Article

Identifying tax evasion in shell companies and fraudulent credits of Brazil's state value-added tax (ICMS)

Gunther Siqueira Lemos Gomes [1]

Remis Balaniuk [2]

[1] Secretaria de Estado de Economia do Distrito Federal, Brasília / DF – Brazil
[2] Universidade Católica de Brasília (UCB), Brasília / DF – Brazil

Companies that issue tax documents to defraud the tax authorities with the transfer of credits of Brazil's state value-added tax (ICMS) without the movement of goods cause financial losses to the government and, therefore, to society as a whole. Several initiatives to combat tax fraud have successfully used data analysis and Machine Learning techniques. This work sought to investigate the use of these techniques in identifying a specific practice of tax fraud, practiced by shell companies, formed exclusively to issue non-due ICMS credits, the tax on operations related to the circulation of goods, and the provision of interstate, intercity, and communication services. Based on document analysis and consultation with auditors and specialists, typologies and variables relevant to identifying tax evasion events carried out by shell companies were identified. Around these variables, data from the Finance Department of the Federal District were collected and prepared. With this data, it was possible to explore the use of predictive models based on Machine Learning capable of pointing out potentially fraudulent behavior. The good results obtained by these models demonstrate their potential as part of systematic monitoring and fiscal audits by tax authorities.

Keywords: ICMS; machine learning; shell companies; tax evasion.

Identificando evasão fiscal em empresas de fachada e em créditos ilegais de ICMS

Empresas que emitem documentos fiscais para fraudar o fisco com a transferência de crédito do ICMS sem a circulação de mercadorias causam prejuízo ao erário público e, por conseguinte, à sociedade. Diversas iniciativas de combate a fraudes fiscais têm utilizado, com sucesso, técnicas de análise de dados e aprendizagem de máquina. Este trabalho buscou investigar o uso dessas técnicas na identificação de uma prática específica de fraude fiscal realizada por empresas popularmente conhecidas como "empresas noteiras", que formadas exclusivamente para emitir créditos não devidos de ICMS, imposto sobre operações relativas à circulação de mercadorias e sobre prestações de serviços de transporte interestadual, intermunicipal e de comunicação. Com base na análise documental e em consulta com auditores e especialistas, foram identificadas tipologias e variáveis relevantes na determinação de eventos de sonegação fiscal realizados pelas empresas noteiras. Em torno dessas variáveis, procedeu-se à coleta e à preparação de dados provenientes da Secretaria de Fazenda do Distrito Federal. Com esses dados, foi possível

ISSN: 1982-3134

DOI: https://doi.org/10.1590/0034-761220230303x
Article received on September 20, 2023 and accepted on June 07, 2024.
[Translated version] Note: All quotes in English translated by this article's translator.
Editor-in-chief:
Alketa Peci (Fundação Getulio Vargas, Rio de Janeiro / RJ – Brazil)
Associate editor:
Sandro Cabral (Insper Instituto de Ensino e Pesquisa, São Paulo / SP – Brazil)
Reviewers:
João Paulo de Oliveira Louzano (Universidade Federal dos Vales do Jequitinhonha e Mucuri, Teófilo Otoni / MG – Brazil)
Ana Carolina Giuberti (Universidade Federal do Espírito Santo, Vitória / ES – Brazil)
Peer review report: the peer review report is available at this URL.

explorar o uso de modelos preditivos baseados em aprendizagem de máquina capazes de apontar comportamentos potencialmente fraudulentos. Os bons resultados obtidos por esses modelos demonstram seu potencial como parte de uma sistemática de monitoramento e auditorias fiscais realizadas pelos órgãos fazendários.

Palavras-chave: ICMS; noteiras; aprendizagem de máquina; empresas de fachada; sonegação fiscal.

Identificación de evasión tributaria en empresas fachada y créditos ilegales del ICMS

Las empresas que emiten documentos tributarios para defraudar al fisco con la transferencia de crédito del ICMS (impuesto a las operaciones relacionadas con la circulación. de bienes y de prestación de servicios interestatales, interurbanos y de comunicaciones) sin movimiento de mercancías causan daños al erario público y, por ende, a la sociedad en su conjunto. Varias iniciativas para combatir el fraude fiscal han utilizado con éxito técnicas de análisis de datos y aprendizaje automático. Este trabajo buscó investigar el uso de estas técnicas en la identificación de una práctica específica de fraude fiscal, practicada por empresas conocidas popularmente como 'empresas factureras', constituidas exclusivamente para emitir créditos no vencidos del ICMS. A partir del análisis documental y la consulta a auditores y especialistas, se identificaron tipologías y variables relevantes para la identificación de eventos de evasión fiscal realizados por empresas factureras. En torno a estas variables se recolectaron y prepararon datos desde la Secretaría de Hacienda del Distrito Federal. Con estos datos fue posible explorar el uso de modelos predictivos basados en *machine learning* capaces de señalar comportamientos potencialmente fraudulentos. Los buenos resultados obtenidos por estos modelos demuestran su potencial como parte de un seguimiento sistemático y auditorías fiscales por parte de las autoridades tributarias.

Palabras clave: ICMS; factureras; aprendizaje automático; empresas de fachada; evasión de impuestos.

1. INTRODUCTION

In the Brazilian economic and political scenario, society's demands for better services and infrastructure, coupled with the perception of an already excessively high tax burden, remain central themes of discussion and pressure. In a context of increasing civic awareness and social participation, Brazilian citizens have been demanding, in an increasingly emphatic manner, solutions to chronic problems such as the lack of investments in health, education, transportation, and security, while also clamoring for relief from the tax burdens that impact their lives and the business environment. The expectation for efficient and transparent public policies that promote sustainable development and social inclusion, while seeking to alleviate the weight of taxes on the population and businesses, has become evident, requiring agile and innovative responses from government authorities.

Contrary to these demands and expectations, in a scenario of high tax burden, the public administration sees revenue losses due to tax fraud gaining ground. It is observed that the greater the gain in not paying taxes, the greater is also the taxpayer's inclination to evade them.

At the heart of the fiscal and tax issue are indirect taxes, which consist of consumption taxes included in the prices of all goods and services. In 2015, these taxes represented 49.4% of the Union's gross tax burden, a level far from that practiced in developed countries (Souza, 2018). Examples of indirect taxes are the Tax on Industrialized Products (IPI) and the Tax on Circulation of Goods and Services (ICMS). About 82% of the tax revenue of the states and the Federal District, in 2020, were based on the ICMS, which, along with Income Tax, were the most evaded taxes that year (Santos et al., 2021). The ICMS is also of particular relevance to small municipalities that depend on the transfer of resources through the constitutional distribution of taxes (Azevedo et al., 2015).

Due to their breadth and materiality, fraudulent schemes in the collection of ICMS often lead to billion-dollar losses to the public coffers. In just one operation carried out by the tax inspection of the state of Paraná, in 2022, 844 shell companies created exclusively to evade taxes were identified. These companies issued invoices for operations that totaled R$ 4.8 billion, resulting in the evasion of R$ 542.8 million in ICMS (Ortiz, 2022).

