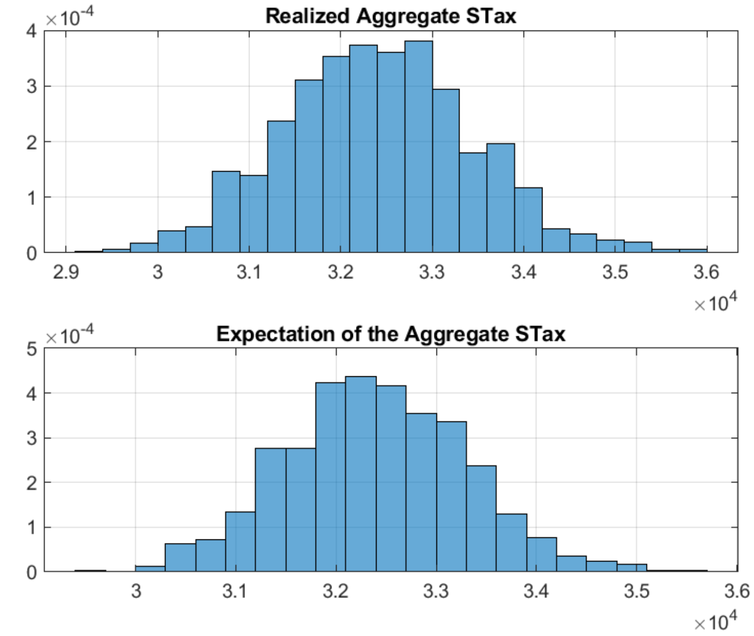

Figure 6: Realized and predicted aggregate STax for 1,000 replications.

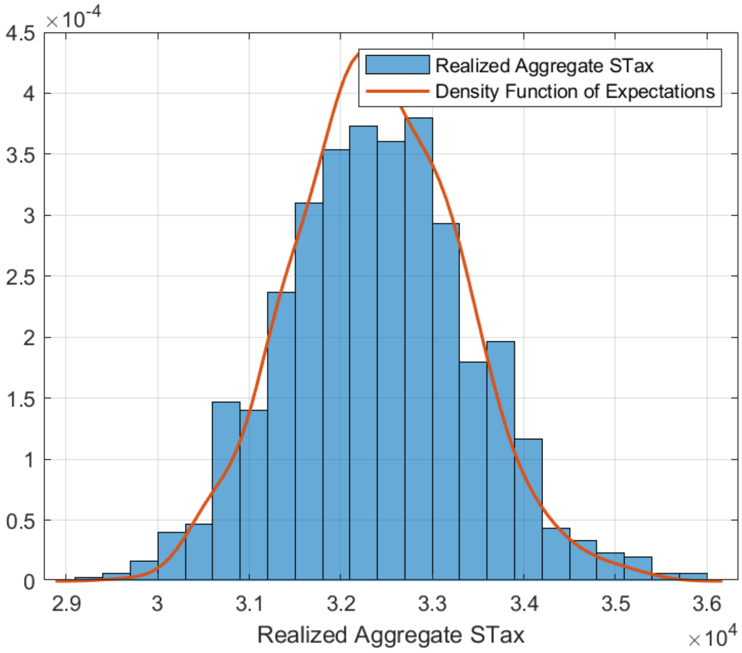

Figure 7: Distribution of realized aggregate STax and its probability distribution function of predicted values.

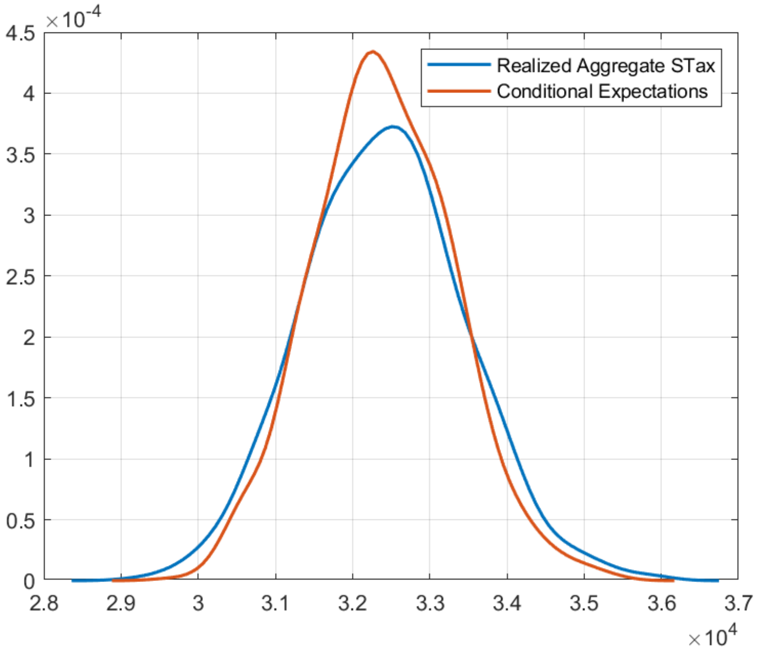

Figure 8: Empirical probability distribution function of realized and predicted aggregate STax.

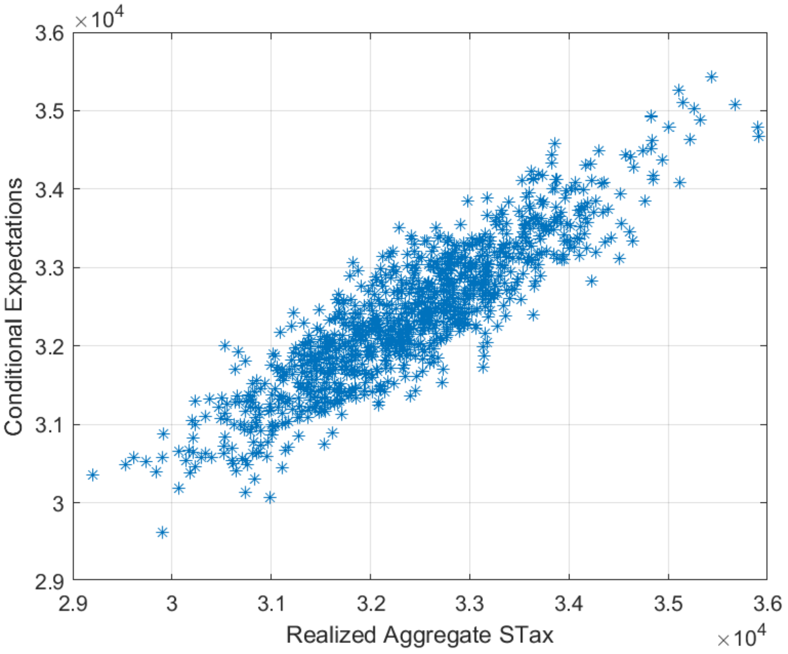

Figure 9: Scatter plot of realized and predicted aggregate STax.

Note that, $D_i = \widetilde{D}_i \quad \forall i : C_i = 1$,

$$M_i^* = \begin{cases} \mathbf{X}_{m,i}\beta_m + \varepsilon_{m,i} & \text{if } \widetilde{D}_i = 1 \\ 0 & \text{otherwise} \end{cases} \quad \forall i = 1, \ldots, n, \tag{26}$$

$$M_i = \begin{cases} \mathbf{X}_{m,i}\beta_m + \varepsilon_{m,i} & \text{if } C_i = 1 \text{ and } D_i = 1 \\ 0 & \text{otherwise} \end{cases} \quad \forall i : C_i = 1 \tag{27}$$

$$\begin{pmatrix} \varepsilon_{c,i} \\ \varepsilon_{d,i} \\ \varepsilon_{m,i} \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{28}$$

As a reminder, only the variables C_i, D_i, and M_i are observable and enter into the construction of the likelihood of the sample $S_n = \left\{ \{C_i\}_{i=1}^n \{D_i, M_i\}_{i:C_i=1} \right\}$. It is also important to note that the definition of the tax adjustment variable M_i is different from the one used previously, since it depends here on the (observable) detection D_i and not on the latent fraud variable $\widetilde{D}_i$.

Remark: *We assume that the effective tax adjustment M_i and detection D_i variables are only defined for the controlled firms, for which $C_i = 1$. We do not assume that these variables are equal to zero for uncontrolled firms, but censored (undefined) when $C_i = 0$.*

In order to better understand the construction of the log-likelihood of the complete system, we proceed in two steps. First, we will limit ourselves to the estimation of the parameters of the control and detection equations, which appear in the form of a *bi-probit model with censorship* or of a *nested probit model with dependence*. Second, we add the tax adjustment equation and complete the likelihood formula. It should be noted that in practice, this two-step decomposition is not necessary.

4.2 Estimation of the bi-probit model with censorship

Under the aforementioned assumptions, the control and detection decisions can be represented in the form of a nested Probit structure with dependence (see Figure 10).

It should be noted that while the fraud ($\widetilde{D}_i$) and control (C_i) decisions are assumed not to be nested (even if they are linked), this is not true for the detection (D_i) and the control (C_i) variables, which are necessarily nested events. This type of model is unusual, because on the one hand, nested models have often logit specifications, and on the other hand, dependence is rarely considered into nested models. This model can also be associated to a bi-Probit model with censorship on D_i for entities for which $C_i = 0$.

