A Machine-Learning Based Measure of Corporate Tax Avoidance

Hui Chen* Yuhan Cheng† Jie Zou‡

July 6, 2025

Abstract

We exploit a unique feature of the Chinese tax system: state-owned enterprises (SOEs) and private firms are subject to the same tax laws, yet SOEs face markedly weaker incentives to avoid taxes. Leveraging this contrast, we train a machine learning model to predict the profits of SOEs and apply it to non-SOE firms (public and private) to infer their actual profits – and thereby estimate the extent of their tax avoidance. Our measure aligns closely with empirical proxies for the costs and benefits of tax avoidance, consistent with the predictions of the classic Allingham-Sandmo tax evasion model. We further validate our approach by examining the effects of a nationwide digital tax administration reform, which acts as an exogenous shock to the cost of tax evasion. Lastly, we demonstrate that our measure is robust to concerns about accounting manipulation and structural differences in production between SOEs and private firms.

Keywords: Corporate Tax Avoidance Measurement, Tax Avoidance of Non-Listed Companies, Automated Machine Learning

*MIT Sloan and NBER
†School of Management, Shandong University
‡School of Economics, Shandong University

1 Introduction

Tax avoidance is inherently covert, posing a fundamental challenge for empirical economics research. Several factors make the measurement of corporate tax avoidance difficult, particularly for private firms. The relationships between firms' tax liabilities and the wealth of financial and accounting data that describe firms' business activities can be highly complex, the data are subject to manipulation due to firms' heterogeneous tax avoidance incentives, and the effectiveness of monitoring is limited.

For listed companies, there are mature methods for measuring tax avoidance in the literature. Two common approaches are: Effective Tax Rate (ETR) and its variant, Book Tax Differences (BTD). Conceptually, both methods rely on the assumption that book income, which is reported to shareholders, is a relatively accurate measure of a firm's true profit and can be used to determine the degree of tax avoidance when compared to the firm's income tax expense. One obvious limitation of these methods is that they are only applicable to listed firms. A widely used proxy for tax avoidance for unlisted firms in the literature is proposed by Cai and Liu (2009). Their core idea is that a firm's true profits can be approximated using the income approach from national accounts. Specifically, they compute the imputed profit by

$$\text{Imputed Profit} = \text{Industrial Output} - \text{Intermediate Inputs} - \text{Labor} - \text{Depreciation} - \text{Taxes}. \tag{1}$$

A comparison between the imputed profit and reported profit submitted to the tax authority can indicate the level of tax avoidance.

An important caveat is that the imputed profit defined above is essentially accounting profit, which is typically different from firms' actual economic profit, on which the tax liabilities are determined. To address the inherent discrepancies between accounting profit and taxable profit, Cai and Liu (2009) propose a profit estimation method grounded in economic theory. Their method has two key limitations. First, the equation (1) is linear, which fails to capture the internal complexity of firms' cost structures and economic activities, making it difficult to accurately estimate true profits and, therefore, the scale of tax avoidance. Second, due to the inherent discrepancy between accounting and taxable profits, as well as the discrepancy between national account-based and accounting-based profit measures, two error terms in their specification, it is infeasible to determine the actual magnitude of tax avoidance. These limitations mean that the method of Cai and Liu (2009) can only be used to estimate the *direction* of the impact of certain factors on tax avoidance, such as government

fiscal pressure (Chen et al. (2022)), labor costs (Xiang et al. (2023)), or organizational capital Hasan et al. (2021), but not the absolute scale of tax avoidance.

Building on the research of Cai and Liu (2009), we propose a framework that utilizes machine learning to measure the extent of corporate tax avoidance. To more accurately estimate a firm's level of tax avoidance, one must first obtain an estimate of its true profit. The true profit, although not directly observable, can be viewed as a function of a set of firm-specific characteristics, denoted as $f(x)$. Importantly, $f(\cdot)$ is a function that maps observable financial and accounting variables, such as sales, operating income, VAT paid, number of employees, etc., to the firm's true profit, which is mainly determined by the tax law, which is common across all firms (listed vs. non-listed, SOEs vs. non-SOEs), and not the firm's production function, which can differ significantly across firms. Therefore, if this mapping function $f(x)$ can be learned in the data, it becomes possible to predict a firm's true profit.

China's rich administrative datasets contain both the profits reported to the tax authorities and a wide range of firm characteristics. We exploit a unique feature of the Chinese tax system: state-owned enterprises (SOEs) and private firms are subject to the same tax laws, yet SOEs face markedly weaker incentives to avoid taxes. This enables us to use the data from the SOE sample to train a machine learning model for $f(x)$. We then apply it to estimate the true profit for non-SOE firms and compare it to the profit reported to the tax authorities to calculate the extent of tax evasion.

Compared to the method of Cai and Liu (2009), which is based on national accounting rules, machine learning models offer significant advantages. They can more effectively exploit the information embedded in high-dimensional features and capture complex nonlinear relationships among variables, thereby enabling more accurate profit prediction in theory.

Another key improvement is that imputed profits derived from national income accounting are closer to a firm's accounting profit. However, due to differences between accounting and tax recognition standards, the gap between accounting profit and tax profit is inherently not indicative of tax avoidance. Our method directly predicts the firm's true taxable profit and uses the difference between predicted taxable profit and reported taxable profit as the measure of tax avoidance, thus avoiding distortions caused by systematic differences between accounting and tax profits.

However, this approach faces a fundamental challenge: supervised learning models require paired input-output data–in other words, not only the input variables x but also the corresponding true profit as the label. Yet true profit is unobservable, creating a seemingly recursive problem: to learn $f(x)$, we need true profit as training data; but to obtain true

profit, we would first need to know $f(x)$.

China's state-owned enterprises (SOEs) offer a solution to the recursive problem described above. Benguria and Saffie (2025) highlight key characteristics of SOEs: as representatives of the government in the market, SOEs are obligated to pay taxes and remit all retained earnings to the state. In practice, SOEs divide their pre-tax profits into two parts—one portion is paid as taxes to the general public budget, and the other is remitted as after-tax profits to the state-owned capital operation budget. Since both portions ultimately go to the government regardless of how the profits are split, SOEs have almost no economic incentive to reduce their tax burden for the sake of retaining more profit.

Therefore, in Chinese administrative datasets, the profits reported by SOEs can be reasonably considered close to their true taxable profits. Based on this feature, we use SOEs as the training set and model the mapping function $f(x)$ between their firm-level characteristics and reported profits. This function can then be applied to other firms (e.g., private enterprises) to estimate their true taxable profits. The discrepancy between these predictions and the firms' reported taxable profits serves as a measure of tax avoidance. In other words, the deviation between a private firm's reported taxable profit and the model's prediction reflects the firm's potential level of tax evasion.

We utilize the Chinese State Administration of Taxation (SAT) dataset, which includes firm-level income statements, balance sheets, cash flow statements, and tax filing data. This dataset covers both SOEs and private firms of various sizes and industries, enabling us to explore heterogeneity in tax avoidance behavior across different types of firms.

We use the ratio of (predicted profit − reported profit) to total assets as our measure of a firm's tax avoidance[1], which reflects how much profit a firm underreports per unit of assets. The results show that state-owned enterprises (SOEs), which serve as our training sample, have reported profits that closely match their predicted profits, with an average avoidance index of approximately 0.005. In contrast, private enterprises exhibit a much higher level of tax avoidance, with an average avoidance index of about 0.217. This suggests that our method effectively captures the tax avoidance behavior of private firms.

We also find that firms awarded an A-grade taxpayer status by the tax authorities tend to have relatively low levels of tax avoidance, with an average avoidance index of around 0.041. Listed non-state-owned enterprises come next, with an average index of approximately 0.085.

However, there are three main concerns regarding the construction of the tax avoidance index. The first issue relates to whether the reported profits of state-owned enterprises (SOEs) accurately reflect their true profits. Although the majority of SOE profits are eventually

[1]Dividing by total assets removes the unit of scale.

remitted to the government, this does not eliminate the possibility that SOE managers have some incentives to engage in tax avoidance. For instance, if managers place greater emphasis on after-tax profits as a key performance indicator, they may have motivation to reduce tax payments in order to inflate after-tax figures.

If SOEs indeed engage in mild tax avoidance, then the tax avoidance index we estimate for private firms—based on a model trained on SOEs—would constitute a **lower bound**, because the model would systematically underestimate the true profit baseline. On the other hand, if SOE managers place more value on tax contributions themselves, or intentionally overpay taxes under the guidance of local governments, the reported profits may exceed true profits, causing our estimate for private firms' tax avoidance to become an **upper bound**. In practice, the former is more likely—that is, SOEs may also engage in some degree of tax avoidance, though their incentives to do so are significantly weaker than those of private firms.

To test the reliability of SOE reported profits, we compare them with the profits predicted by our machine learning model. We find that firms with overreported and underreported profits do not differ systematically across a range of key observable characteristics. These include: the promotion prospects of executives, local government fiscal pressure, regional economic development level, firm age and size, industry classification, the scale of the local tax authority, and the degree of tax digitalization.

Our estimation results show that the predicted tax-adjusted profits of private firms under our method are generally much lower than those estimated using Cai and Liu's method. This is consistent with the interpretation that their approach provides an *upper bound* of tax avoidance, while our method arguably yields a more conservative *lower bound*.

The second concern lies in whether the profit-generating process $f(x)$ is consistent between state-owned enterprises (SOEs) and non-SOEs. If SOEs and private firms follow different underlying functions $f(x)$, then a machine learning model trained on SOE data to estimate $f(x)$ may not generalize to private firms. In such a case, the predicted profits for private firms—obtained via $f(x)$—would no longer serve as a good approximation of their true tax-adjusted profits.

To address this issue, we conduct two discussions. First, we modify the prediction process by changing the outcome variable from the firm's tax profit to the difference between the tax profit and the imputed profit based on national accounts. The benefit of this transformation is that potential inconsistencies in $f(x)$ between SOEs and non-SOEs may be captured by the imputed profit. By subtracting it from the tax profit, we reduce heterogeneity in the outcome variable, allowing the learning algorithm to focus on a more stable relationship with

the predictors.