This research sought to identify a specific type of fraud associated with ICMS, characterized by the use of simulated sales operations of goods practiced by so-called "shell companies": companies created with the purpose of operating tax fraud, mainly through the issuance of false invoices (Carvalho, 2018). It involves the issuance of a fiscal document without the seller – the issuer of the invoice – transacting the merchandise being sold, indicating that there is neither physical nor legal circulation (exchange of ownership) of it, and without the payment of the due tax, that is, it is just a fictitious operation. These are the so-called "cold invoices". For Ferreira (2019), shell companies are also known as "shell companies", since they are created only to provide a legal appearance to a fraudulent transaction.

The mechanism of trading credits contained in cold invoices is explained because ICMS has the characteristic of non-cumulativity, a situation in which a merchant, when acquiring a good for resale, has the right to use the ICMS paid by its supplier as credit, paying to the state in which it is registered only the difference or the added value of the tax in the next sale operation of the same merchandise. In this scheme, the shell company issues the cold invoice, without, obviously, paying the due ICMS in its state, and sells the invoice for a fraction of the value of the tax due to a recipient company – typically from another federation unit – that wants to legalize its own sales without paying the tax in its state.

The main difficulty in identifying such fraud is that the information provided in the issued fiscal documents is, apparently, from normal operations and, to further hinder the action of the tax authorities, the evaders, aware of the difficulty of communication between federated entities, simulate interstate sales, where the issuer and the recipient of the invoice are from different states.

As the identification of these companies and their fraudulent practice requires time and effort, the shell companies end up benefiting from this delay, generating significant amounts of ICMS credits, taken advantage of by their recipients, until they are discovered.

Allingham and Sandmo (1972) emphasize that companies work under a risk limit when evading taxes, and that applying fines is one of the main ways to increase the sense of risk for the taxpayer and prevent evasion.

The experiment conducted by Kleven et al. (2011), in Denmark, which corroborates the studies presented by Allinghan and Sandmo, revealed that the taxpayer who undergoes a tax audit tends to respond better to spontaneous collection. However, Lima (2007) highlights that the regulatory bodies do not have the capacity to inspect all taxpayers. In this context, it is verified that, to imprint a more effective sense of risk in a context of a high number of taxpayers and, consequently, operations to be inspected, it is necessary to use automated methods for large-scale fraud detection, in order to make such process more comprehensive, agile, and with lower cost.

These automated methods obviously need to use robust computational platforms, large databases, and analytical models capable of processing the enormous amounts of daily transactions and identifying atypical behaviors with indications of tax fraud. Therefore, it is a typical application for data science, a multidisciplinary area that uses computational, statistical, and mathematical techniques to solve complex problems and has been used in the most diverse domains to extract meaningful information for businesses and decision-making. In tax inspection, the use of techniques derived from data science has been evolving significantly in recent decades and has been producing results on various fronts and countries (Abrantes & Ferraz, 2016).

In this context, the objective of the present work was to explore the potential of data science methods and machine learning as a basis for a systematic selection of indications of simulation in the issuance of fiscal documents, seeking to circumvent taxation by ICMS through shell companies.

As a differential with respect to existing works in the literature on the topic of tax fraud, this study seeks to contribute with an analysis strategy suitable to the Brazilian tax reality, the particularities of ICMS, and the fraudulent practices practiced in the federation units at present. Moreover, the research is based on the business knowledge of the tax auditors themselves and explores practices already adopted by the state treasury departments, such as the automatic denial of invoices as input for the preparation of predictive models.

The methodology used was CRISP-DM (Chapman, 2000), a reference in data mining projects. The entire understanding of the business, definition of assumptions, and choice of data used in the analysis were carried out with the participation of tax auditors. For the conception of predictive models, only databases typically available in state treasury departments were used, namely: electronic fiscal documents, specifically the Electronic Invoice (NFe), corporate and registration data of companies, and NFe events linked to the automatic denial of invoices. Additionally, data from the Electronic Transportation Knowledge (CTe), Electronic Consumer Invoice (NFCe), Digital Fiscal Bookkeeping (EFD), and Declaration of Payment Methods (Dimp) were used. The data used cover the period between 2014 and 2023 obtained from the Treasury Department of the Federal District.

This article is composed of five sections. In the first, we brought this introduction; in the second, we will present a review of the literature on the topic, in addition to some concepts necessary for understanding the context in which the research unfolded; in the third, we will detail the research methodology; in the fourth section, we will present the results of the research stages; finally, in the fifth and last section, we will bring some concluding remarks.

2. LITERATURE REVIEW

Castellón-González and Velásquez (2013, as cited in De Roux et al., 2018) describe how the selection of companies for tax inspection is traditionally based on rules and the use of intuition by the tax agent. This scenario is similar to that described by Zumaya et al. (2021), who report that traditional manual taxpayer selection techniques for inspection, without the support of computational methods, have limited effectiveness, as various companies do not appear in conventional filters and are not inspected.

With the advancement of the digitalization of fiscal documents, taxpayer registration information, corporate structures of companies, and the record of commercial and financial transactions, huge official databases are being fed daily and are available to regulatory bodies. Dramatic advances in the technological landscape have opened new possibilities for the use of these data and have brought data science into the reality of tax inspection. In recent decades, various initiatives, which proposed the application of a significant diversity of techniques and produced remarkable results, demonstrated that the automation of tax inspection is an irreversible path (Abrantes & Ferraz, 2016).

The literature on the subject is abundant, but some works with an applied focus and based on real cases, such as the study presented here, deserve special mention.

The Tax Administration Diagnostic Assessment Tool (Tadat, 2019), sponsored by the European Union and the International Monetary Fund, among other institutions, points to the use of information on a large scale with massive cross-referencing of information as one of the good practices in combating evasion, improving the tax authority's ability to identify deviations and increase the efficiency of inspection. One of the examples cited is the use of so-called tax meshes, in which the tax authority points out indications of fiscal irregularities for the taxpayer to rectify spontaneously. Therefore, although the practice is not aimed at auditing, it demonstrates the importance of data science in tax inspection.

Matos (2019) raises the importance of considering other information during the selection of audit targets, such as the registration characteristics of taxpayers (features) but highlights the difficult interpretation of these data without the help of machine intelligence, as they have low linear correlations, and frauds use complex strategies that are often not directly identifiable in the available data. The author suggests the technique of graphs as a way to interconnect such characteristics and presents a method of feature selection based on association rules and propositional logic.

Bittencourt (2018) worked with fiscal data from the base of the Treasury Department of the Federal District, the same primary base of this dissertation. His goal was the study of outliers to identify monthly fiscal periods in which taxpayers had a behavior of evasion. The author suggests the use of the Neural Networks method for predictive purposes to determine behaviors based on fiscal registration data, which is proposed in this work for the specific fraud involving shell companies.

Ippolito and Lozano (2020), in turn, demonstrate the applicability of a variety of techniques, such as Random Forests, Naive Bayes, Decision Tree, Logistic Regression, Ensemble Learning, and Neural Networks, in identifying crimes against the tax authority in the city of São Paulo.