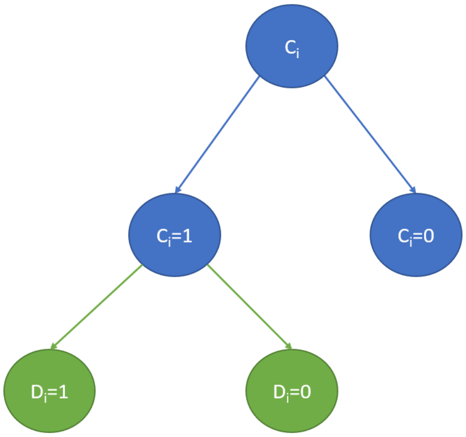

Figure 10: Nested Probit structure with dependence

The likelihood function associated to the sample of detected and controlled entities, $\widetilde{S}_n = \left\{ \{C_i\}_{i=1}^n \{D_i\}_{i:C_i=1} \right\}$ and used to estimate the vector of parameters $\theta = (\beta_c' \ \beta_d' \ \rho_{cd})'$ is written as follows:

$$\begin{aligned} \ell_n(\theta; C, D) &= \sum_{i=1}^n \ln\left(\Pr\left(C_i = 0\right)\right) \times \mathbf{1}_{(C_i=0)} \\ &+ \sum_{i=1}^n \ln\left(\Pr\left(D_i = 0 | C_i = 1\right) \times \Pr\left(C_i = 1\right)\right) \times \mathbf{1}_{(D_i=0)} \\ &+ \sum_{i=1}^n \ln\left(\Pr\left(D_i = 1 | C_i = 1\right) \times \Pr\left(C_i = 1\right)\right) \times \mathbf{1}_{(D_i=1)} \end{aligned} \tag{29}$$

The log-likelihood can equivalently take the form:

$$\begin{aligned} \ell_n(\theta; C, D) &= \sum_{i:C_i=0} \ln\left(\Pr\left(C_i = 0\right)\right) + \sum_{i:D_i=0} \ln\left(\Pr\left(\left(D_i = 0\right) \cap \left(C_i = 1\right)\right)\right) \\ &+ \sum_{i:D_i=1} \ln\left(\Pr\left(\left(D_i = 1\right) \cap \left(C_i = 1\right)\right)\right) \end{aligned} \tag{30}$$

Its elements are then made explicit as a function of the *cdf* of a bivariate normal distribution denoted $\Phi_2(u, v; \rho)$, such as:

$$\Phi_2(u; v; \rho) = \Pr\left(\left(U < u\right) \cap \left(V < v\right)\right) \tag{31}$$

where the vector $(U \ V)$ admits $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_2)$ as joint distribution, with $\mathbf{\Sigma}_2 = [1 \ \rho \ ; \ \rho \ 1]$. Note that, just as in the case of a probit, we normalize the variances to one.

***Remark**: If one wants to write* $\Pr((U<u)\cap(V>v))$*, in this case*

$$\Pr((U<u)\cap(V>v)) = \Pr((U<u)\cap(-V<-v)) = \Phi_2(u;-v;-\rho) \tag{32}$$

because passing from V *to* $-V$ *implies that cor* $(U,-V)=-\rho$.

In order to control for these sign changes, we introduce two new dichotomous variables taking values between 1 and -1, such that:

$$q_{c,i} = 2C_i - 1 = \begin{cases} 1 & \text{if } \; C_i = 1 \\ -1 & \text{if } \; C_i = 0 \end{cases} \tag{33}$$

$$q_{d,i} = 2D_i - 1 = \begin{cases} 1 & \text{if } \; D_i = 1 \\ -1 & \text{if } \; D_i = 0 \end{cases} \tag{34}$$

Hence, the first term of the log-likelihood becomes:

$$\Pr(C_i = 0) = \Pr\left(\frac{\varepsilon_{c,i}}{\sigma_c} < -\frac{\mathbf{X}_{c,i}\beta_c}{\sigma_c}\right) = \Phi\left(-\mathbf{X}_{c,i}\widetilde{\beta}_c\right) = \Phi\left(q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c\right) \quad \forall i : C_i = 0 \tag{35}$$

with $\widetilde{\beta}_c = \beta_c/\sigma_c$, as $q_{c,i} = -1$ for all uncontrolled firms (i.e., $C_i = 0$). Similarly:

$$\begin{aligned} \Pr((D_i=0)\cap(C_i=1)) &= \Pr\left(\left(\frac{\varepsilon_{d,i}}{\sigma_d} < -\frac{\mathbf{X}_{d,i}\beta_d}{\sigma_d}\right)\cap\left(\frac{\varepsilon_{c,i}}{\sigma_c} > -\frac{\mathbf{X}_{c,i}\beta_c}{\sigma_c}\right)\right) \\ &= \Pr\left(\left(\widetilde{\varepsilon}_{d,i} < -\mathbf{X}_{d,i}\widetilde{\beta}_d\right)\cap\left(-\widetilde{\varepsilon}_{c,i} < \mathbf{X}_{c,i}\widetilde{\beta}_c\right)\right) \\ &= \Pr\left(\left(\widetilde{\varepsilon}_{d,i} < q_{d,i}\mathbf{X}_{d,i}\widetilde{\beta}_d\right)\cap\left(-\widetilde{\varepsilon}_{c,i} < q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c\right)\right) \\ &= \Phi_2\left(q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c; q_{d,i}\mathbf{X}_{d,i}\widetilde{\beta}_d; -\rho_{cd}\right) \qquad \forall i : D_i = 0, C_i = 1 \end{aligned} \tag{36}$$

with $\widetilde{\beta}_d = \beta_d/\sigma_d$, $\widetilde{\varepsilon}_{c,i} = \varepsilon_{c,i}/\sigma_c$, and $\widetilde{\varepsilon}_{d,i} = \varepsilon_{d,i}/\sigma_d$. For all undetected but controlled entities $(D_i = 0, C_i = 1)$ we have $q_{c,i} = 1$ and $q_{d,i} = -1$.

The multiplication of the index $\mathbf{X}_{c,i}\widetilde{\beta}_c$ by $q_{c,i}$ does not allow us to reverse the sign, reason for which we are forced to change the sign of the correlation in the cdf of the normal distribution. Finally, concerning the third term, it can be written as following:

$$\begin{aligned} \Pr((D_i=1)\cap(C_i=1)) &= \Pr\left(\left(\frac{\varepsilon_{d,i}}{\sigma_d} > -\frac{\mathbf{X}_{d,i}\beta_d}{\sigma_d}\right)\cap\left(\frac{\varepsilon_{c,i}}{\sigma_c} > -\frac{\mathbf{X}_{c,i}\beta_c}{\sigma_c}\right)\right) \\ &= \Pr\left(\left(-\widetilde{\varepsilon}_{d,i} < \mathbf{X}_{d,i}\widetilde{\beta}_d\right)\cap\left(-\widetilde{\varepsilon}_{c,i} < \mathbf{X}_{c,i}\widetilde{\beta}_c\right)\right) \\ &= \Pr\left(\left(-\widetilde{\varepsilon}_{d,i} < q_{d,i}\mathbf{X}_{d,i}\widetilde{\beta}_d\right)\cap\left(-\widetilde{\varepsilon}_{c,i} < q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c\right)\right) \\ &= \Phi_2\left(q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c; q_{d,i}\mathbf{X}_{d,i}\widetilde{\beta}_d; \rho_{cd}\right) \qquad \forall i : D_i = 1, C_i = 1 \end{aligned} \tag{37}$$

because we have $q_{c,i} = 1$ and $q_{d,i} = 1$ for all controlled and detected individuals ($D_i = 1, C_i = 1$). Note that the correlation between the transformed variables $-\widetilde{\varepsilon}_{d,i}$ and $-\widetilde{\varepsilon}_{c,i}$ is equal to ρ.

The log-likelihood of the bi-Probit model with censorship associated with the control and detection decisions can be written as:

$$\begin{aligned}\ell_n(\theta; C, D) \quad = \quad & \sum_{i:C_i=0} \ln\left(\Phi\left(q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c\right)\right) + \sum_{i:D_i=0} \ln\left(\Phi_2\left(q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c; q_{d,i}\mathbf{X}_{d,i}\widetilde{\beta}_d; -\rho_{cd}\right)\right) \\ & + \sum_{i:D_i=1} \ln\left(\Phi_2\left(q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c; q_{d,i}\mathbf{X}_{d,i}\widetilde{\beta}_d; \rho_{cd}\right)\right) \end{aligned} \tag{38}$$

where $\Phi(u)$ represents the cdf of a standard normal distribution, and $\Phi_2(u; v; \rho)$ the cdf of a bivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_2)$, with $\mathbf{\Sigma}_2 = [1\ \rho\ ;\ \rho\ 1]$.[14]

The maximum likelihood estimator of the vector $\theta = (\beta_c'\ \beta_d'\ \rho_{cd})'$ can be hence defined as:

$$\widehat{\theta}_1 = \underset{\theta \in \Theta}{\arg\max}\ \ell_n(\theta; C, D) \tag{39}$$

This estimator will be mainly used as an initial condition for estimating the maximum likelihood of the complete system. For this, it remains now to include the tax adjustment equation in the model.

4.3 Estimation of the complete system by maximum likelihood

Let us consider now the complete model defined by equations (24) to (28). The objective is to estimate the full vector of $k_c + k_d + k_m + 6$ parameters corresponding to the three equations of control, detection, and tax adjustment, and to the parameters of the covariance matrix, denoted $\theta = (\beta_c'\ \beta_d'\ \beta_m'\ \rho_{cd}\ \rho_{cm}\ \rho_{dm}\ \sigma_m)'$. The structure of the complete model (see Figure 11) is similar to a model with a censorship on the tax adjustment M_i, similar to a Type II Tobit model introduced by Amemiya (1984). The censorship mechanism is represented by a bi-Probit model itself censored, or equivalently by a nested Probit with dependencies.

[14] In the absence of censorship, i.e. if the fraud variable $\widetilde{D}_i$ had been observed for all individuals, the log-likelihood would become:

$$\ell_n(C, D) = \sum_{i=1}^{n} \ln\left(\Phi_2\left(q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c; q_{d,i}\mathbf{X}_{d,i}\widetilde{\beta}_d; \rho_i\right)\right)$$

with $\rho_i = q_{c,i}q_{d,i}\rho$ a correlation term whose sign changes depending on the values of the variables D_i and C_i observed for each individual.