For example, if certain firm characteristics—such as market power—simultaneously increase both the tax profit and the imputed profit, these effects will cancel out in the differenced outcome, leading to a more consistent data-generating process across ownership types.We find that this adjustment has only a minimal effect on the results: the average level of the avoidance index changes by less than 1% regardless of whether we use the original or adjusted outcome variable.

Second, we find that among private firms, those with an A-grade taxpayer rating from China's tax authority exhibit even lower levels of tax avoidance. By including these A-rated private firms in the training sample, we effectively introduce private-sector observations into the learning process. This helps mitigate the concern that the profit-generating function may differ systematically between SOEs and private firms.

We collect the list of A-rated taxpayers and incorporate them into the training data. After adjusting the training sample, we find that the average level of the avoidance index changes by less than 0.2%, suggesting that the model is robust to this modification and that our results are not sensitive to the potential heterogeneity in profit formation across ownership types.

The third potential concern with our method for estimating tax avoidance is that if a firm decides to manipulate its reported tax profit, then the entire financial statement may be affected. In such cases, the input factors used in the prediction model could themselves be distorted and no longer reflect true underlying values. We offer three responses to this concern. First, we exclude input factors that are easily subject to managerial manipulation and retain only those that are verifiable in practice—such as total assets, liabilities, and interest expenses—as model inputs. These variables typically require supporting documentation when submitted to tax authorities and are thus difficult to falsify. After replacing the set of input factors, we find that the average value of the avoidance index changes by less than 0.8%, indicating that the results are not sensitive to this adjustment.

Second, we examine the feature importance scores of all input factors used in the model. We find that the factors with the highest predictive contribution are those that are less prone to manipulation, including sales revenue, total assets, total liabilities, industry classification, city location, and employee compensation. In contrast, factors that are more discretionary in nature—such as administrative expenses, selling expenses, and R&D expenditures—exhibit much lower importance in the prediction process.

Third, we utilize an alternative dataset obtained from the tax authority of a developed province in China. This dataset is comparable in nature to the SAT data, as both are

administrative in origin, although the provincial dataset includes slightly fewer variables. The key distinction, however, is that this dataset is reported at the **monthly** level.

In practice, accounting or tax officers in Chinese firms typically make adjustments to financial statements during the annual tax settlement period. As a result, monthly filings are often submitted before such adjustments are made, and thus may reflect more truthful and timely reporting. Leveraging this monthly administrative dataset, we re-estimate the tax avoidance index and find that the results are highly consistent: the magnitude of the avoidance index remains essentially unchanged.

To validate the effectiveness of our tax avoidance index, we conduct a series of tests from both theoretical and empirical perspectives. Drawing on the classical Allingham-Sandmo model and informed by practical observations, we categorize the determinants of corporate tax avoidance into two main groups: the expected benefit from evasion and the strength of tax enforcement. We construct a set of variables to proxy for these two dimensions. For the expected benefit of evasion, we include the statutory tax rate, industry competition, financing constraints, and preferential tax policies. For enforcement strength, we include the number of local tax officers, turnover of local officials, fiscal pressure, the local tax-sharing ratio, and the degree of tax system digitalization. Our regression results show that these variables are significantly associated with the avoidance index, and the directions of the effects are consistent with the predictions of the Allingham-Sandmo framework: greater potential benefit from evasion is associated with higher levels of tax avoidance, while stronger enforcement is associated with lower levels of avoidance. These findings support the theoretical validity of our index.

In particular, we find that when the government allocates more fiscal resources to tax digitalization and enforcement infrastructure, firms' tax avoidance decreases. In 2010, the Chinese government launched a comprehensive digital reform of its tax administration system—namely, the Golden Tax System Phase III (GTS-III). A growing body of evidence shows that the implementation of GTS-III significantly reduced corporate tax avoidance. Our estimates capture this decline as well: following the rollout of GTS-III, the average amount of tax avoidance per firm decreased by approximately RMB 10,675. Based on this figure and the number of firms in China, we estimate that the total increase in tax revenue amounts to roughly RMB 384.3 billion. In comparison, the total implementation cost of the system was around RMB 100 billion, suggesting that the fiscal return far exceeds the expenditure.The Chinese case illustrates that investing in digital tax enforcement is an effective policy tool for reducing corporate tax avoidance. Our findings support the notion that allocating more resources toward tax system digitalization can yield substantial compliance gains.

We also conduct several additional and insightful tests. First, we examine how effectively Chinese tax authorities detect corporate tax avoidance. We find that the correlation between firms' total tax penalties and our estimated tax avoidance index is only 11%. This suggests that while the tax authority does catch some cases of avoidance, a substantial portion likely goes undetected.

Second, we address a long-standing and contested question: do small firms or large firms engage in more tax avoidance? Our findings indicate that, in China, tax avoidance generally increases with firm size. However, the relationship follows an inverted-U shape: mid-sized firms exhibit the highest levels of tax avoidance. This pattern reflects the specific structure of China's tax system, including its differentiated tax rates and enforcement strategies. It also highlights the importance of the "missing middle" phenomenon in understanding tax compliance dynamics.

The primary contribution of this paper is to propose a novel method for measuring corporate tax avoidance. In public economics research, one of the greatest challenges in studying tax avoidance is the difficulty in accurately quantifying the extent of evasion. As a result, many existing studies are limited to listed firms, and often cannot disentangle the mechanical differences between accounting profits and taxable profits from true tax avoidance behavior.

Our approach offers a new framework for estimating the degree of tax avoidance among unlisted firms. More importantly, it allows us to directly quantify tax avoidance with respect to taxable profits, providing a more targeted and policy-relevant measure of corporate tax compliance.

ETR-based methods assume that the income tax expense reported in financial statements reflects taxable income, while book income, being disclosed to shareholders, is relatively accurate. Thus, tax avoidance can be measured by dividing tax expense by book income. (i) GAAP Effective Tax Rate (GAAP ETR or GETR) is calculated as total income tax expense divided by pre-tax accounting income minus special items. This indicator captures tax avoidance strategies that result in permanent tax differences Cen et al. (2017). (ii) Cash Effective Tax Rate (Cash ETR or CETR) is calculated as cash taxes paid divided by pre-tax accounting income minus special items. This metric captures all cash-saving tax strategies implemented within a fiscal year, addressing the GAAP ETR's limitation of not reflecting actual cash outflows Dyreng, Hanlon, and Maydew (2008). (iii) Current Effective Tax Rate (Current ETR) is defined as total income tax expense minus deferred tax expense, divided by pre-tax accounting income minus special items. This indicator reflects tax savings resulting from both permanent and temporary tax differences (Cheng et al. (2012); Chen et al. (2019)).

The Book-Tax Difference (BTD) approach is conceptually similar to the ETR method. Book income is reported to shareholders and is therefore considered more reliable, while taxable income is reported to tax authorities. A larger gap between the two suggests a higher degree of tax avoidance (Desai and Dharmapala (2006); Desai, Foley, and Hines Jr (2006)). Unlike the ETR approach, the BTD method accounts for differences between accounting and tax standards in how income and expenses are recognized, making it more precise. Specifically, taxable income is calculated as: Taxable income = Current income tax expense / Nominal income tax rate. A higher BTD indicates a greater discrepancy between book profit and taxable income, which in turn implies a higher likelihood of tax avoidance. Desai and Dharmapala (2006) further refine this measure by introducing the Discretionary Book-Tax Difference (DDBTD), which isolates the portion of BTD not explained by accruals. DDBTD is estimated using the following model:

$$BTD_{i,t} = \alpha TACC_{i,t} + \mu_i + \varepsilon_{i,t} \tag{2}$$

TACC represents total accruals, calculated as (Net Profit - Net Cash Flow from Operating Activities) / Total Assets.μ_i denotes the average residual for firm i over the sample period, representing how the residual in year t deviates from the firm's historical average. ε_{it} captures the portion of the Book-Tax Difference (BTD) that cannot be explained by accruals.

The structure of the paper is as follows. Chapter 2 introduces the methodology for measuring corporate tax avoidance using machine learning. Chapter 3 evaluates the predictive performance of the model. Chapter 4 presents applications of the tax avoidance index. Chapter 5 concludes.

2 Methodology for Measuring Corporate Tax Avoidance Using Machine Learning

2.1 Overview of Cai and Liu (2009)

Cai and Liu (2009) propose to measure corporate tax avoidance by considering the following accounting identity:

$$PRO_{i,t} = Y_{i,t} - MED_{i,t} - FC_{i,t} - WAGE_{i,t} - DEP_{i,t} - VAT_{i,t} \tag{3}$$

In this formula, PRO refers to the imputed profit, Y is industrial output, and MED, FC, $WAGE$, DEP, and VAT represent intermediate inputs, financial costs, total wages, current

depreciation, and value-added tax, respectively.

Formula (2) strictly follows the rules of national income accounting to compute the profit that a firm should have earned according to macroeconomic principles. However, the imputed profit derived from this calculation is not equivalent to the firm's true profit and therefore cannot serve as a direct proxy. Nevertheless, in theory, the two should be positively correlated, that is:

$$\pi_{i,t} = \eta_{i,t} + PRO_{i,t} + \theta_{i,t}. \tag{4}$$

Here, π_{it} represents the firm's true profit, and η_{it} is an unknown parameter that captures the inherent differences between the national income accounting method and the accounting standards used by each firm to calculate profit—such as how depreciation is treated. This parameter may be positive or negative. θ is a random disturbance term with an expected value of zero. Assuming the firm's reported profit is denoted as RPRO, the relationship between reported profit and true profit can be expressed as:

$$RPRO_{i,t} = d_{i,t} * \pi_{i,t} + \varepsilon_{i,t} + \zeta_{i,t}. \tag{5}$$

Here, d_{itd} represents the degree of tax avoidance by the firm. A smaller value of d_{itd} indicates that the firm is under-reporting its profits more significantly, and thus engaging in more severe tax avoidance. ε_{it} reflects the systematic differences between accounting standards and tax regulations. ζ_{it} is a random disturbance term with a mean of zero.