Andrade et al. (2021) used data from the Treasury Department of the State of Espírito Santo for fraud detection and also different techniques for the same purpose. For the tests, the algorithms K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), and Neural Network (NN) were used.

Castellón-González and Velásquez (2013) conducted a study for the Tax Administration of Chile, in which they divide taxpayers into clusters by size, using algorithms such as Self Organizing Maps (SOM), Neural Gas (NG), and Decision Trees, and then work with a supervised algorithm for the solution of the fraud classification problem.

Xavier et al. (2022) applied Random Forests, Neural Networks, and Graphs to identify the profile of potential evaders, using only open and public data made available by the Brazilian Federal Revenue, the Administrative Council of Tax Appeals of the State of Goiás, and other public sources.

Finally, Zumaya et al. (2021) tackled fiscal fraud in Mexico and proposed the identification of frauds using Neural Networks and Random Forest. Specifically, regarding the Neural Network, Zumaya et al. (2021) used Dynamic Recurrent Neuronal Network (DRNN) to consider the fiscal documents issued in a distinct manner, through time series.

It is concluded that there is a significant diversity of data science and machine learning techniques applied in tax inspection, especially in fraud detection. Various of these techniques were tested and some, effectively, used to structure the method chosen by this study. However, the work proposed here distinguishes itself from others already published by specifically attacking the scheme of shell companies and cold invoices in fraud against ICMS. Moreover, our technique uses the business knowledge of Brazilian tax auditors and data from the automatic denial of fiscal invoices in the preparation of predictive models.

3. FOUNDATIONAL CONCEPTS ON THE DATA USED IN THE STUDY

Brazil adopted the Public Digital Bookkeeping System in 2007. This system consists of various subprojects, with emphasis on tax registration, fiscal and accounting bookkeeping, and electronic fiscal documents, such as the Electronic Invoice (NFe), the Electronic Consumer Invoice (NFCe), and the Electronic Service Invoice (NFSe).

For the present study, the Tax Registration and the Electronic Invoice are of great relevance. The Tax Registration contains registration information provided by legal entities to state tax authorities, such as address, accounting responsible, share capital, classification in economic activities in a standardized manner (CNAE-Fiscal), and corporate composition.

Each fiscal document has a specific purpose, and the Electronic Invoice (NFe) is the competent document to record an interstate merchandise sale operation or an operation between tax contributors. Thus, when the operation involves ICMS credit, the document to record this transaction is an NFe, and the issuance of other documents is prohibited. Rural producers constitute an exception to this rule, as they can issue the Rural Producer Invoice, which is still sent through forms and does not have an electronic version.

The Tax on Circulation of Goods (ICMS) is state-level and has a constitutional characteristic of being non-cumulative, that is, it compensates for the amount already paid. Shell companies rely on this aspect of the ICMS to offer an advantage to the beneficiary of the scheme. The purchaser of a cold invoice, issued by an invoice company, makes use of this tax to compensate, thereby reducing the tax due. The invoice company, being a shell company, does not collect the tax to any state tax authority, reducing the effective taxation of the products.

A relevant event for this study, which served as a starting point for data annotation and the creation of bases for training and testing predictive models, is the denial of NFes, practiced by most state treasury departments. The denied NFe occurs when an electronic invoice is issued, but the Treasury Department (Sefaz) identifies some problem on the part of the issuer or the recipient. Saying that the electronic invoice was denied means that the Sefaz identified some suspicion of irregularity from the issuer or the recipient of the NFe, and it cannot be invoiced/authorized until it is audited, and this irregularity is associated with the practice of shell companies or simply a registration problem. Through automatic denial, the treasury departments manage to stop the action of suspicious companies even before fraudulent actions are confirmed. Companies whose invoices have been denied are audited and can appeal the denial. In the end, the tax auditor records the standardized reason for the NFe block and whether the suspicions of irregularities were confirmed. Thus, the record of denials provides a rich base of examples of audited companies, some of which were identified as ghost or shell companies and others were confirmed as legitimate and non-fraudulent companies. Bases labeled in this way greatly facilitate the implementation of predictive analytical models based on supervised learning.

4. METHOD

To address the specific problem of tax evasion operated by shell companies, a quantitative and applied methodology was employed, based on fiscal data typical of the tax practices of the federation units. Thus, an exploratory research approach was chosen, considering that the phenomenon, although known to tax auditors, lacks systematization and automation when identifying the fraudulent taxpayer.

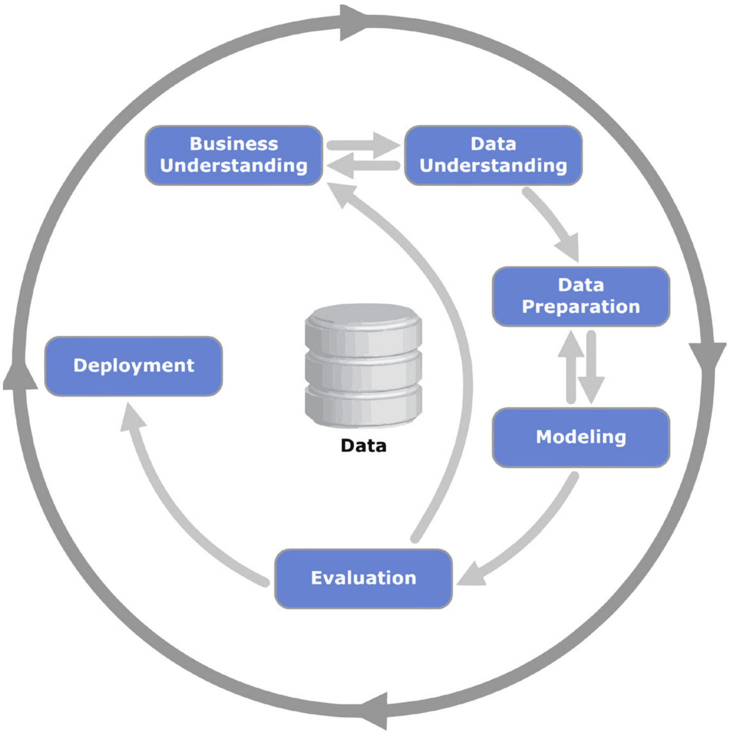
The research was based on documentary fiscal data in possession of the Treasury Department of the Federal District, but can, in principle, be applied in other federation units, since similar data are used by all state treasury departments. In this work, techniques and methods from statistical and data sciences were utilized.

The tasks performed during the study were:

- Mapping the vulnerabilities of the taxation and ICMS collection system exploited by shell companies in the main tax evasion schemes;
- Identification of relevant variables for the evidence of fiscal fraud based on the digital data of electronic fiscal documents contained in the database of the Treasury Department of the Federal District;
- Selection of statistically relevant variables to be used in predictive models;
- Implementation of an automated process for loading and transforming raw attributes from databases to derive the selected analytical variables;
- Experimentation with predictive models based on machine learning, seeking those capable of, on a large scale and quickly, identifying indications of the operation of shell companies in the fraudulent issuance of fiscal documents with the intention of circumventing ICMS taxation;
- Evaluation of the suitability and potential application of the proposed methods in the daily operations of the Treasury Department of the Federal District.