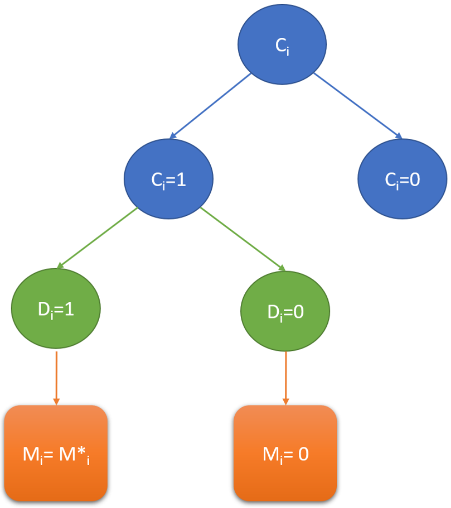

Figure 11: Design of the complete model

***Remark**: Considering a censorship mechanism based on the observation of $D_i = 1$ (Type II Tobit) and not directly on the sign of M_i^* (Type I Tobit) means that we can observe also effective amounts of negative recovery or tax adjustment.*

$$\textit{Type II Tobit:} \;\; M_i = \begin{cases} \mathbf{X}_{m,i}\beta_m + \varepsilon_{m,i} & \textit{if} \;\; D_i = 1 \\ 0 & \textit{otherwise} \end{cases} \qquad \forall i : C_i = 1 \tag{40}$$

$$\textit{Type I Tobit:} \;\; M_i = \begin{cases} \mathbf{X}_{m,i}\beta_m + \varepsilon_{m,i} & \textit{if} \;\; M_i^* > 0 \\ 0 & \textit{otherwise} \end{cases} \qquad \forall i : C_i = 1 \tag{41}$$

The log-likelihood of the sample $S_n = \left\{ \{C_i\}_{i=1}^n \{D_i, M_i\}_{i:C_i=1} \right\}$ associated with the control, detection and adjustment events can be written as follows:

$$\begin{aligned} \ell_n(\theta; C, D, M) &= \sum_{i:C_i=0} \ln\left(\Pr\left(C_i = 0\right)\right) + \sum_{i:M_i=0} \ln\left(\Pr\left(\left(D_i = 0\right) \cap \left(C_i = 1\right)\right)\right) \\ &\quad + \sum_{i:M_i \neq 0} \ln\left(f_{M|C,D}\left(M_i^* | D_i^* > 0, C_i^* > 0\right) \times \Pr\left(\left(D_i = 1\right) \cap \left(C_i = 1\right)\right)\right) \end{aligned} \tag{42}$$

The first two terms of the likelihood, i.e., $\ln\left(\Pr\left(C_i = 0\right)\right)$ and $\ln\left(\Pr\left(\left(D_i = 0\right) \cup \left(C_i = 1\right)\right)\right)$, are identical to those already presented previously.

It remains to characterize the conditional density of M_i knowing that the entity has been controlled and detected as fraudulent, i.e., $D_i^* > 0$ and $C_i^* > 0$, noted $f_{M|C,D}(u, v)$. The latter has the following form:

$$f_{M|C,D}\left(M_i^* | D_i^* > 0, C_i^* > 0\right) = \Pr\left(\left(D_i = 1\right) \cap \left(C_i = 1\right)\right)^{-1} \int_0^\infty \int_0^\infty f_{C,D,M}\left(C_i^*, D_i^*, M_i^*\right) dC_i^* dD_i^*$$

with $f_{C,D,M}(u,v,w)$ the *pdf* of the joint distribution of the triplet (C_i^*, D_i^*, M_i^*). Thus, the last term of the likelihood of equation (43) becomes:

$$f_{M|C,D}(M_i^*|D_i^*>0, C_i^*>0) \times \Pr((D_i=1)\cap(C_i=1)) = \int_0^\infty \int_0^\infty f_{C,D,M}(C_i^*, D_i^*, M_i^*)\, dC_i^* dD_i^*$$

The problem here is that C_i^* and D_i^* are not observable. To solve the problem, Amemiya (1984) proposes to reverse the conditioning issue so as to work on the distribution of the observed variables C_i^* and D_i^* conditional on the observation of $M_i^* = m_i^*$, as follows:

$$\begin{aligned}\int_0^\infty \int_0^\infty f_{C,D,M}(C_i^*, D_i^*, M_i^*)\, dC_i^* dD_i^* &= f_M(M_i^*) \int_0^\infty \int_0^\infty f_{C,D|M}(C_i^*, D_i^*|M_i^*)\, dC_i^* dD_i^* \\ &= f_M(M_i^*) \Pr(C_i^*>0, D_i^*>0|M_i^*) \qquad (43)\end{aligned}$$

where $f_M(u)$ denotes the marginal distribution of variable M_i^* and $f_{C,D|M}(u,v)$ is the *pdf* of the conditional distribution of the couple (C_i^*, D_i^*), knowing that $M_i^* = m_i^*$. Under the assumption that the vector (C_i^*, D_i^*, M_i^*) is normally distributed, we know that the marginal and conditional distributions are also normal with:

$$f_M(M_i^*) = \frac{1}{\sigma_m}\phi\left(\frac{M_i^* - \mathbf{X}_{m,i}\beta_m}{\sigma_m}\right) \qquad (44)$$

with $\phi(.)$ the *pdf* of a standard Normal distribution. Concerning the conditional distribution $f_{C,D|M}(u,v)$, we can show that:

$$\begin{pmatrix} C_i^* \\ D_i^* \end{pmatrix} |_{M_i^*=m_i^*} \sim \mathcal{N}\left(\mu_{CD|M,i}, \mathbf{\Sigma}_{CD|M}\right) \qquad (45)$$

$$\mathbf{\Sigma}_{CD|M} = \mathbf{\Sigma}_{CD} - \frac{1}{\sigma_m^2}\mathbf{\Sigma}_{CD,M}\mathbf{\Sigma}'_{CD,M} \qquad (46)$$

$$\mu_{CD|M,i} = \mu_{CD,i} + \frac{1}{\sigma_m^2}\mathbf{\Sigma}_{CD,M}(M_i^* - \mathbf{X}_{m,i}\beta_m) \qquad (47)$$

where the vectors $\mu_{CD,i}$, $\mathbf{\Sigma}_{CD,M}$ and $\mathbf{\Sigma}_{CD}$ are defined as:

$$\mu_{CD,i} = \begin{pmatrix} \mathbf{X}_{c,i}\beta_c \\ \mathbf{X}_{d,i}\beta_d \end{pmatrix} \quad \mathbf{\Sigma}_{CD} = \begin{pmatrix} \sigma_c^2 & \sigma_{cd} \\ \sigma_{cd} & \sigma_d^2 \end{pmatrix} \quad \mathbf{\Sigma}_{CD,M} = \begin{pmatrix} \sigma_{cm} \\ \sigma_{dm} \end{pmatrix} \qquad (48)$$

Hence, if we denote by $\Phi_2\left(u; v; \mathbf{\Sigma}_{CD|M}\right)$ the *cdf* of the bivariate normal distribution $\mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{CD|M}\right)$, we show immediately that:

$$\begin{aligned} f_M(M_i^*) \Pr(C_i^*>0, D_i^*>0|m_i^*) &= f_M(M_i^*) \Pr((\widetilde{\varepsilon}_{c,i} > -\mu_{c,i}) \cap (\widetilde{\varepsilon}_{d,i} > -\mu_{d,i}) > 0|M_i^*) \\ &= f_M(M_i^*) \Pr((-\widetilde{\varepsilon}_{c,i} < \mu_{c,i}) \cap (-\widetilde{\varepsilon}_{d,i} < \mu_{d,i})\,|M_i^*) \\ &= \frac{1}{\sigma_m}\phi\left(\frac{M_i^* - \mathbf{X}_{m,i}\beta_m}{\sigma_m}\right)\Phi_2\left(\mu_{c,i}; \mu_{d,i}; \mathbf{\Sigma}_{CD|M}\right) \qquad (49)\end{aligned}$$

Finally, we can deduce the log-likelihood of the complete system.

Definition 2 *The log-likelihood of the full model associated with the control, detection and recovery/adjustment decisions is written:*

$$\begin{aligned}\ell_n(\theta; C, D, M) &= \sum_{i:C_i=0} \ln\left(\Phi\left(q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c\right)\right) + \sum_{i:M_i=0} \ln\left(\Phi_2\left(q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c; q_{d,i}\mathbf{X}_{d,i}\widetilde{\beta}_d; -\rho_{cd}\right)\right) \\ &\quad + \sum_{i:M_i\neq 0} \ln\left(\frac{1}{\sigma_m}\phi\left(\frac{M_i^* - \mathbf{X}_{m,i}\beta_m}{\sigma_m}\right)\Phi_2\left(\mu_{c,i}; \mu_{d,i}; \mathbf{\Sigma}_{CD|M}\right)\right) \qquad (50)\end{aligned}$$

where $\Phi(u)$ is the cdf of the standard normal distribution, $\Phi_2(u; v; \rho)$ the cdf of the bivariate normal $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_2)$ with $\mathbf{\Sigma}_2 = [1\ \rho\ ;\ \rho\ 1]$, and $\Phi_2\left(u; v; \widetilde{\mathbf{\Sigma}}_2\right)$ the cdf of the bivariate normal $\mathcal{N}\left(\mathbf{0}, \widetilde{\mathbf{\Sigma}}_2\right)$.

Note that the coefficients σ_c and σ_d are not identifiable, but that the variance σ_m^2 of the error term of the adjustment equation is identifiable. The maximum likelihood estimator of the vector $\theta = (\beta_c'\ \beta_d'\ \beta_m'\ \rho_{cd}\ \rho_{cm}\ \rho_{dm}\ \sigma_m)'$ is then defined by:

$$\widehat{\theta} = \underset{\theta\in\Theta}{\arg\max}\ \ell_n(\theta; C, D, M). \qquad (51)$$

4.4 Validation of the Maximum Likelihood Estimation

Monte Carlo simulations are used in order to show the validity of the log-likelihood function (equation 50). Thus, we take the data generating process presented in Section 3, and simulate from the equations (24) to (28) the unobservable variables (C_i^*, D_i^*, M_i^*), as well as the observable ones (C_i, D_i, M_i), and this for a large number of firms $(n = 1{,}000{,}000)$.