Combining Equations (3) and (4), we obtain:

$$RPRO_{i,t} = d_{i,t} \times (\eta_{i,t} + PRO_{i,t} + \theta_{i,t}) + \varepsilon_{i,t} + \zeta_{i,t}. \tag{6}$$

Moreover, d_{itd} is influenced by a variety of factors:

$$d_{i,t} = \beta_0 + \beta_X X_{i,t} + \beta_2 \text{Control}_{i,t} + \beta_3 T_t + \beta_4 L_l + \upsilon_{i,t} \tag{7}$$

Therefore, to examine the relationship between $X_{i,t}$ and d_{it}, one only needs to estimate the following regression: the dependent variable is the firm's reported profit (RPRO_{it}), and the key explanatory variable is $X_{i,t} \times \text{PRO}_{it}$. If the estimated coefficient is significantly positive at the 10% level, it suggests that the factor reduces tax avoidance; conversely, a significantly negative coefficient indicates that it increases tax avoidance.

2.2 Our New Method to Measure Tax Avoidance

To overcome the limitations of Cai and Liu (2009), two conditions must be met. First, if administrative data are used, and they already contain the firm's reported taxable profits, it becomes essential to accurately obtain the firm's true taxable profit so that the difference between the two can be used to precisely quantify the degree of tax avoidance. Second, when predicting the firm's true taxable profit, the model must capture the nonlinear and complex relationships between profit and various input features.

Compared with econometric models, machine learning offers a purely data-driven approach. It eliminates the need for manual variable selection and model specification, and instead relies on complex model structures and a large number of parameters to automatically uncover the nonlinear interactions between firm characteristics and true profits. Additionally, it employs regularization techniques to mitigate overfitting, enabling more accurate predictions. Clearly, the relationship between taxes and underlying financial variables is highly nonlinear and difficult to model manually—thus, machine learning is a necessary tool.

The uniqueness of the Chinese market makes this research feasible. If all firms engaged in tax avoidance to the same extent, machine learning would not be able to distinguish between evading and non-evading firms. Fortunately, structural features of China's market make this differentiation possible.

First, China operates a dual-track system with one of the largest state-owned sectors in the world. In state-owned enterprises (SOEs), executives are appointed by the government, and all profits are remitted to the state. As a result, managers have no incentive to engage in tax avoidance **?**. Tan and Wang (2010) found that SOEs use fewer tax shields compared to private firms. Shaheer et al. (2019) showed that SOE tax avoidance decreases further during economic downturns, helping to offset government revenue shortfalls.

Based on this evidence, we assume that SOEs have a lower degree of tax avoidance, and that their reported profits more closely approximate their true profits. These firms can therefore serve as a training set to model the mapping $f(x)$ between firm features and true profit.

Machine learning has already been applied in the literature to label tax avoidance behavior. For example, Pérez López(2019) used deep neural networks to predict individual tax evasion probabilities and achieved improved accuracy over prior studies. Murorunkwere et al. (2022) used neural networks to detect tax fraud in income tax data and found that artificial neural networks performed well. Bao et al. (2020) applied machine learning to predict corporate financial statement restatements. In these cases, the data clearly identified which entities

were committing tax fraud and which were not.

2.3 Data

This study utilizes data from the Chinese State Administration of Taxation (SAT) covering the years 2007 to 2017. The dataset was compiled by the Ministry of Finance and the SAT, encompassing enterprises across all regions and industries in China. It includes not only large enterprises but also a substantial number of small firms.

The survey collects information related to firms' financials, taxation, operations, investments, and import/export activities, and contains a wealth of variables that can be used as predictors. The dataset includes records of firms' reported profits. Since this data is collected by agencies with tax enforcement authority, firms typically ensure that the profits they report in the survey match those in their official tax filings to avoid triggering audits. However, for firms that engage in tax avoidance, their reported profits may still deviate from their true profits.

We cleaned the data as follows: we apply winsorization to key variables at the top and bottom 1%; we remove samples where basic accounting or logical relationships between indicators were violated, such as current assets being greater than total assets, or current liabilities being greater than total liability; we also remove samples with obvious data entry errors, such as negative values for indicators that should not be negative (e.g., negative operating income or negative total assets).

Data Name	Administrative	Variable	Sample
Chinese State Administration of Taxation Data	State Administration of Taxation and Finance	Having all the tax records(tax profits) of the enterprise, financial statement variables, especially tax penalties and the adjustment amounts made by the tax authority to the tax base in the enterprise's reports	National sample, large enterprises are the panel, SMEs are the cross section data, approximately one million samples a year

Table 1: **Data Description**

2.4 Machine Learning Workflow

To estimate the magnitude or degree of corporate tax avoidance, we must first estimate a firm's true profit. The difference between reported and true profit is the basis for our measure

of tax avoidance. While reported profit is directly observed in the data, true profit must be inferred.

We begin by identifying a group of firms—referred to as the *real group*—whose reported profits are believed to closely reflect their actual profits because they have little incentive to engage in tax avoidance. This group forms the training set, from which we derive the mapping between firm characteristics and true profits, denoted as:

$$\text{profit}_{it} = f(x_{it}) \tag{8}$$

This function captures how much profit a firm is expected to generate given its assets, liabilities, production, and spending profile. For firms outside the real group—labeled the *evasion group*—if they share the same profit-generation process, then their true profits can also be approximated using $f(x_{it})$. The gap between their reported profits and model-predicted profits becomes our measure of tax avoidance.

Fortunately, China provides a clear source of low-avoidance firms: state-owned enterprises (SOEs). In these firms, executives are appointed by the government and all profits are remitted to the state, removing the incentive to underreport profits. Studies by **??** show that SOEs use fewer tax shields compared to private firms. **?** found that SOEs even reduce tax avoidance during economic downturns to help meet government revenue targets.

Based on this evidence, we assume that SOEs have lower levels of tax avoidance and that their reported profits are closer to the truth, making them suitable for training the profit estimation model $f(x)$.

We employ an expanding window strategy for model training. Starting in 2014, for each year k, we use 70% of the SOE data from 2007 to year k to train the model that learns the relationship between firm characteristics and profit, reserving the remaining 30% for model selection. We then retrospectively estimate the tax avoidance of private firms in year k. This process is repeated by rolling the window forward one year at a time.

Figure 1 illustrates this workflow.

2.5 Automated Machine Learning

In choosing technical solutions, we aimed to minimize the influence of manual model selection and hyperparameter tuning on tax avoidance estimation. Typically, when predicting corporate profits, one might construct models based on various firm-level factors (e.g., sales, production costs, market demand, macroeconomic indicators). However, traditional manual modeling

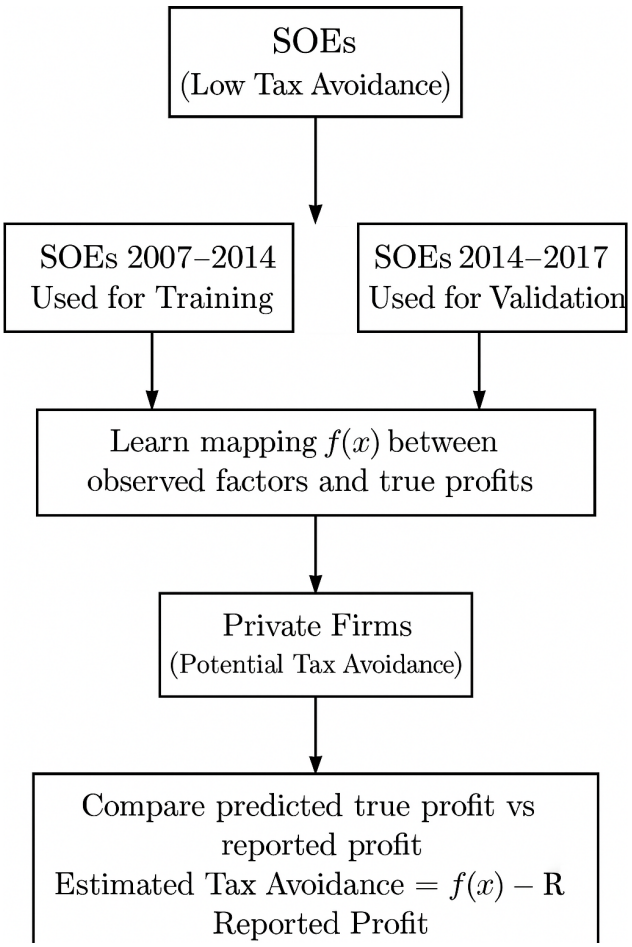


Figure 1: **Workflow** The flowchart presents the modeling framework of the paper, which differs significantly from the conventional approach.

requires substantial expert knowledge and repeated parameter tuning, which inevitably introduces subjectivity and high time costs. Different modeling approaches often yield varying results, so a unified and credible pipeline—covering preprocessing, feature engineering, model selection, and hyperparameter tuning—is necessary to reduce human intervention in the modeling process.

Automated Machine Learning (AutoML) offers significant advantages in this context, making it an ideal choice for corporate profit prediction and allowing for consistent, comparable outputs across models. AutoML automates the full pipeline, integrating data preprocessing, feature engineering, model selection, hyperparameter optimization, and evaluation. Its goal is to find the optimal or near-optimal modeling strategy with minimal manual input, thereby avoiding the biases or premature assumptions that can arise from human model selection. For instance, defaulting to linear or tree-based models may overlook more predictive alternatives, while AutoML searches and compares a wide space of algorithms.

Unlike traditional workflows, where data cleaning, feature construction, model selection, and tuning are handled manually, AutoML manages all steps systematically. It employs a range of search strategies (e.g., random search, Bayesian optimization, genetic algorithms, reinforcement learning) to improve efficiency and model performance.