To guide the data analysis process and the execution of tasks, the Cross Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000) was used. CRISP-DM is a methodology divided into stages and based on cycles.

FIGURE 1 CRISP-DM OVERVIEW



Source: Chapman et al. (2000).

4.1. Detailing the phases of the CRISP-DM Process

4.1.1 Business understanding

The first stage of "business understanding" involves getting to know the business domain, defining the problem to be addressed, and the objectives from the business perspective. This stage requires several interactions with the knowledge holders of the activity or process under study. In the present case, where the aim is to support the audit process of the treasury departments, the experts are, obviously, the tax auditors. In this phase, potentially useful data sources for analyzing the problem are also identified. For this, the participation of technology experts from these bodies, who are familiar with the existing databases, is necessary.

4.1.2 Data understanding

The "data understanding" phase is the longest phase of the project, where the aim is to identify quality problems and their interesting subsets for analysis, as well as to form hypotheses based on their visualization.

Chapman et al. (2000) describe the data understanding phase as the phase of selecting tables, records, attributes, and transforming and cleaning those considered not suitable for analysis.

Data from different sources are integrated, summarized, and prepared for the subsequent phases of analysis.

4.1.3 Data preparation

Once the attributes present in the databases related to the research context and relevant for identifying the typical fraud schemes operated by invoice companies are identified, it is necessary to define the variables to be used by the predictive models and the process of transforming the raw attributes from the databases into these analytical variables. This phase is known as "feature engineering."

The correct relationship between tables is a relevant step to make sense of the data. In some cases, raw data must be grouped, such as the set of incoming or outgoing invoices from a taxpayer.

Chapman et al. (2000) highlight the relevance of working with descriptive statistics and also the use of Business Intelligence dashboards for a better understanding of the data, including checking for correlation and dependency between variables.

4.1.4 Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. As a rule, there are several methods for the same type of data mining problem. Some have specific requirements in the form of input and output data. Therefore, it is often necessary to return to the data preparation phase and reformulate the training and testing bases. Each tested model is evaluated according to accuracy metrics and predictive potential. Once trained and evaluated, the model can then be used to make predictions on new data.

4.1.5 Evaluation

In the evaluation, the most important thing is to verify if the studied model meets the business objectives, that is, if it will be able to solve the proposed issue. For this, methods such as cross-validation and performance metrics are used to evaluate how well the models perform on unseen data. Based on this evaluation, it is possible to adjust and refine the models, if necessary.

4.1.6 Deployment

If the evaluation indicates the suitability of the created models, the result should be put into production, in order to add value to the business. The way this is done varies greatly and should integrate into the organization's work processes. Ideally, the models used should be continuously re-evaluated, through the comparison of prediction results with the practical reality of the business, to feed back into the model conception process and its evolution.

5. DEVELOPMENT

5.1 Business understanding

With the aim of mapping the vulnerabilities of the current taxation and ICMS collection system used by the federation units, a questionnaire (see Annex I) was sent to state tax auditors involved in combating shell companies. The questionnaire was sent to all state treasury departments. In total, 16 state tax auditors from ten different states responded. Representatives from the states of Alagoas, the Federal District, Espírito Santo, Goiás, Maranhão, Minas Gerais, Pernambuco, Piauí, Rio Grande do Norte, and Tocantins participated. The respondents have between six and 28 years of service dedicated to tax inspection.

From the responses collected through the questionnaire, relevant information to the study was extracted, among which we highlight:

a) To date, no state uses artificial intelligence-based analytical techniques for the detection of shell companies, which reinforces the pioneering nature of this work regarding tax inspection among Brazilian federated entities;
b) Six states declare the use of Business Intelligence (BI) in the analysis and visualization of data related to the tax issue;
c) As mentioned, an important action in combating shell companies is the Automatic Denial of the Invoice. Of the nine states interviewed, only three do not use this procedure routinely. All six federated units that use denial do so after a reasoned dispatch to ensure the legal security of the procedure. One state declared the use of Indication of Serious Irregularity (IGI), proposed by it, as a way to reduce discretion in the process.

Regarding the nature and integration of different inspection bodies in combating shell companies, the majority opinion is clear that the problem should be addressed with the integration of the Judiciary, the Public Prosecutor's Office, the Civil Police, and tax inspection. Only two tax auditors believe that the issue is resolved solely within the scope of tax inspection.

It was also questioned about what action states effectively adopt when identifying that a company is an shell company, allowing multiple responses:

a) Five states directly fine the shell company, but of these, only one does not list other jointly responsible parties, justifying that shell companies do not have assets and do not pay fines issued. In the same vein, one state does not list the partners of the social contract, and the others list all possible ones;
b) Eight states inform others that the company behaves like an shell company, which allows the destination state of the invoice to treat the tax credit of the entries, disallowing them;
c) Two states do not publish the act of ineligibility in the state official gazette, which hampers the tax inspection work of the destination state, and even one state that makes the publication does not inform the others, which makes it difficult for the destination state to act legally;

d) The real beneficiary of the fraud is the recipient of the invoice, who takes advantage of an undue tax credit. Only four states claim to reverse these credits, which makes the replication of the fraud advantageous for its authors;
e) Four states offer tax complaints to the Public Prosecutor's Office and bring to the attention of the Judiciary the frauds committed. This shows that, for part of the states studied, such practice does not have a criminal nature.

Respondents were encouraged to indicate typical behaviors of shell companies, rating them between 1 (Least relevant) and 5 (Most relevant), in terms of an indication of tax evasion practice. Among the behaviors assessed as "Most relevant," there is a tendency for a partner or accountant of an invoice company to replicate this fraud in other companies of their economic group. Another recurring fact is a company transacting with shell companies and, subsequently, adopting such a posture itself.

The tax auditors consulted do not see as an indication of fraud the fact that the company is not up to date with its ancillary obligations to record the tax. According to Tadat (2019), there is a distinction between the evader and the non-compliant. In general, the "non-compliant" may not be up to date with the ancillary obligation, and an inspection action or a stimulus, such as the tax mesh, can bring it into compliance. The evader, from the inspection experience, can simulate records in the submission of the Digital Fiscal Bookkeeping (EFD) to appear compliant. The fact that the company has partners in other states was also not considered a relevant indicator of shell company practice.

In addition to the proposals, tax auditors were encouraged to suggest other quantitative indicators of possible fraud, and the results were:

- Check if the partner is a beneficiary of a Government social program;
- Analyze the legal forms used for the company's constitution, basically if the company is individual or a partnership;
- Highlight companies located in apartments and commercial rooms instead of warehouses, especially for grain wholesalers;
- Validate the phones and addresses listed in the Tax Registration and if there is a registered accountant;
- Suspect companies in the same location where a company was previously identified as being an shell company.