As a reminder, we consider $k_c = 4$ explanatory variables for the equation of the latent variable C_i^*, $k_d = 2$ explanatory variables for the equation of the latent variable D_i^* and $k_m = 1$ explanatory variables for the equation of the latent variable M_i^*. We assume that the explanatory variables $\mathbf{X}_{c,i}$, $\mathbf{X}_{d,i}$ and $\mathbf{X}_{m,i}$ are i.i.d. for all $i = 1, \ldots, n$, and verify $\mathbf{X} = (\mathbf{X}_{c,i} : \mathbf{X}_{d,i} : \mathbf{X}_{m,i}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{k_c+k_d+k_m})$. The parameters have been set so that the conditional probability of control for fraudsters is equal to 5.30%, while the conditional probability of control for non-fraudsters is equal to 4.61%. This configuration implies also that the average ratio of effective adjustments on the global $STax$ is equal to 6%. As for the parameters, they have been fixed to the following values:

$$\beta_c' = \begin{pmatrix} -9.27 & 1 & -2 & 3 & -4 \end{pmatrix}$$

$$\beta_d' = \begin{pmatrix} -10.13 & 5 & -6 \end{pmatrix}$$

$$\beta_m' = \begin{pmatrix} 33.75 & 7 \end{pmatrix}$$

The covariance matrix of the error terms has been calibrated as follows (see Section 3 for more details):

$$\mathbf{\Sigma} = \begin{pmatrix} 2.0000 & 1.6000 & 0.9487 \\ 1.6000 & 2.0000 & 1.5811 \\ 0.9487 & 1.5811 & 5.0000 \end{pmatrix}$$

In addition, we recall that only the ratios β_c/σ_c and β_d/σ_d are identifiable, i.e., can be estimated. The variances σ_c^2 and σ_d^2 of the control and detection equations are not identifiable. By convention, they are fixed to one. On the other hand, the correlations and the variance σ_m^2 of the error term corresponding to the adjustment equation are identifiable and can be estimated.

As we saw, the log-likelihood of the full model (Equation 50) is strongly nonlinear in the parameters and not globally concave. Therefore, we may deal with problems of numerical convergence in the optimization process (See Lee, 2020). In order to limit these issues, we propose a 3-steps procedure for setting the initial conditions.

Step 1 Estimation of the two univariate Probit models associated with the control decision (on all firms) and the detection decision (on controlled firms only), without taking into account the selection bias.

Step 2 Maximization of the log-likelihood of the bi-Probit model with censorship on the control and detection decisions, considering $\widehat{\beta}_c$ and $\widehat{\beta}_d$, as initial conditions. We note $\widehat{\theta}_1$ the ML estimator thus obtained.

Step 3 Maximization of the log-likelihood of the complete model (Equation 50), which gives the estimates of the parameters $\beta_c', \beta_d', \beta_m', \rho_{cd}, \rho_{cm}, \rho_{dm}$ et σ_m, considering $\widehat{\theta}_1$ for the initial condition.

The MLE results are reported in Table (2) for the nested Probit model and in Table (3) for the full model. Results show that the ML estimation approach allows not only to estimate the correlations between shocks, but also to obtain unbiased estimates of the model parameters.[15]

[15] Appendix D shows that in the case of non-linear models (such as the probit model used for the detection equation) taking into account the selection bias via Heckman's method is not valid (Heckman, 1976, 1979). Not only it does not allow to take into account different correlations, but it also leads to biased parameter estimates.

Table 2: Nested Probit model estimated by maximum likelihood

This table reports the estimated coefficients by maximum likelihood issued of the nested probit model and their respective true value (ratios).

		Estimated	True
Control Equation	β_c'/σ_c	-6.5175	-6.5549
		0.7049	0.7071
		-1.4062	-1.4142
		2.1029	2.1213
		-2.8079	-2.8284
Detection Equation	β_d'/σ_d	-7.1433	-7.1630
		3.5263	3.5355
		-4.2571	-4.2426
Correlation	ρ_{cd}	0.7683	0.8000

Table 3: Full model estimated by maximum likelihood

This table reports coefficient estimates by maximum likelihood issued of the complete model and their respective true value (ratios).

		Estimated	True
Control Equation	β_c'/σ_c	-6.5172	-6.5549
		0.7049	0.7071
		-1.4059	-1.4142
		2.1029	2.1213
		-2.8079	-2.8284
Detection Equation	β_d'/σ_d	-7.1433	-7.1630
		3.5034	3.5355
		-4.2350	-4.2426
Adjustment Equation	β_m'	33.6295	33.7500
		7.0433	7.0000
Correlation	ρ_{cd}	0.7738	0.8000
	ρ_{cm}	0.3004	0.3000
	ρ_{dm}	0.5626	0.5000
Error term std dev. of the adjustment equation	σ_m	2.2651	2.2361

5 Empirical application

This section presents estimation results obtained using real data provided by the MSA. We first define robust initial conditions, then implement an approach to identify the best configurations of the three model equations, and finally calculate the tax shortfall. Several configurations and specifications of the full model were considered and tested.

5.1 Data

In addition to the methodological contribution presented above, an empirical application is performed on real data collected by the French agricultural social security agency (MSA) as a result of regular controls carried out on firms in the agricultural system. More precisely, in order to combat illegal activities, MSA collects data systematically from their beneficiaries and organizes regular controls on a subsample of their taxpayers. The database matches all data retained for the scope of the accounting control for the years 2014 to 2016. The data were collected by different local control entity funds and consolidated at the national level by the Department of Statistics, Studies and Funds of the control entity. The main analysis covers the year 2014, and the other two years are used for the robustness check analysis. Table 4 presents simple descriptive statistics for the three years to better understand the scope of the study.

Table 4: Descriptive statistics

	2014	2015	2016
Geographic coverage	Metropolitan France	Metropolitan France	Metropolitan France
Total number of firms	187 646	185 808	183 093
Total number of controlled firms	5 468	4 925	4 055
Observed frequency of control	2.91%	2.65%	2.21%
Total number of detected fraudulent firms	1 614	1 496	1 172
Observed frequency of fraud (among all firms)	0.86%	0.81%	0.64%
Observed frequency of fraud (among the controlled firms)	29.52%	30.38%	28.90%
Percentage of the applied tax adjustment with respect to the total social contribution amount recorded for controlled firms	2.10%	2.10%	2.02%

Considering the reference year of 2014, the database regroups data for 187 646 firms and contains 34 explanatory variables to run the three equations of the model. The data dictionary is available in Table E1. We observe that 2.91% of the total taxpayers made the object of a control. Fraud was detected for 29.52% of the firms audited and corrective actions were applied. This proportion should not be generalized to the entire sample, since, as already mentioned, the decision of control is not random but based on internal considerations.[16] The amount of detected fraud represents 2.10% of the total amount of social security contributions recorded for the firms subject to an audit. For the uncontrolled firms (the remaining 97.09% of the sample), an estimation of the tax shortfall (i.e. the tax adjustments that could have been imposed on the defrauding firms if they had been effectively controlled and detected by the control authority) is carried out by using our econometric model. For confidentiality reasons, all results are presented in relative terms, i.e. expressed with respect to the total social contribution amount recorded for uncontrolled firms.

5.2 Looking for optimal specifications

As explained in Section 4.4, the log-likelihood of the complete model is strongly non-linear in the parameters and not globally concave. In order to limit numerical convergence problems in the optimization of the log-likelihood, we rely on the 3-steps procedure for setting robust initial conditions. Moreover, estimating the complete model by considering 34 explanatory variables does not spare us from estimation issues.[17] One solution is to adopt parsimonious alternative specifications, with a reduced number of explanatory variables, while remaining close to the performance of the model that includes all available information. Therefore, we reduce progressively the number of explanatory variables to four, five or six variables, for the three equations. The search of the best configurations is thus done based both on business expert knowledge and statistical criteria, by following the next steps:

1. $C_{34}^{k_1}$ and $C_{34}^{k_2}$ standard Probit models are estimated (i.e., all combinations of k_1 variables of the control equation and k_2 for the detection equation among the 34 variables initially proposed) for the control and detection equations, respectively.[18] For each estimated model, we calculate the area under the ROC curve (ROC-AUC) as a performance indicator and select the configuration that maximizes it.

2. For the adjustment equation, we estimate $C_{34}^{k_3}$ linear regression models, calculate the

[16] The detected fraudulent firms represent only 0.86% of the total number of firms under analysis.

[17] Considering the entire set of explanatory variables, we end up with a total of 109 parameters to estimate, since we have to take also into account the constant in each equation, the three correlation coefficients, $\rho_{cd}, \rho_{cm}, \rho_{dm}$, as well as the standard deviation of the third equation, σ_m).

[18] As a reminder, $C_n^k = \frac{n!}{k!(n-k)!}$.

(adjusted) coefficient of determination (R^2 and $\bar{R}^2$) as a measure of performance and keep the specification that maximizes it.

The best specifications of the three equations (for a different number of explanatory variables) are presented in Table 5.

Table 5: Best configurations retained for each equation of the model

1st Equation	2nd Equation	3rd Equation	
AUC & Variables	AUC & Variables	R2 & Variables	R2 adjusted & Variables
Best 4			
71.24%	62.62%	30.86%	30.69%
’coti_tot’	’coti_tot’	’coti_tot’	’coti_tot’
’saison3cl_3’	’LCOTITOT_REG8’	’segment1’	’segment1’
’LCOTITOT_REG3’	’LCOTITOT_REG10’	’LCOTITOT_REG1’	’LCOTITOT_REG1’
’LCOTITOT_REG_sq8’	’LCOTITOT_REG_sq8’	’LCOTITOT_REG_sq1’	’LCOTITOT_REG_sq1’
Best 5			
71.58%	63.19%	31.84%	31.63%
’coti_tot’	’coti_tot’	’ancentr’	’ancentr’
’nbsal41’	’LCOTITOT_REG8’	’coti_tot’	’coti_tot’
’saison3cl_3’	’LCOTITOT_REG10’	’segment1’	’segment1’
’LCOTITOT_REG3’	’LCOTITOT_REG_sq3’	’LCOTITOT_REG1’	’LCOTITOT_REG1’
’LCOTITOT_REG8’	’LCOTITOT_REG_sq8’	’LCOTITOT_REG_sq1’	’LCOTITOT_REG_sq1’
Best 6			
71.91%	63.64%	32.81%	32.56%
’coti_tot’	’cotietat_cotitot’	’ancentr’	’ancentr’
’nbsal41’	’nbsal44’	’coti_tot’	’coti_tot’
’saison3cl_3’	’LCOTITOT_REG1’	’segment1’	’segment1’
’LCOTITOT_REG8’	’LCOTITOT_REG4’	’segment3’	’segment3’
’LCOTITOT_REG_sq1’	’LCOTITOT_REG8’	’LCOTITOT_REG1’	’LCOTITOT_REG1’
’LCOTITOT_REG_sq2’	’LCOTITOT_REG_sq8’	’LCOTITOT_REG_sq1’	’LCOTITOT_REG_sq1’
All variables			
75.07%	66.08%	35.40%	34.01%

5.3 Results on the estimated tax shortfall

All combinations $\{k_1,\ k_2,\ k_3\}$ were estimated and the results on the estimated tax shortfall (with respect to the total amount of contributions recorded for uncontrolled firms) are reported on Table 6.[19]

[19] We only report the configurations that converged. The models are estimated with Matlab 2022a.