We are given a dataset $\mathcal{D}$ (which can be further divided into a training set and a validation set), and a search space $\mathcal{A}$ consisting of a set of feasible models and parameter configurations. For any configuration $\alpha \in \mathcal{A}$, we obtain the model h_α by training on the training set. At this point, we are concerned with minimizing the loss or evaluation metric $\mathcal{L}\big(h_\alpha, \mathcal{D}\big)$ on the validation set. Therefore, the core of AutoML can be represented as the following optimization process:

$$\hat{\alpha} \;=\; \arg\min_{\alpha \in \mathcal{A}} \mathcal{L}\big(h_\alpha, \mathcal{D}\big). \tag{9}$$

where $\mathcal{L}$ is the chosen loss function (e.g., Mean Squared Error) on the validation data D_{val}.

AutoML iteratively explores combinations of values and possible model architectures or algorithms on the user-provided dataset, aiming to find the configuration that minimizes MSE within a limited search time. We adopt a rolling window strategy (as described earlier) to avoid forward-looking bias from future data.

Firm profits are affected by numerous factors, and the data exhibit high dimensionality, nonlinearity, and temporal dynamics. Without a deep understanding of these complex interactions, manually built models may lack accuracy or robustness to distributional shifts. AutoML addresses this by exploring a broad range of algorithms—including neural networks, ensemble methods (e.g., random forests, gradient boosting machines), support vector machines, and more advanced architectures. It compares their performance systematically and can also apply ensemble learning to combine predictions from multiple models, further enhancing stability and generalization.

2.6 Feature Selection

In the accounting literature, there are generally two approaches to selecting input variables for profit prediction: one is the traditional use of ratio data (e.g., debt-to-asset ratios). The other, which is more prevalent in recent years, uses raw financial data.

In our study, we adopt raw data as input and verify that the selected features are sufficiently robust—adding new features (including ratio data) does not significantly improve

performance on either the training or validation sets.

The raw features used include: accounts receivable, inventory, total assets, fixed assets, depreciation, investment amount, total liabilities, current liabilities, sales revenue, cost of goods sold, paid value-added tax (VAT), export value, export rebate, selling expenses, administrative expenses, R&D expenditures, outsourced R&D and technical services, purchased intangible assets, interest expenses, wage expenses, number of employees, net cash flows from operating, financing, and investing activities, industry (four-digit industry code), district, regional GDP growth, population size, industrial structure, ownership type.

2.7 Key Issues and Solutions

One major concern with using SOEs to estimate the mapping between firm fundamentals and taxable profits—and then applying this to non-SOEs—is whether SOEs and non-SOEs follow the same profit-generating process. To address this, we explore two strategies: Instead of predicting profit directly, we predict the difference between a firm's reported profit and its imputed profit, calculated as:

$$\text{Imputed Profit} = \text{Industrial output} - \text{Intermediate input} - \text{Labor} - \text{Depreciation} - \text{Taxes} \tag{10}$$

This adjustment allows the imputed profit to absorb systematic differences in the profit-generation process between SOEs and non-SOEs. For example, if SOEs have lower labor costs, both their profits and imputed profits would increase proportionally, preserving comparability in the residual component. We find that this transformation does not materially affect our core results.

Second, we include A-rated firms—those awarded a top compliance rating by Chinese tax authorities—as part of the training set. We collect lists of A-rated firms from 2014 to 2020. The A-rating signifies strong tax compliance, so these firms are expected to have low levels of tax avoidance. While literature confirming this is limited, we treat the inclusion of these firms as a robustness check.

It is important to note that whether we use SOEs or A-rated firms, their reported profits—while close to actual profits—are unlikely to be overstated. More likely, reported profits are slightly lower than true profits, making our tax avoidance estimates a conservative (lower-bound) measure. In theory, if firms that overreport profits could be identified, an upper bound could be estimated—but such firms are rare and hard to observe.

Another issue is that if a firm manipulates its taxable profit, its entire financial report may be affected, potentially distorting the input variables. To address this, we take two

actions: First, we remove input variables such as accounts receivable, administrative expenses, selling expenses, R&D expenses, outsourced R&D expenses, intangible assets, and cash flow variables. These are prone to discretionary adjustments. The remaining inputs—such as fixed assets, liabilities, and interest expenses—are less susceptible to manipulation and often require supporting documentation during tax filing.

Second, we validate our model using an alternative administrative dataset from a provincial tax bureau. Although this dataset contains slightly fewer variables, it is monthly rather than annual. Since firms typically revise financial reports during annual tax reconciliation, monthly submissions (filed earlier) are less likely to have been adjusted. We re-estimate tax avoidance using this monthly dataset and find consistent results.

3 Empirical Results

3.1 Prediction Performance

Table 3 shows the prediction results of the linear model and different machine learning models. This paper uses R^2 to evaluate the predictive performance of the models:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{11}$$

Where y_i is the reported profit of the enterprise, $\bar{y}$ is the mean of the reported profit of the enterprise, and $\hat{y}_i$ is the predicted value of the true profit of the enterprise from the machine learning model. The enterprises in both the training set and validation set are state-owned enterprises, so the reported profit of the enterprise can serve as a good proxy for its true profit. Therefore, the higher the R^2 in the training set, the better the model fits the true profit of the enterprise. Similarly, the higher the R^2 in the validation or test set, the better the model predicts the true profit of the enterprise on new data. In our results, the R^2 is 0.849 for the SOE training sample and 0.802 for the validation sample; on the private-firm dataset, it is 0.381. We do not expect the model to achieve an exceptionally high R^2 on the non–SOE dataset; otherwise, it would imply that private firms engage in no tax avoidance at all, an assumption that clearly defies common sense.

Figure 2 reports the mean values of the estimated profit (Imputed) and reported profit (Report) in the decile range for both aotuML (Pre_ML) and the method used by Cai et al. (2009). The numbers from 0 to 9 represent the decile ranges from the lowest to the highest. In each decile range, the mean predicted profit (Pre_ML) from aotuML is greater than the

mean reported profit, indicating that aotuML effectively identifies corporate tax avoidance. Additionally, the imputed profit is greater than Pre_ML at any decile, suggesting that using the method from Cai et al. (2009) estimates an upper bound, while our method provides a lower bound.

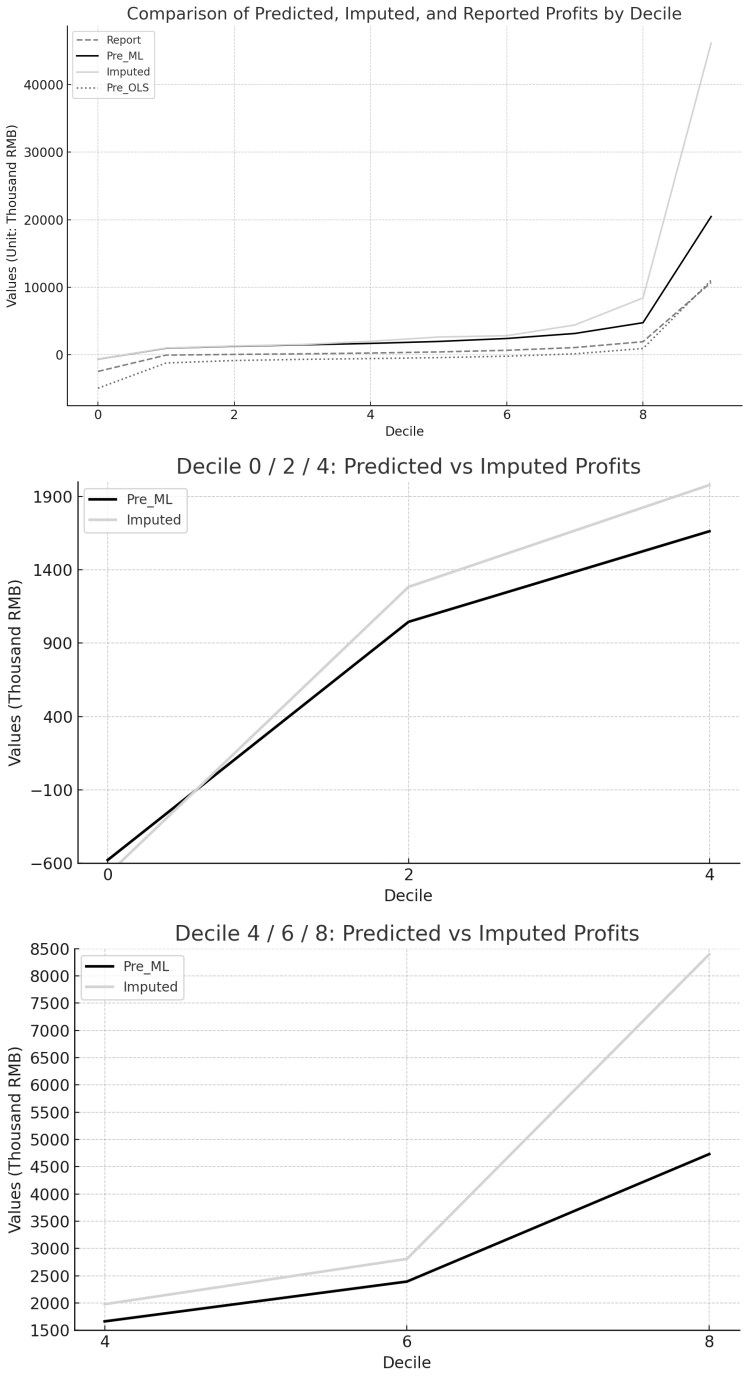


Figure 2: **Comparison With the Tax Avoidance Measures in Previous Studies** In the three figures, the middle and bottom panels provide magnified views of the top panel.

Figure 3 reports the distribution of the difference between true profit (imputed profit) and reported profit in both state-owned and non-state-owned enterprise samples. The majority of the samples have a positive difference, but the difference between the two profits is much smaller for state-owned enterprises compared to non-state-owned enterprises. Since enterprises should not have an incentive to overstate their profits to the tax authorities, we removed the companies where the predicted profit is lower than the reported profit. These companies may have some estimation bias (about 0.6% of the samples were removed).