Interviewees were also asked to evaluate the relevance of some suggested quantitative variables, which can be estimated from typical data in the possession of treasury departments, rating them between 1 (Least relevant) and 5 (Most relevant) in terms of detecting shell company practice. For the vast majority of auditors consulted, the following variables have high predictive power in this regard:

- The relationship between the company's share capital and its total outputs: seeks to verify if the company has outputs (sales) in an amount much higher than its initial capital;
- The relationship between the ICMS to be collected, which is a measure that evaluates the subtraction of the ICMS highlighted in the output deducted from the ICMS credited in the entry, and the total outputs of the company;
- The number of documents issued;

- The relationship between the book value of the entry and the book value of the output, because, in a normal relationship, the acquisition cost of the goods, added to a margin of added value, should be close to the value of the outputs.

There was no consensus regarding the relevance of the following variables:

- Net revenue, which is a measure that discounts costs from total sales (gross revenue);
- Total expenditure on salaries;
- Total number of customers, measured in a distinct count of recipients;
- The ratio between the number of employees and the total outputs;
- Total number of suppliers, measured in a distinct count of issuers.

Auditors were invited to suggest other quantitative variables that may indicate shell company practice. Among the responses, the following stand out:

- The number of days between the release of the tax registration and the first invoice issued;
- Concentration of economic movement in a few days of the month;
- Coherence between the expected economic activity (CNAE) and the sales products;
- Coherence between products in the entry and products in the output;
- Verification of reactivated state registration, that is, a company that had lost the right to issue invoices requests that its registration be reactivated.

Finally, the survey asked tax auditors if they had other relevant points that could contribute to the research. In this regard, there were responses related to allowing greater integration between the federation units and also between the Judiciary, the Civil Police, the Public Prosecutor's Office, and tax inspection in combating shell companies.

It is important to note that the responses obtained, as they do not represent even 50% of the state treasury departments, cannot be used as a complete picture of the issue of tax fraud perpetrated by shell companies in the country, nor of its combat by the federation units. However, there is also a common perception among respondents of the severity, breadth, and complexity of the problem, as well as a convergence in the selection of indicators and variables that can assist in identifying these frauds. This convergence proved robust enough for these selected indicators and variables to be used as a basis for the subsequent stages of this research.

But, obviously, this study, conducted in the context of a single treasury department, does not claim to propose a solution, or a method, to support the audit of shell companies, generalizable to all states, being only an exploratory study that tried, as far as possible, to use data and understandings common to a larger number of inspection units.

5.2 Data understanding

The databases used in this study are limited to those that a typical state tax authority has access to, in order to propose a solution that can be replicated in a larger number of states. An Extraction, Transformation, and Load (ETL) process was implemented that extracted raw data from the data lake of the Treasury Department of the Federal District in XML format, which were transformed into tables in a relational database. These tables were then processed using the Structured Query Language (SQL), generating data collections that were loaded and processed on the department's own equipment using the Python programming language.

The selection of databases for extraction was based on highlights and suggestions of variables and risk indicators made by the tax auditors themselves in their responses to the questionnaire. These databases contain data from electronic fiscal documents, specifically NFes and their events, as well as registration and corporate data. Additionally, data from Electronic Transportation Knowledge (CT-e), Electronic Consumer Invoice (NFCe), Digital Fiscal Bookkeeping (EFD), and Declaration of Payment Methods (DIMP) were used.

The events of the NFes indicate whether there was automatic denial and what type of occurrence led to the immediate denial of the issuance of invoices. The tax auditor initially records a standardized reason for the NFe block, which may be a consequence of a registration irregularity or a volume of operations incompatible with the size, registration status, or share capital of the company. Often, taxpayers who suffer an automatic denial action by the tax authority, but who are bona fide taxpayers, seek the tax authority, which, in turn, reverses this situation preventing the issuance of the fiscal document. A subsequent investigation by the auditors of the Treasury Department of the Federal District may identify some tax fraud, classifying the company as a shell company, if applicable.

Of the approximately 655,000 taxpayers registered in the Federal District, only 174,000 issued invoices from January 2014 to June 2023. For this study, a set of 2,801 companies initially indicated as suspicious of some irregularity was selected and later classified as shell companies or not by the tax auditors of the Federal District, through field validations. This classification was then used as a label to define the training base of the predictive models.

Thus, the indication of denial alone is not sufficient to determine if there was tax fraud, as it can occur for other reasons, but confirmation through audit strengthens the selection of fraud cases. Similarly, it cannot be stated that companies that were not analyzed by tax auditors and do not have an automatic denial event are not shell companies, which would make it difficult to select "non-shell companies" simply by sampling among those without automatic denial.

The next step was to merge the database of identified and validated shell and non-shell companies by tax auditors, with the other electronic fiscal documents issued by them.

For the study, data from Electronic Invoice (NFe) from January 2014 to June 2023 and Electronic Consumer Invoice (NFCe) from January 2016 (NFCe is a document with more recent adoption by taxpayers) to June 2023 were extracted.

The percentage of companies without a record of entry or exit by Electronic Fiscal Document (DFE) was 4.1%, and of the companies without exit documents or where these were zeroed, 9.7% of the sample in the data, as can be observed in Table 1. For these companies, the only available data are registration and corporate, which hinders their comparison to other companies contained in the database. For this reason, this study proceeded only with companies that have electronic fiscal documents of entry or exit with ICMS values.

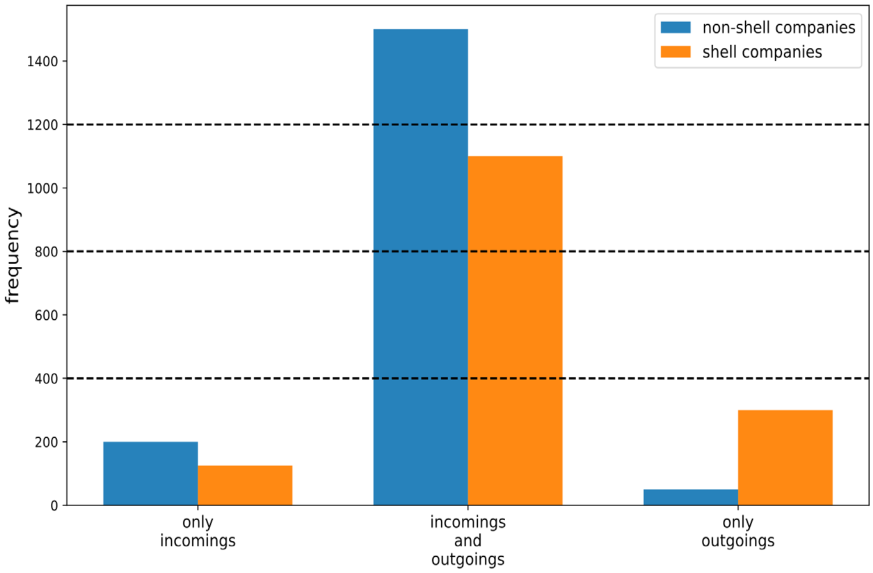
TABLE 1 EXISTENCE OR NOT OF AN ELECTRONIC TAX DOCUMENT (DFE)

	Count	Without exit tax document or with zero value	No entry or exit registration
Shell company	1.369	144	42
Non-shell company	1.432	173	92
Total	2.801	317	134

Source: Elaborated by the authors.