Table 6: Estimation of $STax$

Configuration	**% Stax (+/-)**	**% Stax (+)**	**Log-Likelihood**
ALL (k1=k2=k3=34)	**7.43%**	**7.43%**	**-43385.84**
k1=k2=k3=4	2.12%	4.19%	-45311.55
k1=4, k2=4, k3=5	1.94%	4.43%	-45300.87
k1=4, k2=4, k3=6	1.98%	4.37%	-45289.10
k1=4, k2=5, k3=4	5.74%	8.23%	-45304.08
k1=4, k2=5, k3=5	5.09%	7.70%	-45293.30
k1=4, k2=6, k3=4	3.51%	5.53%	-45260.89
k1=4, k2=6, k3=5	3.07%	5.41%	-45250.34
k1=4, k2=6, k3=6	3.21%	5.53%	-45238.38
k1=5, k2=4, k3=4	2.96%	5.16%	-45036.97
k1=5, k2=4, k3=6	2.87%	5.33%	-45014.50
k1=5, k2=6, k3=4	2.40%	4.18%	-44989.83
k1=5, k2=6, k3=6	2.19%	4.37%	-44967.33
k1=6, k2=5, k3=4	3.43%	5.56%	-44839.83
k1=6, k2=6, k3=5	4.10%	6.38%	-44782.72
k1=6, k2=6, k3=6	**4.26%**	**6.49%**	**-44770.82**

The tax shortfall was calculated both on the entire set of individual tax shortfalls (which can take both positive and negative values), as well as only on the positive individual tax shortfalls. The actual database used for the estimation of the third equation contains only the observed positive amounts of adjustment corresponding to the detected fraud, but in practice the control entity might record also negative amounts due to the reimbursement of overpayments observed after certain regularizations.[20] The results are quite consistent for all the configurations tested and in line with the expectations of the control authority. The estimated $STax$ represents between 4.18% and 8.23% of the total social security contributions paid by taxpayers.

We use the reported Log-likelihood to compute also Likelihood Ratio (LR) tests since we face nested models.[21] These tests lead to reject systematically the null hypothesis, meaning than we always prefer the unconstrained model, the one with more parameters. It leads us to

[20] Our econometric model is able to deal with both positive and positive/negative amounts. A robustness check analysis is done on data including also negative adjustments.

[21] For instance, the model with $k_1 = k_2 = k_3 = 4$ is a constrained version of the model with $k_1 = k_2 = 4, k_3 = 5$, since it is enough to set to 0 the coefficient associated to the explanatory variable "ancientr" in order to go from the latter to the former.

select the last configuration with $k_1 = k_2 = k_3 = 6$, that we will keep for the robustness check as well.

5.4 Robustness check

To check the robustness of our results, we carry out several exercises described below:

1. We extend the analysis for 2015 and 2016. In a first setting, we predict $STax$ for 2015 and 2016 by keeping constant the estimated parameters of 2014 and applying the formulas on the new datasets of 2015 and 2016. In a second setting, we re-estimate the preferred configuration of 2014 ($k_1 = k_2 = k_3 = 6$) on new data of 2015 and 2016. We only obtain convergence and consistent results for 2015. Table 7 reports the main results of these experimental frameworks. We observe that the results are robust, the estimated and predicted $STax$ are in line with the results of the first period.

Table 7: Estimation and prediction of $STax$ on different periods

	Configuration	% Stax (+/-)	% Stax (+)
First setting			
Predicted $STax$ on 2015, 2016 using 2014 parameters estimates			
2014	k1=6, k2=6, k3=6	4.26%	6.49%
2015	k1=6, k2=6, k3=6	4.07%	6.42%
2016	k1=6, k2=6, k3=6	3.87%	6.25%
Second setting			
Estimated $STax$ on 2015, 2016 using 2014 configurations			
2014	k1=6, k2=6, k3=6	4.26%	6.49%
2015	k1=6, k2=6, k3=6	4.38%	5.37%
2016	k1=6, k2=6, k3=6	NA	NA

2. Estimation and calculation of $STax$ on a new database, including negative values for the amount of the adjustments in the third equation. Based on this new dataset for the year 2014, we look for the best configurations and we estimate all the possible configurations. The results (for the models that converged) are presented in Table 8. Results are quite consistent with those obtained previously.

Table 8: Estimation of $STax$

Configuration	% Stax (+/-)	% Stax (+)	Log-Likelihood
ALL (k1=k2=k3=34)	3.24%	3.24%	-48112.87
k1=k2=k3=4	0.87%	3.45%	-49926.25
k1=4, k2=4, k3=5	1.13%	4.18%	-49920.75
k1=4, k2=4, k3=6	1.89%	4.63%	-49906.35
k1=4, k2=5, k3=4	4.88%	7.58%	-49918.92
k1=4, k2=5, k3=5	4.26%	7.28%	-49913.24
k1=4, k2=5, k3=6	4.50%	7.51%	-49899.06
k1=4, k2=6, k3=5	2.11%	4.80%	-49861.61
k1=4, k2=6, k3=6	1.91%	4.69%	-49847.08
k1=5, k2=4, k3=4	2.35%	5.02%	-49615.04
k1=5, k2=4, k3=5	2.42%	5.51%	-49609.59
k1=5, k2=4, k3=6	3.44%	6.17%	-49595.14
k1=5, k2=5, k3=4	4.48%	7.15%	-49607.16
k1=5, k2=5, k3=5	4.24%	7.18%	-49601.38
k1=5, k2=5, k3=6	5.08%	7.78%	-49587.11
k1=5, k2=6, k3=4	1.64%	3.80%	-49560.61
k1=5, k2=6, k3=5	1.44%	4.27%	-49555.04
k1=5, k2=6, k3=6	1.49%	4.28%	-49540.51
k1=6, k2=4, k3=4	1.80%	4.36%	-49416.93
k1=6, k2=4, k3=5	2.13%	5.00%	-49411.26
k1=6, k2=4, k3=6	2.93%	5.54%	-49396.89
k1=6, k2=5, k3=4	2.24%	4.80%	-49412.29
k1=6, k2=5, k3=5	2.91%	5.56%	-49406.48
k1=6, k2=5, k3=6	2.19%	5.22%	-49392.70
k1=6, k2=6, k3=6	3.13%	5.84%	-49338.30

6 Conclusion

The main goal of this paper is to estimate the tax shortfall defined as the potential sum of the tax adjustments that could have been imposed on firms having defrauded (or made erroneous social declarations), if they had been effectively checked, whereas they were not in reality. To this end, we define first the shortfall in tax revenue from a statistical point of view, and then we validate the methods used to estimate such an amount. To highlight the tractability of the theoretical approach proposed, we use Monte Carlo simulations before relying on real data.

An empirical application is finally performed on real data for the estimation of the tax shortfall on two different periods. An approach based both on statistical indicators and expert knowledge is designed to identify the best configurations of the model, and finally the tax shortfall is calculated for each of the configurations retained. The results show that the estimated $STax$ represents between 4.18% and 8.23% of the total social security contributions actually paid by taxpayers. The results are in line with the expectations of the control body and the main conclusions of the study reinforced by a robustness check analysis. The analysis was also extended to the periods of 2015 and 2016.

References

Amemiya, T. (1984): "Tobit models: A survey," *Journal of Econometrics*, 24(1), 3–61. 3, 9, 26, 28

Baesens, B., V. Van Vlasselaer, and W. Verbeke (2015): *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection*. John Wiley & Sons. 4

Banulescu-Radu, D., and M. Yankol-Schalck (2021): "Fraud detection in the era of Machine Learning: a household insurance case," Discussion paper, Orleans Economics Laboratory/Laboratoire d'Economie d'Orleans. 4

Begier, M. H., and M. A. Hamdan (1971): "Correlation in a Bivariate Normal Distribution with Truncation in Both Variables," *Australian Journal of Statistics*, 13(2), 77–82. 14

Chung, J. H., and K. G. Goulias (1995): "Sample selection bias with multiple selection rules: application with residential relocation, attrition, and activity participation in Puget sound transportation panel," *Transportation Research Record*, 1493, 128–135. 47

Dyer, D. D. (1973): "On Moments Estimation of the Parameters of a Truncated Bivariate Normal Distribution," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22(3), 287–291. 14

Greene, W. (2006): "A General Approach to Incorporating Selectivity in a Model," . 47, 48

Heckman, J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," in *Annals of Economic and Social Measurement*, ed. by S. V. Berg, vol. 5, pp. 475–492. National Bureau of Economic Research. 15, 30, 46

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153–161. 15, 30, 46

Horrace, W. C. (2005): "Some results on the multivariate truncated normal distribution," *Journal of Multivariate Analysis*, 94(1), 209–221. 14

Kan, R., and C. Robotti (2017): "On Moments of Folded and Truncated Multivariate Normal Distributions," *Journal of Computational and Graphical Statistics*, 26(4), 930–934. 14, 43, 45

Khatri, C. G., and M. C. Jaiswal (1963): "Estimation of Parameters of a Truncated Bivariate Normal Distribution," *Journal of the American Statistical Association*, 58(302), 519–526. 14

Lee, S. C. (2020): "Moments calculation for truncated multivariate normal in nonlinear generalized mixed models," *Communications for Statistical Applications and Methods*, 27(3), 377–383. 30

Manjunath, B., and S. Wilhelm (2010): "tmvtnorm: A Package for the Truncated Multivariate Normal Distribution," *The R Journal*, 2. 14, 43, 45

Manjunath, B., and S. Wilhelm (2012): "Moments Calculation For the Doubly Truncated Multivariate Normal Density," *SSRN Electronic Journal*. 3, 14, 15, 43

Muthén, B. (1990): “Moments of the censored and truncated bivariate normal distribution,” *British Journal of Mathematical and Statistical Psychology*, 43(1), 131–143. 14, 44

Rosenbaum, S. (1961): “Moments of a Truncated Bivariate Normal Distribution,” *Journal of the Royal Statistical Society. Series B (Methodological)*, 23(2), 405–408. 14, 44

Shah, S. M., and N. T. Parikh (1964): “Moments of Single and Doubly Truncated Standard Bivariate Normal Distribution,” *Vidya*, 7, 82–91. 14

Van Vlasselaer, V., T. Eliassi-Rad, L. Akoglu, M. Snoeck, and B. Baesens (2017): “Gotcha! Network-based fraud detection for social security fraud,” *Management Science*, 63(9), 3090–3110. 4

West, J., and M. Bhattacharya (2016): “Intelligent financial fraud detection: a comprehensive review,” *Computers & security*, 57, 47–66. 4

Appendix A STax moments based on the informational structure hypothesis

Case 1: Complete information.