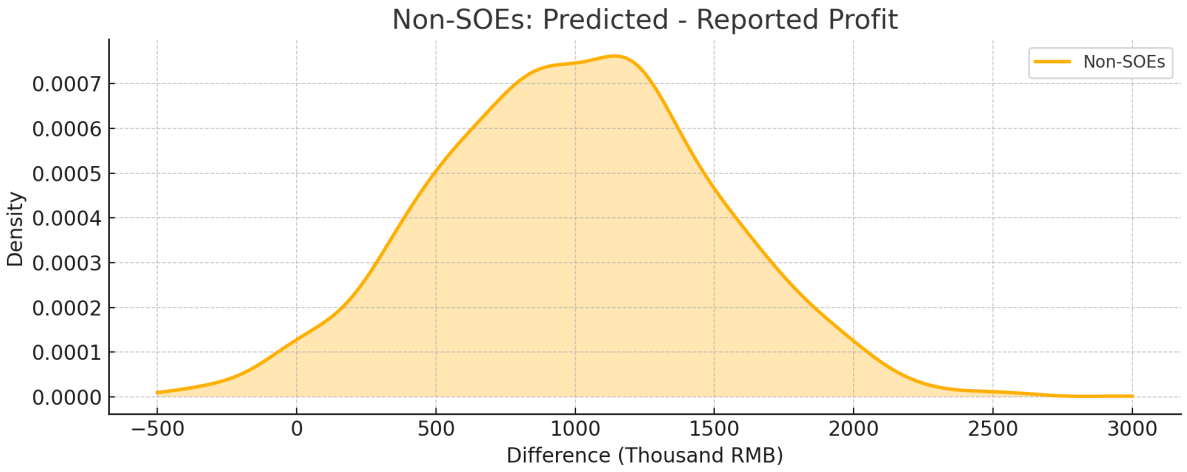


Figure 3: **Predicted profits - Reported profits** We plot the empirical density function of the estimated tax avoidance level, using all firm-/year observations of non-/state-/owned enterprises in the dataset.

We use the following indicator to measure the tax avoidance degree of an enterprise:

$$\frac{\text{predicted profit} - \text{reported profit}}{\text{total assets}}. \tag{12}$$

The larger this value, the further the reported profit deviates from the predicted profit, which indicates a higher degree of tax avoidance. We use total assets to scale the value, so the final tax avoidance indicator measures how much tax avoidance occurs for each unit of assets.

Figure 4 shows the measurement of tax avoidance degrees for state-owned enterprises, Class A enterprises, listed companies, and non-state-owned enterprises. State-owned enterprises are our training set, where the reported profit is close to the true profit, and the average tax avoidance degree is about 0.005. In the test set, non-state-owned Class A enterprises have the lowest tax avoidance degree, with an average of about 0.041, followed by non-state-owned listed companies, with an average of about 0.085. On the one hand, this demonstrates that the machine learning method used in this study has good predictive performance. Theoretically, Class A enterprises and listed companies in the test set should have relatively low tax

avoidance degrees, and indeed the tax avoidance indicators calculated in this study are also low. The remaining non-state-owned general enterprises have an average tax avoidance degree of about 0.217.

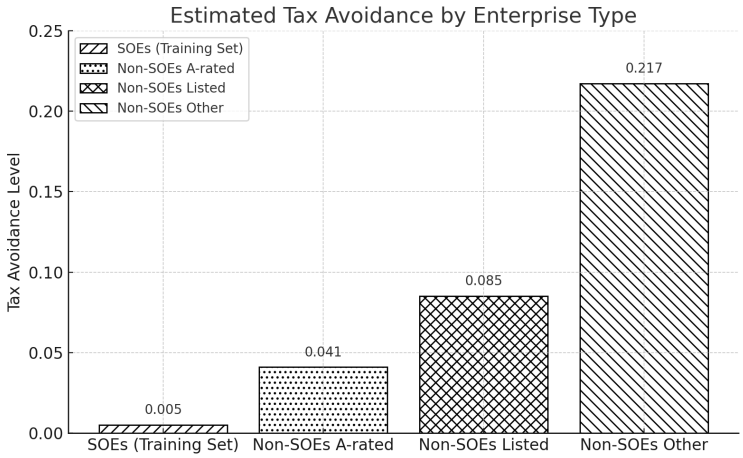


Figure 4: **Tax Avoidance Index** The figure reports the average level of tax avoidance across firms with different ownership types and tax credit ratings.

3.2 Can We Trust Our Results?

3.2.1 Inconsistency in the Profit Generation Process Between State-Owned and Non-State-Owned Enterprises

To address the issue that the profit generation process for state-owned and non-state-owned enterprises may not be the same, we conducted two robustness checks. First, we changed the output of the predictive model. In the baseline results, the reported (true) profit of state-owned enterprises[2] was used as the output in the training set. We change the predicted outcome variable to the company's reported (true) profit minus the imputed profits from the national accounts (imputed profits from Cai and Liu (2009)). We consider that if state-owned and non-state-owned enterprises have different production functions, this difference would be reflected in the imputed profit. By subtracting the imputed profits, the remaining part of the reported profit should be due to differences in the degree of tax avoidance among enterprises. By changing the model output, we can still obtain the predicted true profit of non-state-owned enterprises in the test set, which in turn allows us to measure the tax avoidance degree of non-state-owned enterprises.

To verify the robustness of the results, we use the following indicators to measure

[2]For state-owned enterprises, this is also an approximate value of the true profit.

the deviation between the tax avoidance degree obtained under the new method and the benchmark result:

$$\text{Deviation} = \frac{\text{TaxAvoidance}_{\text{robust}} - \text{TaxAvoidance}_{\text{baseline}}}{\text{TaxAvoidance}_{\text{baseline}}} \tag{13}$$

Figure 5 shows the deviation in the tax avoidance degree after changing the model output to reported profits minus imputed profits. To reduce the impact of erroneous data and outliers, the deviation plot removed the top and bottom 10% of outlier samples. The results show that the deviation is controlled within the range of [-0.9%, 0.8%], with most values clustering around 0. After changing the output, the tax avoidance degree of enterprises remains generally unchanged.

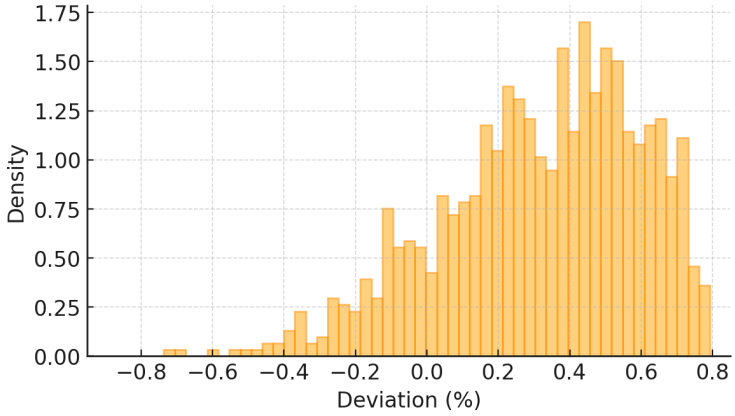


Figure 5: **Eliminating Potential Differences In The Profit-Generating Processes Between State-Owned And Non-State-Owned Enterprises.** The table reports the difference between the adjusted and original tax avoidance indicators. The adjusted indicator is constructed by interpolating between the predicted profits and the imputed profits, and then using the sum of the predicted value and the imputed profit as the new measure of tax avoidance.

The second discussion is about changing our training and testing sets. If we are concerned that the training set consists entirely of state-owned enterprises and the test set consists entirely of non-state-owned enterprises, leading to potential systematic differences in the profit generation functions of the two types of enterprises, we can include some non-state-owned enterprises in the training set. According to the results in Figure 4, the tax avoidance degree of Class A private enterprises is also very low, meaning their reported profits are very close to their true profits. Therefore, Class A enterprises can also be used as part of the training set. There are a large number of private enterprises within the Class A group, especially some smaller enterprises, so this can effectively address the issue of differing profit generation functions between the training and test sets. The figure shows the results after including

Class A enterprises in the training set, Figure 6 shows the deviation in the tax avoidance degree obtained from the new training set compared to the baseline results. The deviation is controlled within the range of [-0.13%, 0.2%], with most values clustering around -0.06%. The left skew is due to the fact that Class A enterprises still have an incentive to engage in tax avoidance because their profits are private. Including them in the training set leads to a lower bound with a lower tax avoidance degree.

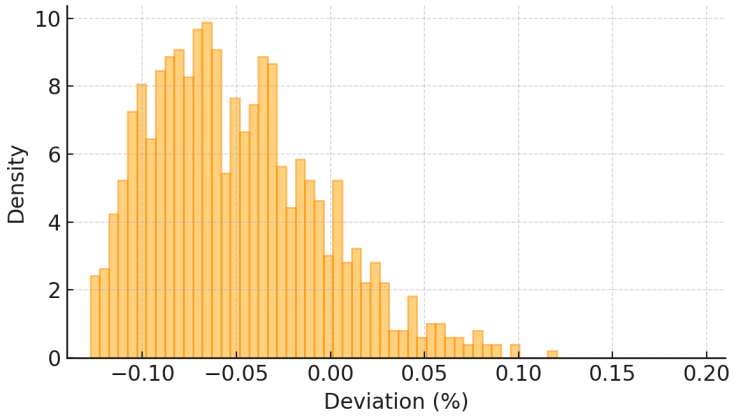


Figure 6: **Changing Training Set** We calculate the difference between the re-estimated results after adding Class A enterprises to training set and the original results.