In Figure 2, we have the frequency of contributors from the selected base who promote only the entry of goods, only the exit or both.

FIGURE 2 INCOMINGS VS. OUTGOINGS

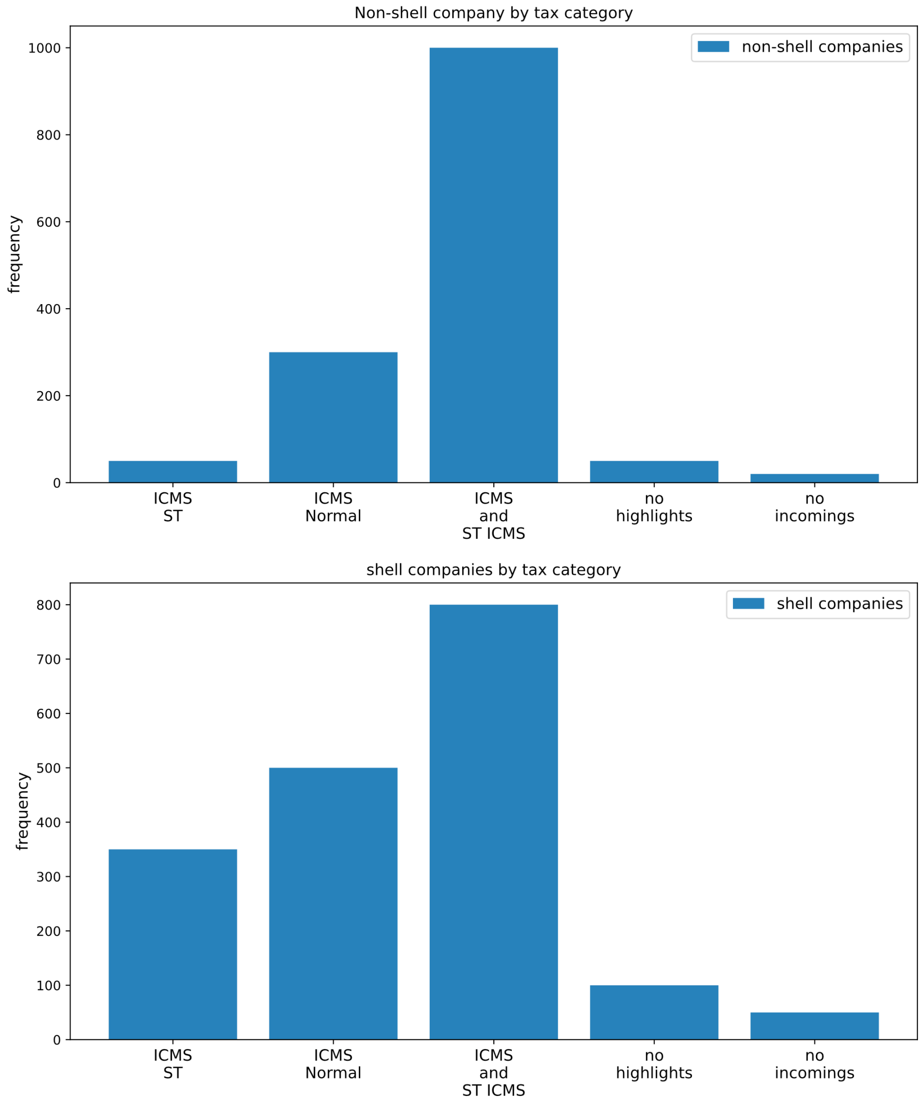


Source: Elaborated by the authors.

It is observed, then, that companies which only have incoming goods are less frequent, and there is a balance between shell companies and regular ones. The shell companies, however, have a marked tendency to sell goods they have not purchased, as noted in the third histogram of Figure 2, which displays the frequency of companies for which there is only outgoing goods. The majority of cases involve companies that both buy and sell (incoming and outgoing), and in this scenario, it is not as evident to differentiate the entire universe of shell companies.

Another significant characteristic, illustrated in Figure 3, concerns the tax that is highlighted on the invoice, whether it is the Normal ICMS, the ICMS Tax Substitution (ICMS-ST[1]), which anticipates the collection of ICMS at an earlier stage, whether there is no tax highlighted, or whether both are present. Again, the characteristic that stands out for invoice issuers is the receipt of invoices with highlighted ICMS and ICMS-ST, followed by the highlight of ICMS only, and finally, those that do not have incoming invoices (purchases). Meanwhile, the tendency to buy without ICMS highlight is more pronounced among non-shell companies.

FIGURE 3 ICMS HIGHLIGHT OR TAX SUBSTITUTION ON THE INVOICE



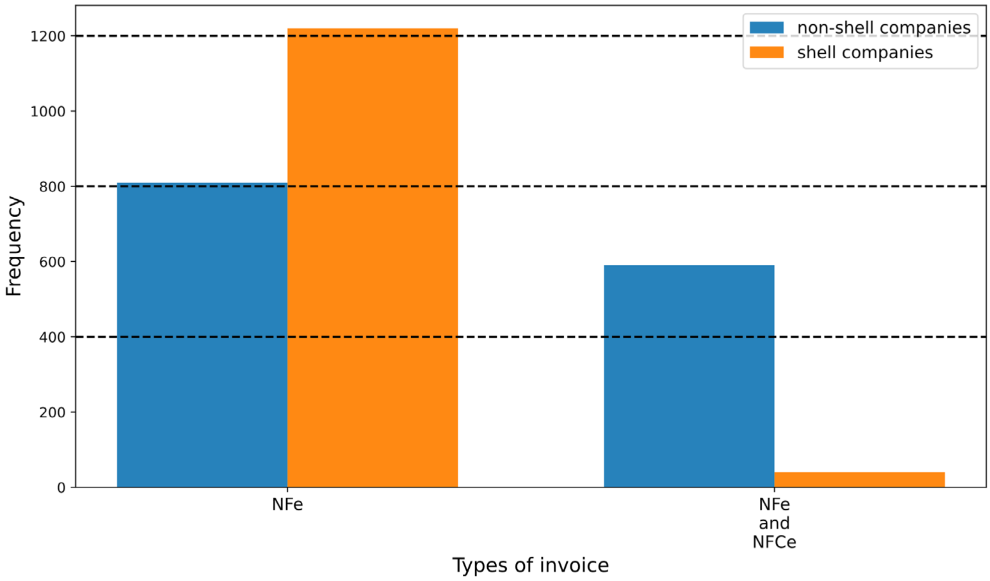
Source: Elaborated by the authors.

[1] ICMS Tax Replacement (ST) is the regime by which responsibility for the tax due in relation to operations or services provided is assigned to another taxpayer.

The data confirm an empirical understanding that shell companies work solely with Normal ICMS, thereby not utilizing goods subject only to the ICMS Tax Substitution regime. In non-shell companies, there is a greater division between Normal ICMS and Tax Substitution. A shell company registered under the Simples Nacional regime highlights ICMS on outgoing transactions, a provision only anticipated for larger companies within this regime under very specific circumstances. Conversely, a non-shell company from the Simples Nacional regime exhibits the opposite behavior of not highlighting ICMS, which is expected according to the tax regulations of these companies.