Once the control C_i and fraud $\widetilde{D}_i$ decisions are observed for all firms, the aggregate $STax$ can be deduced as follows:

$$STax = \sum_{i:(C_i=0)\cap(\widetilde{D}_i=1)} \underbrace{M_i^*}_{\text{r.v.}} = \sum_{i:C_i=0} \underbrace{M_i^*}_{\text{r.v.}} \times \mathbf{1}_{(\widetilde{D}_i=1)} = \sum_{i=1}^{n} \underbrace{M_i^*}_{\text{r.v.}} \times \mathbf{1}_{(C_i=0)} \times \mathbf{1}_{(\widetilde{D}_i=1)} \tag{A1}$$

where the variable M_i^* is random. Under the Assumptions A1-A5 and the data generating process (DGP) described by equations 1-9, the conditional moments of the aggregate $STax$ verifies:

$$\mathbb{E}_X\left(STax\right) = \sum_{i:(C_i=0)\cap(\widetilde{D}_i=1)} \mathbb{E}_X\left(M_i^*|\left(C_i=0\right)\cap\left(\widetilde{D}_i=1\right)\right) \tag{A2}$$

$$\mathbb{V}_X\left(STax\right) = \sum_{i:(C_i=0)\cap(\widetilde{D}_i=1)} \mathbb{V}_X\left(M_i^*|\left(C_i=0\right)\cap\left(\widetilde{D}_i=1\right)\right) \tag{A3}$$

where the expressions of the conditional moments $\mathbb{E}_X\left(M_i^*|\left(C_i=0\right)\cap\left(\widetilde{D}_i=1\right)\right)$ and $\mathbb{V}_X\left(M_i^*|\left(C_i=0\right)\cap\left(\widetilde{D}_i=1\right)\right)$ are similar to those reported in Equations (13) and (16).

Case 2: Observed control and unobserved fraud decisions.

Once the control decision C_i are observed for all firms, the aggregate $STax$ is defined by the following formula:

$$STax = \sum_{i:C_i=0} \underbrace{M_i^*}_{\text{r.v.}} \times \underbrace{\mathbf{1}_{(\widetilde{D}_i=1)}}_{\text{r.v.}} = \sum_{i=1}^{n} \underbrace{M_i^*}_{\text{r.v.}} \times \mathbf{1}_{(C_i=0)} \times \underbrace{\mathbf{1}_{(\widetilde{D}_i=1)}}_{\text{r.v.}} \tag{A4}$$

where the variables M_i^* and $\widetilde{D}_i$ are random. Under the assumptions A1-A5 and the DGP described by equations 1-9, the conditional moments of the aggregate $STax$ verifies:

$$\mathbb{E}_X\left(STax\right) = \sum_{i:C_i=0} \mathbb{E}_X(M_i^*|\left(C_i=0\right)\cap(\widetilde{D}_i=1)) \times \Pr(\widetilde{D}_i=1|C_i=0) \tag{A5}$$

$$\mathbb{V}_X\left(STax\right) = \sum_{i:C_i=0} \mathbb{V}_X(M_i^*|\left(C_i=0\right)\cap(\widetilde{D}_i=1)) \times \Pr(\widetilde{D}_i=1|C_i=0) \tag{A6}$$

Case 3: Unobserved control and fraud decisions.

If control decisions C_i and fraud decisions $\widetilde{D}_i$ are not observed, the aggregate $STax$ is defined as follows:

$$STax = \sum_{i:C_i=0} \underbrace{M_i^*}_{\text{r.v.}} \times \underbrace{\mathbf{1}_{(\widetilde{D}_i=1)}}_{\text{r.v.}} = \sum_{i=1}^{n} \underbrace{M_i^*}_{\text{r.v.}} \times \underbrace{\mathbf{1}_{(C_i=0)}}_{\text{r.v.}} \times \underbrace{\mathbf{1}_{(\widetilde{D}_i=1)}}_{\text{r.v.}} \tag{A7}$$

where the variables M_i^*, C_i and $\widetilde{D}_i$ are random. Under the assumptions A1-A5 and the DGP described by equations 1-9, the conditional moments of the aggregate $STax$ verifies:

$$\mathbb{E}_X\left(STax\right) = \sum_{i=1}^{n} \mathbb{E}_X(M_i^* | \, (C_i = 0) \cap (\widetilde{D}_i = 1)) \times \Pr((C_i = 0) \cap (\widetilde{D}_i = 1)) \tag{A8}$$

$$\mathbb{V}_X\left(STax\right) = \sum_{i=1}^{n} \mathbb{V}_X(M_i^* | \, (C_i = 0) \cap (\widetilde{D}_i = 1)) \times \Pr((C_i = 0) \cap (\widetilde{D}_i = 1)) \tag{A9}$$

Appendix B Moments of a multivariate normal distribution with double truncature

We consider a random vector $\mathbf{X} = (x_1, \ldots, x_d)'$ such that $\mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}\right)$. The moments of this variable of dimension d truncated between $\mathbf{a}$ in $\mathbb{R}^d$ and $\mathbf{b} \in \mathbb{R}^d$, such that $\alpha = \Pr\left(\mathbf{a} \leq \mathbf{X} \leq \mathbf{b}\right)$, are written as:

$$\mathbb{E}\left(X_i\right) = \sum_{k=1}^{d} \sigma_{i,k}\left(F_k\left(a_k\right) - F_k\left(b_k\right)\right) \tag{B1}$$

$$\begin{aligned} \mathbb{E}\left(X_i X_j\right) \quad = \quad & \sigma_{i,j} + \sum_{k=1}^{d} \sigma_{i,k} \frac{\sigma_{j,k}\left(a_k F_k\left(a_k\right) - b_k F_k\left(b_k\right)\right)}{\sigma_{k,k}} \\ & + \sum_{k=1}^{d} \sigma_{i,k} \sum_{q \neq k} \left(\sigma_{j,q} - \frac{\sigma_{k,q}\sigma_{j,k}}{\sigma_{k,k}}\right) \left[\left(F_{k,q}\left(a_k, a_q\right) - F_{k,q}\left(a_k, b_q\right)\right)\right. \\ & \left. - \left(F_{k,q}\left(b_k, a_q\right) - F_{k,q}\left(b_k, b_q\right)\right)\right] \end{aligned} \tag{B2}$$

where $F_i\left(x, y\right)$ and $F_{i,j}\left(x, y\right)$ are the marginal densities of the truncated laws and the joint densities such that:

$$F_i\left(x\right) = \int_{a_1}^{b_1} \ldots \int_{a_{i-1}}^{b_{i-1}} \int_{a_{i+1}}^{b_{i+1}} \ldots \int_{a_d}^{b_d} \varphi_{\alpha,\mathbf{\Sigma}}\left(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_d\right) \mathsf{d}x_d \ldots \mathsf{d}x_{i+1} \mathsf{d}x_i \ldots \mathsf{d}x_1 \tag{B3}$$

$$F_{k,q}\left(x, y\right) = \int_{a_1}^{b_1} \ldots \int_{a_{k-1}}^{b_{k-1}} \int_{a_{k+1}}^{b_{k+1}} \ldots \int_{a_{q-1}}^{b_{q-1}} \int_{a_{q+1}}^{b_{q+1}} \ldots \int_{a_d}^{b_d} \varphi_{\alpha,\mathbf{\Sigma}}\left(x, y, \mathbf{x}_{-k,-q}\right) \mathsf{d}\mathbf{x}_{-k,-q} \tag{B4}$$

where $\mathbf{x}_{-k,-q}$ denotes the vector of dimension $(d-2)$ defined by $(x_1,\ldots,x_{k-1},x_{k+1},\ldots,x_{q-1},x_{q+1},\ldots,x_d)'$, with

$$\varphi_{\alpha,\mathbf{\Sigma}}\left(\mathbf{x}\right)=\begin{cases}\frac{\varphi_{\mathbf{\Sigma}}(\mathbf{x})}{\alpha} & \text{if } \mathbf{a}\leq\mathbf{x}\leq\mathbf{b}\\ 0 & \text{otherwise}\end{cases}\qquad\forall i=1,\ldots,n \tag{B5}$$

where $\varphi_{\mathbf{\Sigma}}\left(\mathbf{x}\right)$ denotes the pdf of the vector normal distribution $\mathcal{N}\left(\mathbf{0},\mathbf{\Sigma}\right)$. For example, if $d=3$, we have:

$$F_1\left(a_k\right)=\int_{a_2}^{b_2}\int_{a_3}^{b_3}\varphi_{\alpha,\mathbf{\Sigma}}\left(a_k,x_2,x_3\right)\mathsf{d}x_2\mathsf{d}x_3 \tag{B6}$$

$$F_2\left(a_k\right)=\int_{a_1}^{b_1}\int_{a_3}^{b_3}\varphi_{\alpha,\mathbf{\Sigma}}\left(x_1,a_k,x_3\right)\mathsf{d}x_1\mathsf{d}x_3 \tag{B7}$$

$$F_3\left(a_k\right)=\int_{a_1}^{b_1}\int_{a_2}^{b_2}\varphi_{\alpha,\mathbf{\Sigma}}\left(x_1,x_2,a_k\right)\mathsf{d}x_1\mathsf{d}x_2 \tag{B8}$$