3.2.2 Authenticity of the Input Predictive Factors

If a company engages in tax avoidance and adjusts profits, it will also adjust other variables in the financial statements. If the input features (X) are biased, it will similarly lead to estimation problems. When a company adjusts profits, it does not change financial indicators that are easily detectable as fraudulent. Instead, it only "corrects" indicators that are harder for the tax authorities to verify, in order to ensure the consistency of the entire financial statement. The indicators that are easier to "adjust" are mainly expense-related indicators. These indicators include many detailed items, and when expenses occur, they often do not generate easily verifiable third-party evidence (for example, when purchasing assets, an invoice is generated, and when incurring financial expenses, a voucher is produced by the bank). We remove these expense-related indicators to reduce the probability of errors in the prediction factors. Specifically, these include management expenses, sales expenses, R&D expenditures, outsourced R&D expenditures, intangible assets, accounts receivable, and cash flow items. The figure shows the deviation in the corporate tax avoidance results recalculated after removing these factors, compared to the baseline results. The results of Figure 7 show that in the training set, the prediction performance deteriorated, with R^2 dropping by about

8%. In terms of bias, the tax avoidance index derived from the new factor sequence is higher, and the deviation shows a right-skewed trend. This is because the removed factors may have been used as indicators for manipulating tax avoidance. After removing them, the predicted profits are higher, and consequently, the tax avoidance degree is also higher. However, overall, the deviation is controlled within the range of [-0.6%, 0.7%], indicating that the size of the tax avoidance indicator is robust.

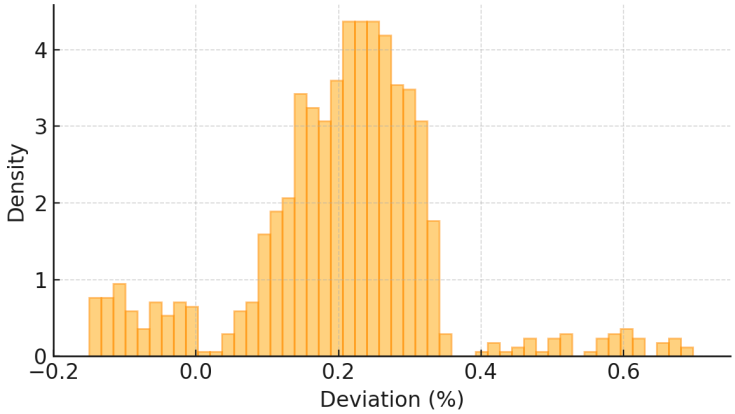


Figure 7: **Removing Easily Manipulated Explanatory Variables** We calculate the difference between the re-estimated results after removing easily manipulated variables such as selling expenses and the original results.

Additionally, we also examined the feature importance of the predictive factors. Figure 8 shows that in the sample of state-owned enterprises (SOEs), the feature importance of easily modifiable predictive factors (such as management expenses, sales expenses, R&D expenses, outsourced R&D expenses, intangible assets, accounts receivable, operating cash flow, investment cash flow, and financing cash flow) is very low, all in the range of [0, 0.04]. Figure B shows that in the SOE sample, the top 10 predictive factors with relatively high feature importance are: operating income, total assets, total liabilities, industry, city, employee salaries, interest expenses, number of employees, VAT paid, and fixed assets. These factors are all relatively difficult to modify. For the private enterprise (PE) sample, we directly trained the model again and recalculated the feature importance. The feature importance of the easily modifiable predictive factors remains very low. While these factors are very easily adjustable in private enterprises, their contribution to profits is minimal.

The third discussion applies a set of monthly tracking survey data from small and medium-sized enterprises (SMEs) in a relatively developed province of China (data Z). Using the same machine learning method, we are able to obtain the annual tax avoidance indicator for enterprises. However, since the data does not include state-owned enterprises, we use tax-rated Class A enterprises in the training set and non-Class A enterprises in the test

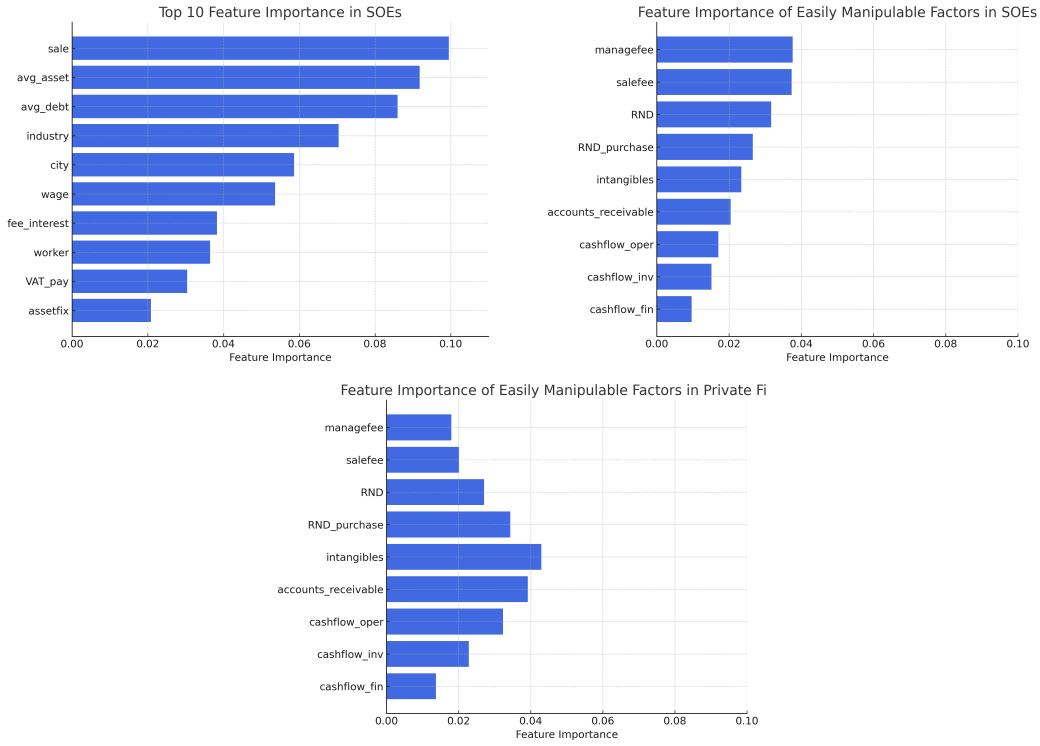


Figure 8: **Feature Importance** This figure compares the feature importance of easily manipulable input variables with that of other input variables.

set. We predict monthly profits and then aggregate them to obtain the annual predicted profits and reported profits, ultimately deriving the annual tax avoidance index. Since the enterprise samples in data Z and SAT cannot be matched one-to-one, we are unable to demonstrate whether the tax avoidance indicators of the same enterprise are consistent across both datasets. Instead, we can only present the overall situation. Figure 9 shows that the tax avoidance indicators of SMEs in SAT are quite similar to those of enterprises in data Z, with the mean tax avoidance indicator fluctuating around 0.10, and the median fluctuating around 0.08.

4 Application of the Tax Avoidance Indicator

Next, this article will validate the effectiveness of the tax avoidance indicator we have calculated. The term "effectiveness" refers to whether our tax avoidance indicator can truly reflect the extent of tax avoidance in reality. If our tax avoidance indicator is effective, it should have a significant correlation with factors that greatly influence the degree of tax avoidance in enterprises.

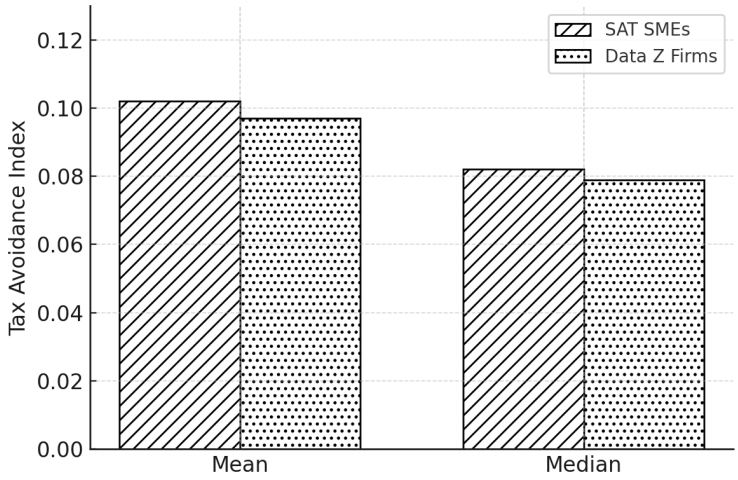


Figure 9: **Tax Avoidance Indicators Of SMEs In SAT and Data Z**

4.1 Can Tax Avoidance be Detected by the Tax Authorities?

We will examine the correlation between the level of tax avoidance and tax avoidance penalties. If a company's level of tax avoidance is higher, the probability of being penalized by the tax authorities should also be higher. The correlation coefficient between the two is 0.11, with a p-value close to 0. These two indicators are indeed correlated, but not all tax evasion cases are caught by the tax authorities.

4.2 The Classical A-S Tax Evasion Model

The classical tax evasion model developed by Allingham and Sandmo (1972) provides a theoretical foundation for analyzing optimal evasion decisions under uncertainty. Although originally proposed for individual taxpayers, its logic has been widely extended to corporate settings. In the baseline model, a risk-averse taxpayer with true income I chooses how much to report to the tax authority. Let $R \leqslant I$ be the reported income, and let t be the statutory tax rate. If the taxpayer is audited (with probability p), a penalty proportional to the evaded tax is imposed at rate $\theta > 0$. The expected utility function is:

$$EU(R) = (1-p) \cdot U(I - tR) + p \cdot U(I - tR - \theta t(I - R)), \tag{14}$$

where $U(\cdot)$ is strictly concave. The taxpayer chooses R to maximize expected utility. The model yields several key predictions:

- **Enforcement effect**: A higher audit probability p raises the expected cost of evasion,