It is also observable that there is a significant distinction between the issuance of NFe and NFCe by shell companies and regular ones, as illustrated in Figure 4. The volume of transactions with NFe only is much higher among shell companies than among regular ones, as the latter maintain a closer relationship with consumer sales, hence, with NFCe. However, the set of companies classified as shell companies and non-shell companies by the tax auditors of Federal District is not sufficient to assert that shell companies do not issue NFCe. Similarly, it is not possible to claim that all non-shell companies issue both documents, since, due to their nature of commerce, they may only transact with legal entities using only NFes.

FIGURE 4 EXCLUSIVE USE OF NFE OR IN CONJUNCTION WITH NFCE



Source: Elaborated by the authors.

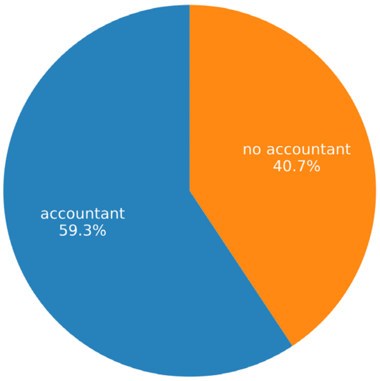
Companies that benefit from the shell company scheme are those that purchase invoices to credit themselves on ICMS and pay less tax. It is expected that the recipient companies of a given shell company buy from various shell companies. With the aim of developing the study, issuers and distinct recipients in all invoices were extracted from the NFe database. The extraction proceeded with the following algorithm:

1) For the known shell companies their recipients were identified;
2) From the recipients in step 1, companies that sold (issuers) to them, different from the company in step 1, were extracted;
3) The classification of the companies in step 2 was verified.

It is observed that the recipient companies of a given shell company tend to buy from other shell companies in an atypical proportion: nearly half of their suppliers (46.3%) are shell companies. This corroborates the findings in the questionnaire. It is relevant information, but, in isolation, does not have predictive power.

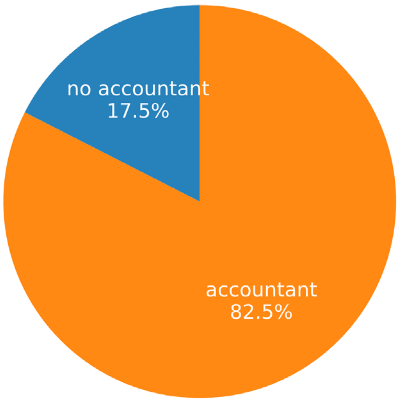
The accountant-client relationship can indicate various fiscal and accounting practices and the company's behavior. The absence of a registered accountant does not necessarily mean that the company does not have the guidance of an accountant, but possibly that this professional was not reported to the Finance Secretariat of the Federal District.

FIGURE 5 ACCOUNTANT IN NON-SHELL COMPANY



Source: Elaborated by the authors.

FIGURE 6 ACCOUNTANT IN SHELL COMPANY

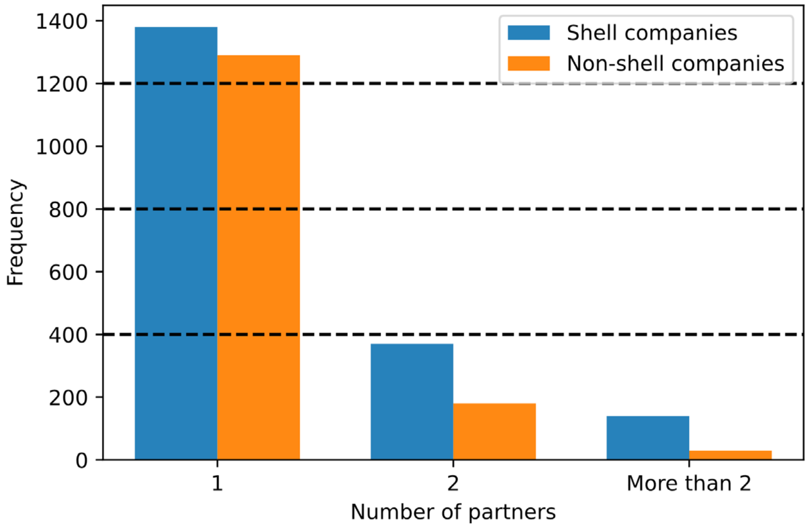


Source: Elaborated by the authors.

When comparing the charts of Figures 5 and 6, it is evident that the percentage of companies with an accountant is significantly lower in shell companies. As it generally involves possible indictment for a crime against the tax order, the tendency of the hidden accountant is marked so that, in case the scheme is discovered, they are not attributed a crime. However, some accountants seek to provide service in a way that gives a certain air of legality to the company.

The analysis of corporate composition is another relevant point within the study. Individual Microentrepreneur and Individual Limited Liability Company (Eireli) are companies with only one holder (or sole partner). Other companies are formed as a partnership, where two or more partners join to achieve the company's social objectives. Figure 7 shows the number of companies by number of partners in the base companies.

FIGURE 7 NUMBER OF PARTNERS PER COMPANY

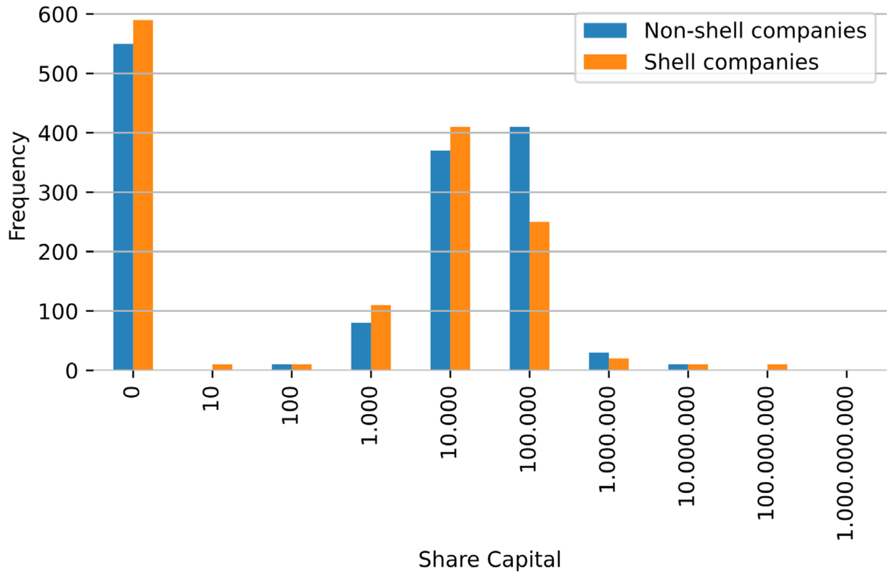


Source: Elaborated by the authors.

It is observed that shell companies tend to have a single partner, and that the occurrence of more than two partners is noticeably low in this category.

Regarding the company's capital, it was observed that there is a slight tendency for regular companies to have a higher capital than shell companies, as can be seen in Figure 8. Even among those companies that do not declare capital, the presence of shell companies is quite substantial.