In the case of a double truncation such that $a_q\leq x_q\leq b_q$ and $a_r\leq x_r\leq b_r$ with $q\neq r$, Manjunath and Wilhelm (2012) show that the quantity $F_{q,r}\left(x,y\right)$ can be written as a product of a bivariate standard normal density $\varphi\left(x_q,x_r\right)$ and a normal integral of dimension $d-2$, noted as dimension $d-2$, denoted Φ_{d-2}:

$$\begin{aligned}F_{k,q}\left(c_q,c_r\right) &= \int_{a_1}^{b_1}\cdots\int_{a_{q-1}}^{b_{q-1}}\int_{a_{q+1}}^{b_{q+1}}\cdots\int_{a_{r-1}}^{b_{r-1}}\int_{a_{r+1}}^{b_{r+1}}\cdots\int_{a_d}^{b_d}\varphi_{\alpha,\mathbf{\Sigma}}\left(x_s,c_q,c_r\right)\mathsf{d}x_s &&\text{(B9)}\\ &= \alpha^{-1}\varphi\left(c_q,c_r;\rho_{qr}\right)\Phi_{d-2}\left(A^q_{rs};B^q_{rs};\mathbf{R}_{qr}\right) &&\text{(B10)}\end{aligned}$$

with $\mathbf{R}_{qr}$ the partial correlation matrix for $s\neq q\neq r$, and

$$A^q_{rs}=\frac{\left(a_s-\beta_{sq.r}c_q-\beta_{sr.q}c_r\right)}{\sqrt{\left(1-\rho^2_{sq}\right)\left(1-\rho^2_{sq.r}\right)}} \tag{B11}$$

$$B^q_{rs}=\frac{\left(b_s-\beta_{sq.r}c_q-\beta_{sr.q}c_r\right)}{\sqrt{\left(1-\rho^2_{sq}\right)\left(1-\rho^2_{sq.r}\right)}} \tag{B12}$$

where $\beta_{sq.r}$ is the partial regression coefficient of x_r in the regression of x_s on x_q and x_r, $\beta_{sr.q}$ is the partial regression coefficient of x_q in the regression of x_s on x_q and x_r. The calculation of these moments is implemented in the R package tmvtnorm (Manjunath and Wilhelm, 2010) or in the dtmvnmom function of Matlab (Kan and Robotti, 2017).

In the particular case $\sigma_c = \sigma_d = 1$, using the results of Rosenbaum (1961) and Muthén (1990), we can establish analytical results for the expectations:

$$\Delta_{c,i} = \mathbb{E}_X \left(\varepsilon_{c,i} \middle| \left(\varepsilon_{c,i} > a_{c,i} \right) \cap \left(\varepsilon_{d,i} > a_{d,i} \right) \right)$$

and

$$\Delta_{d,i} = \mathbb{E}_X \left(\varepsilon_{d,i} \middle| \left(\varepsilon_{c,i} > a_{c,i} \right) \cap \left(\varepsilon_{d,i} > a_{d,i} \right) \right)$$

such that:[22]

$$\alpha_i \Delta_{c,i} = \phi\left(a_{c,i}\right) \times \left(1 - \Phi\left(\frac{a_{d,i} - \rho_{cd} a_{c,i}}{\sqrt{1 - \rho_{cd}^2}}\right)\right) + \rho_{cd} \phi\left(a_{d,i}\right) \times \left(1 - \Phi\left(\frac{a_{c,i} - \rho_{cd} a_{d,i}}{\sqrt{1 - \rho_{cd}^2}}\right)\right) \quad \text{(B13)}$$

$$\alpha_i \Delta_{d,i} = \phi\left(a_{d,i}\right) \times \left(1 - \Phi\left(\frac{a_{c,i} - \rho_{cd} a_{d,i}}{\sqrt{1 - \rho_{cd}^2}}\right)\right) + \rho_{cd} \phi\left(a_{c,i}\right) \times \left(1 - \Phi\left(\frac{a_{d,i} - \rho_{cd} a_{c,i}}{\sqrt{1 - \rho_{cd}^2}}\right)\right) \quad \text{(B14)}$$

where $\alpha = \Pr\left(\left(\varepsilon_{c,i} > a_{c,i} \right) \cap \left(\varepsilon_{d,i} > a_{d,i} \right) \right) = \Phi_2\left(-a_{c,i}; -a_{d,i}; \rho_{cd} \right)$ denotes the probability associated with the truncation. We check that in the case where the shocks $\varepsilon_{c,i}$ and $\varepsilon_{d,i}$ are not correlated ($\rho_{cd} = 0$), the quantities $\Delta_{c,i}$ and $\Delta_{d,i}$ are the inverse of the Mills ratio, and since $\alpha_i = \left(1 - \Phi\left(a_{c,i}\right)\right) \times \left(1 - \Phi\left(a_{d,i}\right)\right)$, it comes:

$$\Delta_{c,i} = \frac{\phi\left(a_{c,i}\right) \times \left(1 - \Phi\left(a_{d,i}\right)\right)}{\left(1 - \Phi\left(a_{c,i}\right)\right) \times \left(1 - \Phi\left(a_{d,i}\right)\right)} = \frac{\phi\left(a_{c,i}\right)}{\left(1 - \Phi\left(a_{c,i}\right)\right)} = \lambda\left(a_{c,i}\right) \quad \text{(B15)}$$

$$\Delta_{d,i} = \frac{\phi\left(a_{d,i}\right) \times \left(1 - \Phi\left(a_{c,i}\right)\right)}{\left(1 - \Phi\left(a_{c,i}\right)\right) \times \left(1 - \Phi\left(a_{d,i}\right)\right)} = \frac{\phi\left(a_{d,i}\right)}{\left(1 - \Phi\left(a_{d,i}\right)\right)} = \lambda\left(a_{d,i}\right) \quad \text{(B16)}$$

Indeed, when $\rho_{cd} = 0$, we verify that:

$$\Delta_{c,i} = \mathbb{E}_X \left(\varepsilon_{c,i} \middle| \left(\varepsilon_{c,i} > a_{c,i} \right) \cap \varepsilon_{d,i} > a_{d,i} \right) = \mathbb{E}_X \left(\varepsilon_{c,i} \middle| \left(\varepsilon_{c,i} > a_{c,i} \right) \right) = \lambda\left(a_{c,i}\right) \quad \text{(B17)}$$

But in the case where the shocks $\varepsilon_{c,i}$ and $\varepsilon_{d,i}$ are correlated ($\rho_{cd} > 0$), the double truncation expectations $\Delta_{c,i}$ and $\Delta_{d,i}$ can be very different from single-truncated expectations based on Mills ratios.

[22] We do not have analytical results when the variances are different from 1. Similarly, we do not have analytical results on the variances and covariances with double truncation.

Appendix C   Numerical Illustration

We assume that:

$$\begin{pmatrix} \varepsilon_{c,i} \\ \varepsilon_{d,i} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}\right) \qquad \mathbf{\Sigma} = \begin{pmatrix} 1.5 & \rho \\ \rho & 3 \end{pmatrix} \tag{C1}$$

We set $b_{c,i} = \sigma_1 \Phi^{-1}(0.20) = -1.0308$ and $a_{d,i} = \sigma_2 \Phi^{-1}(0.70) = 0.9083$. Figure C1 displays the expectation of $\Delta_{c,i}$ and $\Delta_{d,i}$ with the function dtmvnmom from Matlab (Kan and Robotti, 2017)[23], and truncated expectation are defined by:

$$\mathbb{E}_X\left(\varepsilon_{c,i} \middle| \left(\varepsilon_{c,i} < b_{c,i}\right)\right) = -\sigma_c \frac{\phi\left(b_{c,i}/\sigma_c\right)}{\Phi\left(b_{c,i}/\sigma_c\right)} \tag{C2}$$

$$\mathbb{E}_X\left(\varepsilon_{d,i} \middle| \left(\varepsilon_{d,i} > a_{d,i}\right)\right) = \sigma_d \frac{\phi\left(a_{d,i}/\sigma_d\right)}{1 - \Phi\left(a_{d,i}/\sigma_d\right)} \tag{C3}$$

We observe that larger the correlation coefficient, higher difference between the Mills ratio and the expectation with double truncation is. For example, when $\rho = 0.6$, we get:

$$\begin{aligned} \mathbb{E}_X\left(\varepsilon_{c,i} \middle| \left(\varepsilon_{c,i} < b_{c,i}\right) \cap \left(\varepsilon_{d,i} > a_{d,i}\right)\right) &= -1.4337 \quad \text{against} \quad \mathbb{E}_X\left(\varepsilon_{c,i} \middle| \left(\varepsilon_{c,i} < b_{c,i}\right)\right) = -1.7144 \\ \mathbb{E}_X\left(\varepsilon_{d,i} \middle| \left(\varepsilon_{c,i} < b_{c,i}\right) \cap \left(\varepsilon_{d,i} > a_{d,i}\right)\right) &= 1.5118 \quad \text{against} \quad \mathbb{E}_X\left(\varepsilon_{d,i} \middle| \left(\varepsilon_{d,i} > a_{d,i}\right)\right) = 2.0074 \end{aligned}$$

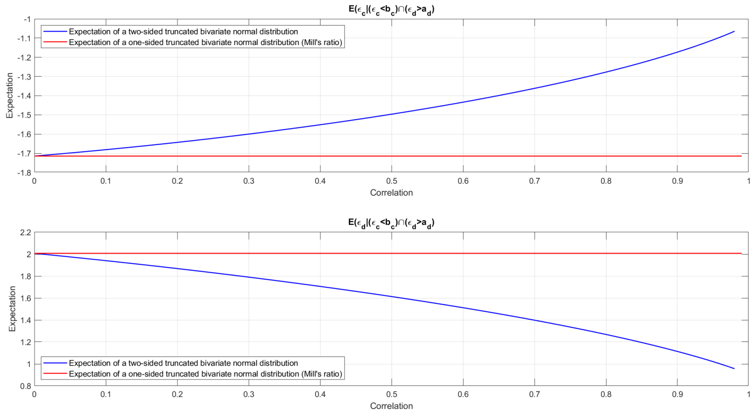

Figure C1: Mills ratio and expectation with double truncation.