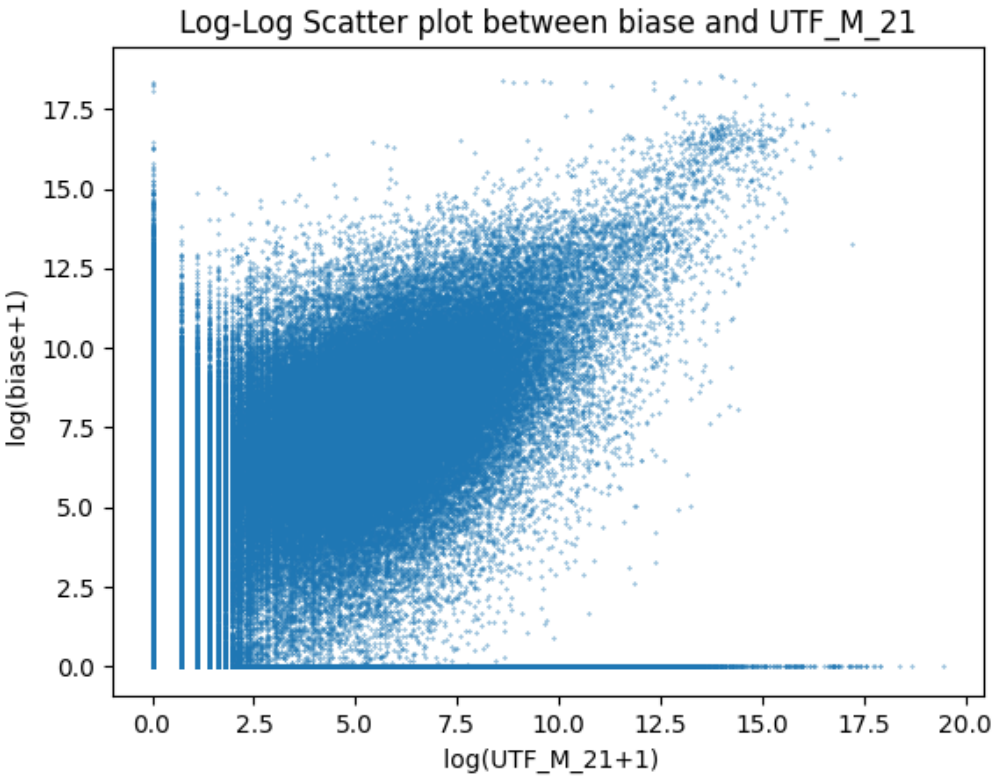


Figure 10: **Fines and the Level of Tax Avoidance** We report the scatter plot of fines and the level of tax avoidance.

leading to greater compliance.

- **Penalty effect**: A higher penalty rate θ increases the marginal cost of evasion if detected, discouraging non-compliance.
- **Tax rate effect (ambiguous)**: The effect of the statutory tax rate t is theoretically ambiguous. On the one hand, a higher t increases the return to evasion (substitution effect), encouraging concealment. On the other hand, it reduces after-tax income, increasing risk aversion and possibly reducing the willingness to evade (income effect). The overall direction depends on the taxpayer's degree of risk aversion and the curvature of the utility function.

When applied to firms, the model must account for financial frictions. In particular, firms facing credit constraints or high financing costs may rely more heavily on internal funds. Since tax evasion increases retained earnings by reducing tax outflows, it may act as an informal channel of self-financing. Let the firm's true profit be Π, and let $\tau \in [0, 1]$ denote the fraction of profit not reported. Then the reported profit is $R = (1 - \tau)\Pi$. If the firm requires capital K for investment and must obtain external financing F, its budget constraint becomes:

$$K = (1 - t)(1 - \tau)\Pi + \tau t\Pi + F, \tag{15}$$

where $\tau t\Pi$ reflects the cash saved through evasion. The firm minimizes the total cost, which includes taxes, expected penalties, and financing costs $c(F)$:

$$\min_{\tau, F} \quad t(1 - \tau)\Pi + p \cdot \theta t\tau\Pi + c(F), \tag{16}$$

$$\text{subject to} \quad (1 - t)(1 - \tau)\Pi + \tau t\Pi + F \geqslant K. \tag{17}$$

This setup shows that greater evasion reduces the need for costly external financing. Therefore, firms facing tighter financial constraints have stronger incentives to evade taxes, especially when the expected penalty is low or enforcement is weak. In summary, the A-S model highlights how enforcement intensity and penalty severity shape tax compliance, while the effect of the tax rate is theoretically ambiguous due to opposing substitution and income effects. When extended to the corporate setting, financing constraints emerge as a critical determinant of evasion behavior, with tax evasion acting as a potential substitute for costly external capital.

4.3 Theoretical Validation of Our New Measurement

According to the classic A-S tax evasion model, the factors affecting the extent of corporate tax evasion can be divided into two categories. One category is the benefits brought by tax evasion itself, and the other category is the intensity of tax administration. The benefits of tax evasion for a company are influenced by the following factors: (1) the statutory tax rate of the company (Taxrate); the higher the tax rate, the higher the level of tax evasion, because the benefits of tax evasion are greater. The statutory tax rate is determined according to the company's size, industry, and the "Corporate Income Tax Law" as well as relevant regulations and policies.[3] (2) Industry competition level (Compet), according to the conclusion of Cai and Liu (2009), the more intense the competition, the higher the level of tax avoidance. (3) Corporate financing constraints indicator (Cash), represented by the company's operating net cash flow (divided by total assets); generally, the more sufficient the cash flow, the lower the level of tax avoidance. (4) Tax incentives enjoyed by the company (Reduce). If the company has already enjoyed sufficient tax incentives, its level of tax avoidance should also be lower, represented by the ratio of the company's enjoyed income tax incentives to the tax base.

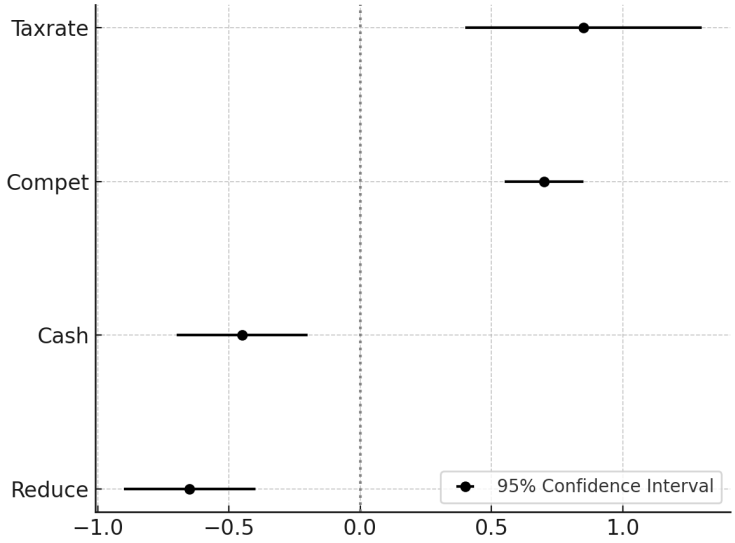


Figure 11: **Theoretical Validation** The figure presents the confidence intervals for the correlation coefficients between the predicted level of tax avoidance and the variables from the classic Allingham–Sandmo tax-evasion model.

The variables influencing the intensity of tax administration are mainly related to the behavior of local governments, including: (1) the number of staff in the tax authority; the greater the number, the lower the level of tax avoidance, as the tax administration intensity is

[3]The relevant regulations and policies come from the Peking University Legal Information Network.

higher. (2) Whether the local city's administrative head has just taken office; newly appointed administrative heads tend to have a lower tolerance for tax avoidance. (3) The fiscal pressure of the local area, as measured by the revenue-expenditure gap; the greater the fiscal pressure, the stronger the motivation for tax administration, and the lower the level of corporate tax avoidance. (4) The local government's share of tax revenue; if the local government receives a larger share, it has a stronger motivation for tax administration, and the level of corporate tax avoidance is lower. (5) The amount of electronic information tax administration systems purchased by the government, which indicates the level of government tax digitalization; the higher the level, the stronger the tax administration capability.

The regression model uses simple linear regression, with the dependent variable being the level of corporate tax avoidance and the independent variables being several influencing factors, controlling for industry, region, and time fixed effects. The results remain valid if the regression form of Cai and Liu (2009) is used.

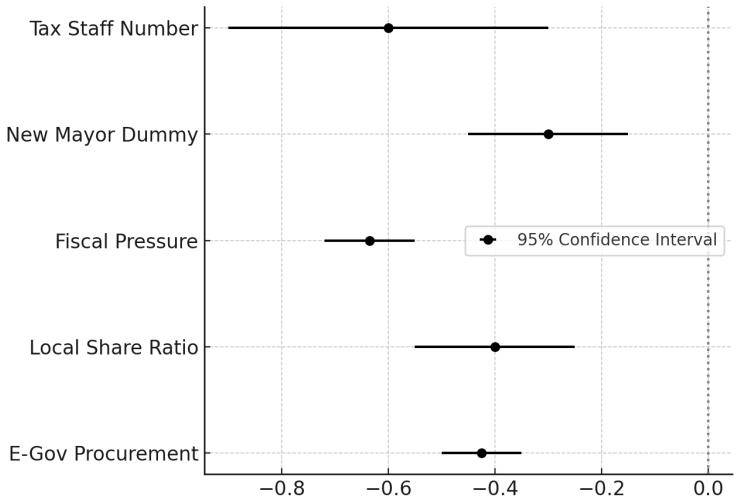


Figure 12: **Theoretical Validation** The figure presents the confidence intervals for the correlation coefficients between the predicted level of tax avoidance and the variables from the classic Allingham–Sandmo tax-evasion model.

4.4 Who Evades More?

The extent to which large firms or small and medium-sized enterprises (SMEs) engage in greater tax evasion remains a subject of empirical debate, with existing studies offering conflicting conclusions. Our analysis suggests that, on average, tax evasion increases with firm size, but peaks among mid-sized firms.

To investigate this relationship, we rank firms by size percentiles and compute the average

value of the tax evasion indicator across different size groups. Firm size is measured using three alternative proxies: total assets, number of employees, and profitability. All three measures yield consistent results. Specifically, firms in the middle of the size distribution (i.e., between the 40th and 80th percentiles) exhibit the highest levels of tax evasion, while both small firms (below the 40th percentile) and large firms (above the 80th percentile) show lower levels of evasion.

This pattern can be explained by institutional characteristics of China's tax system. Small firms benefit from preferential tax policies, with effective tax rates of around 10%, which substantially reduce their incentive to evade taxes. Although large firms face higher statutory tax rates (around 25%), they are also subject to strict tax enforcement. In practice, Chinese tax authorities adopt a "focus-on-large, neglect-small" enforcement strategy, under which large firms are labeled as "key tax source enterprises" and are closely monitored and frequently audited. This intensive scrutiny limits their ability to evade taxes.