This figure describes to what extent the expectation with a double truncation is different from an expectation with a simple truncation (Mills ratio).

[23] Similar results are obtained with the function mtmvnorm from the R package tmvtnorm (Manjunath and Wilhelm, 2010)

Appendix D Estimation by the Heckman correction method

This section presents the estimation approach based on the Heckman correction method (Heckman, 1976, 1979). Given the specificities of the research question, it can be designed around three steps:

Step 1: We consider all the firms in the sample ($i = 1, \ldots, n$) and we estimate by maximum likelihood the parameters β_c of the probit model related to the firms that have been controlled:

$$C_i = \begin{cases} 1 & \text{if } C_i^* = \mathbf{X}_{c,i}\beta_c + \varepsilon_{c,i} > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \forall i = 1, \ldots, n, \tag{D1}$$

with $\Pr\left(C_i = 1 \middle| \mathbf{X}_{c,i}\right) = \Phi\left(\mathbf{X}_{c,i}\beta_c\right)$. We denote by $\widehat{\beta}_c$ the maximum likelihood estimator of the vector β_c/σ_c, and by $\lambda_{1,i} = \lambda(\mathbf{X}_{c,i}\widehat{\beta}_c)$ the corresponding Mills ratios

Step 2: We consider only the n_1 firms that have been controlled and we estimate by maximum likelihood the following probit model:

$$D_i = \begin{cases} 1 & \text{if } D_i^* = \mathbf{X}_{d,i}\beta_d + \delta\lambda_{1,i} + \varepsilon_{d,i} > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \forall i : C_i = 1 \tag{D2}$$

with $\Pr\left(D_i = 1 \middle| C_i = 1, \mathbf{X}_{c,i}, \mathbf{X}_{d,i}\right) = \Phi\left(\mathbf{X}_{d,i}\beta_d + \delta\lambda_{1,i}\right)$. Consider $\widehat{\beta}_d$ and $\widehat{\delta}$ the maximum likelihood estimators of the vector β_d/σ_d and of the parameter δ. We denote by $\lambda_{2,i} = \lambda\left(\mathbf{X}_{d,i}\widehat{\beta}_d + \widehat{\delta}\widehat{\lambda}_{1,i}\right)$ the corresponding Mills ratios.

Step 3: We consider the firms that have been controlled and detected as fraudulent and estimate by ordinary least squares method the following linear model:

$$M_i = \mathbf{X}_{m,i}\beta_m + \gamma_1\lambda_{1,i} + \gamma_2\lambda_{2,i} + v_i \qquad \forall i : D_i = 1, C_i = 1 \tag{D3}$$

where the parameters γ_1 and γ_2 are functions of σ_v. We denote by $\widehat{\beta}_m$, $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ the estimators obtained at this stage.

The simulation results shows that 101 079 out of 1 000 000 companies, committed fraud, 50 432 were audited, of which 5 797 had committed fraud and were redressed. The estimated parameters are reported in Table (D1). The first panel reports the estimated parameters $\widehat{\beta}_c/\widehat{\sigma}_c$ of the Probit model associated with the control decision, as well as the true values of

Table D1: Heckman estimation.

This table reports coefficient estimates by the 3-steps approach of Heckman. The upper panel reports estimated coefficients for the control equation by using a Probit model. The middle panel reports estimated coefficients for the detection equation by using a Probit model without Mills ratio, while the bottom panel reports estimated coefficients for the detection equation by using a Pprobit model with Mills ratio. Estimated, true (ratios) and true (structural) coefficients are displayed.

	Estimated = $\hat{\beta}/\hat{\sigma}$	True ratios = β/σ	True structural = β
Control	-6.5168	-6.5549	-9.2700
Equation	0.7055	0.7071	1.0000
	-1.4052	-1.4142	-2.0000
	2.1031	2.1213	3.0000
	-2.8073	-2.8284	-4.0000
Detection	-6.8972	-7.1630	-10.1300
Equation	3.6079	3.5355	5.0000
without Mills ratios	-4.3528	-4.2426	-6.0000
Detection	-8.0973	-7.1630	-10.1300
Equation	4.0092	3.5355	5.0000
with Mills ratio	-4.8436	-4.2426	-6.0000
	0.8386	0.8000	1.1314

the ratios β_c/σ_c, and the values of β_c (True structural). We observe that the estimators are very close to true values.[24]

In the second panel, we report the estimators of the Probit model associated with the detection estimated only from the controlled firms ($C_i = 1$) without taking into account the selection bias (hence, without including Mills ratio). There is a slight shift from the true values. The shift is quite small because the censorship is not very important, but this result is not general. Beyond the importance of the lag as such, what is interesting to note is that the bias is less important than when a Mills ratio is introduced to take into account the selection bias. In this case, we paradoxically observe a greater bias than when we do not include a Mills ratio.

These observations show that the introduction of Mills ratio in the Probit model (at Step 2) is not judicious and leads to biased estimators. The Heckman correction method is in fact only suitable in the case of a linear model (See Chung and Goulias, 1995) and it is no longer valid in the case of a non-linear model such as the probit model (Greene, 2006). In turn, the Mills ratios at Step 3 are poorly estimated and do not allow to take into account the selection

[24] Recall that the ML estimators are convergent in this case, which means that their realizations tend to be close to the true values when n goes to infinity.

bias. Following Greene (2006):

"*Based on the wisdom in Heckman's (1979) treatment of the linear model, there seems to be a widespread tendency (temptation) to extend his approach to other frameworks by mimicking his two step approach. Thus, for example, Wynand and van Praag (1981), in an early application, proposed to fit a probit model with sample selection with the following two steps: Step 1. Fit the probit model for the sample selection equation. Step 2. Using the selected sample, fit the second step probit model merely by adding the inverse Mills ratio from the first step to the main probit equation as an additional independent variable. This approach is inappropriate for several reason*", Greene (2006), page 1.

Appendix E Data dictionary

Table E1: List of variables

Model variable	Variable name	Type	Description
Y1	ctrlex_2014	Dummy	Companies audited on the 2014 accounting base (over at least two quarters)
Y2	redres_2014	Dummy	Companies audited and adjusted (positively) on the 2014 accounting base (over at least one quarter)
Y3	Mt_redress_tot	Numerical	Annual (positive) adjustment amount (in euros)
$Y3_{bis}$	Mt_redress_tot_brut	Numerical	Annual (positive and negative) adjustment amount (in euros)
X	’ancentr’	Numerical	Company seniority (for companies with multiple locations, company seniority of the location with the highest employer contributions)
	’coti_tot’	Numerical	Total contribution amount
	’cotietat_cotitot’	Numerical	Share of exempted contributions acomputed as Amount of contribution exemptions/Total amount of contributions
	’ETP2015’	Numerical	Full-Time equivalent controllers per MSA office
	’segment1’	Dummy	Business segment (MSA or Crédit Agricole)
	’segment2’	Dummy	Business segment (Companies in LUCEA other than MSA and Crédit Agricole)
	’segment3’	Dummy	Business segment (Other)
	’nbsal41’	Dummy	Number of employees = 1
	’nbsal42’	Dummy	Number of employees = 2
	’nbsal43’	Dummy	Number of employees between 3 and 7
	’nbsal44’	Dummy	Number of employees ≥ 8
	’saison3cl_1’	Dummy	Categorical variable equal to 1 when (Number of seasonal contracts/Number of contracts) = 0; 0 otherwise
	’saison3cl_2’	Dummy	Categorical variable equal to 1 when 0 < (Number of seasonal contracts/Number of contracts) < 1; 0 otherwise
	’saison3cl_3’	Dummy	Categorical variable equal to 1 when (Number of seasonal contracts/Number of contracts) = 1; 0 otherwise
	’LCOTITOT_REG1’	Numerical	log (Total contributions) for firms located in the departments 28-51-75
	’LCOTITOT_REG2’	Numerical	log (Total contributions) for firms located in the departments 27-59-80
	’LCOTITOT_REG3’	Numerical	log (Total contributions) for firms located in the departments 21-52
	’LCOTITOT_REG4’	Numerical	log (Total contributions) for firms located in the departments 21-52
	’LCOTITOT_REG5’	Numerical	log (Total contributions) for firms located in the departments 11-48-84-13-20
	’LCOTITOT_REG6’	Numerical	log (Total contributions) for firms located in the departments 12-32-64-33
	’LCOTITOT_REG7’	Numerical	log (Total contributions) for firms located in the departments 24-87-63-26
	’LCOTITOT_REG8’	Numerical	log (Total contributions) for firms located in the departments 17-86-41-72-14
	’LCOTITOT_REG9’	Numerical	log (Total contributions) for firms located in the departments 22-35
	’LCOTITOT_REG10’	Numerical	log (Total contributions) for firms located in the departments 49-85
	’LCOTITOT_REG_sq1’	Numerical	Squared log (Total contributions) for firms located in the departments 28-51-75
	’LCOTITOT_REG_sq2’	Numerical	Squared log (Total contributions) for firms located in the departments 27-59-80
	’LCOTITOT_REG_sq3’	Numerical	Squared log (Total contributions) for firms located in the departments 21-52
	’LCOTITOT_REG_sq4’	Numerical	Squared log (Total contributions) for firms located in the departments 21-52
	’LCOTITOT_REG_sq5’	Numerical	Squared log (Total contributions) for firms located in the departments 11-48-84-13-20
	’LCOTITOT_REG_sq6’	Numerical	Squared log (Total contributions) for firms located in the departments 12-32-64-33
	’LCOTITOT_REG_sq7’	Numerical	Squared log (Total contributions) for firms located in the departments 24-87-63-26
	’LCOTITOT_REG_sq8’	Numerical	Squared log (Total contributions) for firms located in the departments 17-86-41-72-14
	’LCOTITOT_REG_sq9’	Numerical	Squared log (Total contributions) for firms located in the departments 22-35
	’LCOTITOT_REG_sq10’	Numerical	Squared log (Total contributions) for firms located in the departments 49-85