In contrast, mid-sized firms face the same high statutory tax rates as large firms but are not subject to the same level of enforcement. The combination of a heavy tax burden and relatively lax oversight provides both the incentive and the opportunity to evade taxes, making mid-sized firms the most aggressive evaders in the firm size distribution.

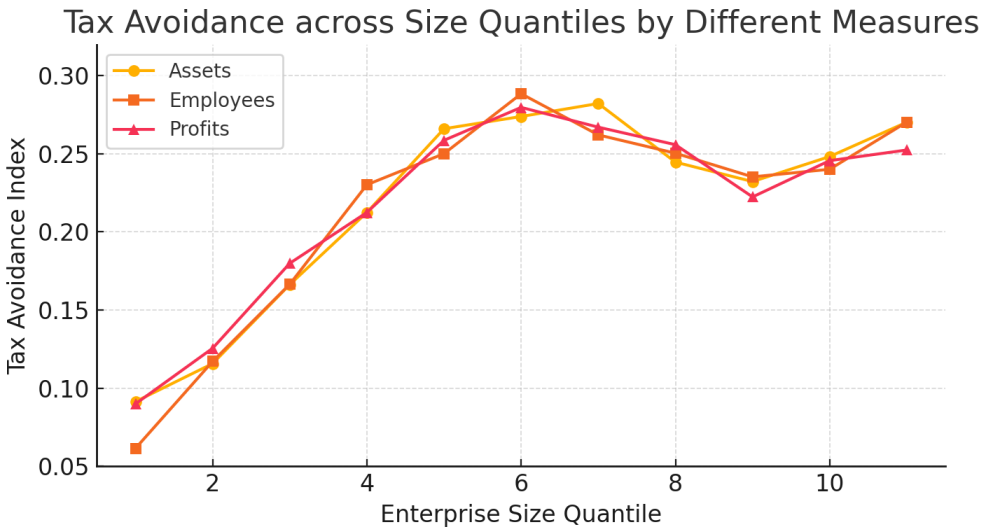


Figure 13: **Tax Avoidance Across Size Quantiles** We divide firms into ten groups—separately by asset size, number of employees, and corporate profits—and calculate the average level of tax avoidance within each group.

4.5 Validation Using the Policy Shock: The Golden Tax System Phase III (GTS-III) in China

In recent years, the Chinese government has undertaken a comprehensive effort to digitize its tax administration system as part of broader state capacity modernization. A centerpiece of this effort is the Golden Tax System Phase III (GTS-III), officially launched in 2016. As the latest and most advanced iteration of the Golden Tax Project, GTS-III represents a significant technological upgrade over its predecessors. The system aims to enhance the real-time collection, integration, and analysis of tax-related data from multiple sources, including tax filings, financial statements, invoice systems, customs records, and banking transactions.

GTS-III is designed to centralize information across local tax bureaus and other government departments through a unified national database. It employs automated algorithms to cross-check corporate financial data, detect inconsistencies, and flag firms for audit based on risk scores. Compared to previous phases, GTS-III features greater interoperability, more stringent control over value-added tax (VAT) invoices, and increased enforcement precision. The system also facilitates data sharing between tax authorities and regulatory bodies such as the State Administration for Market Regulation and the People's Bank of China.

By reducing information asymmetry between firms and tax authorities, GTS-III significantly strengthens the government's enforcement capacity and reduces the space for tax evasion. It marks a shift from reactive to proactive tax administration, enabling real-time monitoring and automated risk targeting. As such, the rollout of GTS-III is widely regarded as a watershed moment in China's tax enforcement regime, with particularly strong implications for firms with historically low transparency or weak internal controls.

Following the implementation of the Golden Tax System Phase III (GTS-III), the scale of tax evasion among Chinese firms is expected to decline substantially. Our empirical analysis confirms this expectation by detecting a significant reduction in corporate tax evasion after the rollout of GTS-III. We employ a difference-in-differences (DID) model to estimate the policy effect:

$$\text{Evasion}_{it} = \alpha + \beta \cdot \text{Treat}_{it} + \gamma X_{it} + \mu_i + \lambda_t + \varepsilon_{it} \tag{18}$$

Here, Treat_{it} is a binary indicator equal to 1 if firm i is located in a region that had implemented GTS-III by year t, and 0 otherwise. The coefficient β captures the average treatment effect of GTS-III on firm-level tax evasion.

Our results indicate that GTS-III significantly reduced tax evasion among private firms. Based on our estimates, the average reduction in tax evasion per firm is approximately RMB 10,675. Given that there are roughly 36 million private enterprises in China, the aggregate increase in tax revenue amounts to approximately RMB 384.3 billion. Compared to the estimated implementation cost of GTS-III—around RMB 100 billion—the policy yields a net fiscal gain of over RMB 280 billion, suggesting that its benefits significantly outweigh its costs.

Table 2: **GTS-III DID** We present the results of a difference-/in-/differences (DiD) regression that exploits the introduction of *Golden Tax System Phase III* as the policy shock, with the tax-avoidance indicator serving as the dependent variable.

DID method						
$treat_{i,t}$	-0.000802***	-0.000802***	-0.000824***	-0.000820***	-0.000764***	-0.000790***
	(43.12)	(44.21)	(38.08)	(38.06)	(37.30)	(38.02)
N	5631830	5631680	5195560	5217140	5217080	5217120
R-sq	0.137	0.158	0.651 0.680	0.683	0.683	
EJ method						
$treat_{i,t} \times pro_{i,t}$	0.00737***	0.00738***	0.00742***	0.00748***	0.00767***	0.00761***
	(61.61)	(61.88)	(55.34)	(50.16)	(50.25)	(51.19)
N	5631830	5631680	5195560	5217140	5217080	5217120
R-sq	0.026	0.029	0.274	0.286	0.287	0.288

5 Conclusion

This paper uses automated machine learning to train on state-owned enterprises (SOEs) in China and infers the tax avoidance level of private enterprises in China. This method has significant advantages over previous methods. In earlier literature, the measurement of tax avoidance for non-listed companies only provided a relative tax avoidance level and could not quantify the absolute tax avoidance amount for a specific point in time, for a specific company or type of company. Our method can measure the absolute tax avoidance level of non-listed companies in China.

This paper considers a series of potential challenges. We address the issue of the profit generation process between SOEs and private enterprises by using imputed profit, and demonstrate that adding A-rated private enterprises with a tax rating to the training sample has a minimal impact on the results. Regarding the issue of systematic financial fraud by enterprises, we show that removing variables prone to manipulation, such as management expenses and sales reception expenses, has a limited impact on the results, as the model assigns these indicators low feature importance.

The tax avoidance indicator in this paper has good economic properties. We verified several parameters in the Allingham-Sandmo (A-S) tax evasion model that influence the extent of corporate tax evasion and found that our tax avoidance indicator can be validated by the A-S model. This indicator can be used to statistically measure and evaluate some significant economic facts about private enterprises in China. The paper finds that the tax avoidance level is highest for medium-sized enterprises, which is consistent with China's policy of "grabbing the big and releasing the small" and the low tax burden policy for micro-enterprises. We also find that the cost of digital transformation of government tax administration is smaller than the benefits derived from curbing corporate tax avoidance, making it a highly cost-effective approach.

References

Bao, Y., B. Ke, B. Li, Y. J. Yu, and J. Zhang. 2020. Detecting accounting fraud in publicly traded us firms using a machine learning approach. *Journal of Accounting Research* 58:199–235.

Benguria, F., and F. Saffie. 2025. Beyond tariffs: How did china's state–owned enterprises shape the us–china trade war? Working Paper, National Bureau of Economic Research.

Cai, H., and Q. Liu. 2009. Competition and corporate tax avoidance: Evidence from chinese industrial firms. *The Economic Journal* 119:764–95.

Cen, L., E. L. Maydew, L. Zhang, and L. Zuo. 2017. Customer–supplier relationships and corporate tax avoidance. *Journal of Financial Economics* 123:377–94.

Chen, S., Y. Huang, N. Li, and T. Shevlin. 2019. How does quasi-indexer ownership affect corporate tax planning? *Journal of Accounting and Economics* 67:278–96.

Chen, T., Y. Tan, J. Wang, and C. Zeng. 2022. The unintended consequence of land finance: Evidence from corporate tax avoidance. *Management Science* 68:8319–42.

Cheng, C. A., H. H. Huang, Y. Li, and J. Stanfield. 2012. The effect of hedge fund activism on corporate tax avoidance. *The accounting review* 87:1493–526.

Desai, M. A., and D. Dharmapala. 2006. Corporate tax avoidance and high-powered incentives. *Journal of financial Economics* 79:145–79.

Desai, M. A., C. F. Foley, and J. R. Hines Jr. 2006. The demand for tax haven operations. *Journal of Public economics* 90:513–31.

Dyreng, S. D., M. Hanlon, and E. L. Maydew. 2008. Long-run corporate tax avoidance. *The Accounting Review* 83:61–82.

Hasan, M. B., M. Mahi, M. K. Hassan, and A. B. Bhuiyan. 2021. Impact of covid-19 pandemic on stock markets: Conventional vs. islamic indices using wavelet-based multi-timescales analysis. *The North American Journal of Economics and Finance* 58:101504–.

Murorunkwere, B. F., O. Tuyishimire, D. Haughton, and J. Nzabanita. 2022. Fraud detection using neural networks: A case study of income tax. *Future Internet* 14:168–.

Shaheer, N., J. Yi, S. Li, and L. Chen. 2019. State-owned enterprises as bribe payers: The role of institutional environment. *Journal of Business Ethics* 159:221–38.

Tan, J., and L. Wang. 2010. Flexibility–efficiency tradeoff and performance implications among chinese soes. *Journal of business research* 63:356–62.

Xiang, J., H. Zhang, D. Dang, and J. Guan. 2023. Involuntary political connections and private firms' tax avoidance. *Pacific-Basin Finance Journal* 79:102025–.