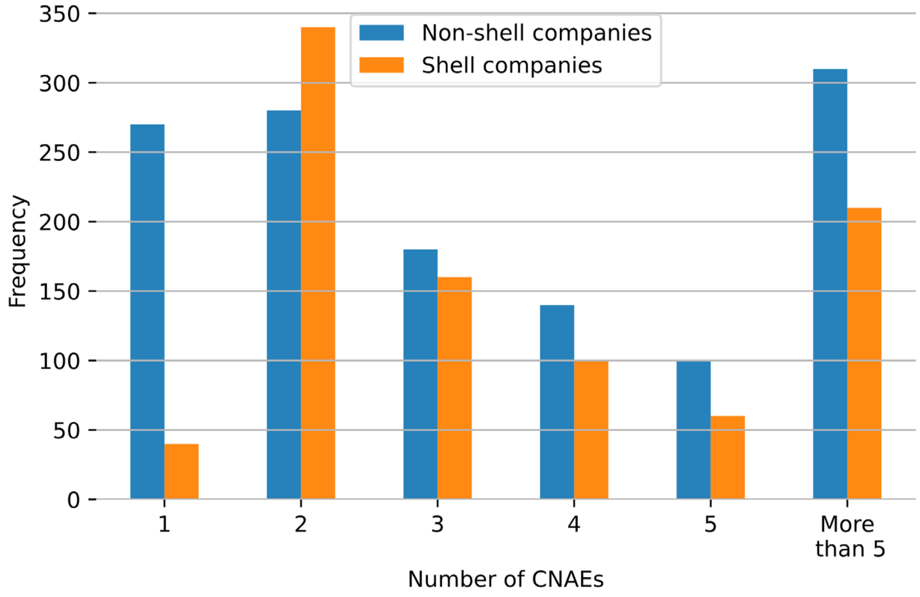
FIGURE 8 SHARE CAPITAL OF COMPANIES



Source: Elaborated by the authors.

The National Classification of Economic Activities (CNAE) also revealed that there is a subtle tendency for shell companies to declare fewer CNAEs than regular ones, as can be verified in Figure 9.

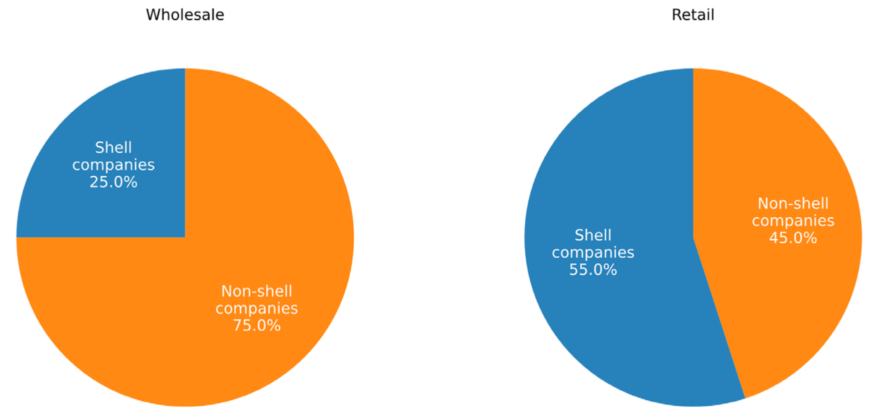
FIGURE 9 NUMBER OF CNAES BY TYPE OF COMPANY



Source: Elaborated by the authors.

Regarding economic activity, shell companies mainly operate in commerce and generally prefer retail, as shown in Figure 10.

FIGURE 10 COMPARISON BETWEEN WHOLESALE AND RETAIL



Source: Elaborated by the authors.

5.3 Data preparation

The objective of this phase is to prepare a data collection that will be used in the modeling phase. This collection contains data summarized by company, meaning each company has a unique record with its aggregated data for the period it remained active. For the training of predictive models, this collection is augmented with the indication of whether the company is a shell company or not. Once trained, the predictive models are capable of suggesting this classification based on input data.

5.3.1 Definition of attributes (features)

The companies selected for the training and testing database issued approximately 7 million invoices between 2014 and 2023. To create a profile of each company's activity, an aggregation was made, where each one has its values summarized in a single tuple.

The summarization of each company's activity took into account the following criteria:

- Economic Activity
    - Operates in retail and/or wholesale?
- Clientele Profile
    - Do clients purchase from companies already identified as shell companies?
- Accountant
    - Does the company have an accountant?
    - Has the accountant worked for companies already identified as shell companies?
- Corporate Structure
    - Number of partners.
    - Has any partner been linked to a company identified as a shell company?

- Financial Data
    - Accounting values of the company's operations:
        - □ differentiated between incoming and outgoing goods;
        - □ discriminates highlighted ICMS.
    - Count of issued and received invoices.
    - Issues invoice to the final consumer (NFCE)?
    - Purchases from a company identified as a shell company and from how many?
    - Number of suppliers of the target company that also sells to a shell company.
    - Number of months in which it received and sold goods:
        - □ difference between the first and last month with fiscal issuance.
    - Company's capital.

To avoid very large differences in values between companies of different sizes and transaction volumes, attributes were normalized whenever possible. For this, criteria that have justifications from the business point of view were adopted to the maximum extent to maintain the semantics of each attribute.

To define appropriate rules for each attribute, they were separated into three categories:

- Financial origin;
- Distinct count of person (partner, client, supplier);
- Distinct count of issued electronic fiscal documents.

For financial data, for example, an interesting value to base normalization on is the sum of the accounting value (VNF) of all incoming and outgoing invoices. In the volume of outgoing transactions, sales to the final consumer with the Electronic Consumer Invoice (NFCe) are also considered. Based on this value, proportions of incoming and outgoing are calculated. Thus, a taxpayer who only sells (outgoing) without buying (incoming), when having their incoming normalized by VNF, will obtain a value close to zero, and their normalized outgoing will have a value close to one. A typical taxpayer, who purchases goods for resale, will have their incoming and outgoing values normalized with values around 0.5.

The previously described aggregation sought to reflect, again, the elements indicated by state tax auditors as relevant for the analysis of fiscal fraud in the invoice issuer scheme, described in section 5.1.

For the final selection of attributes with the highest predictive value, a correlation analysis was conducted, using the Pearson coefficient, between the input variables, which describe the company, and between each input variable and the output variable, which indicates the company's class (invoice issuer or not).

Shimakura (2006) defines variables with coefficients greater than 0.90 as highly correlated. Input variables strongly correlated with each other mean duplication of information, indicating that the elimination of one of them does not harm the predictive model. Meanwhile, input variables weakly correlated to the output indicate attributes with weak predictive power. Thus, for each pair of input variables whose Pearson coefficient was above 0.90, one of them was discarded, and input variables whose correlation with the target variable presented a Pearson coefficient below 0.90, were also discarded. After analyzing the informative value of each attribute, 11 numeric and eight binary attributes were selected to compose the final base, described in Box 1.

BOX 1 DESCRIPTION OF SELECTED ATTRIBUTES

Description	Type
If the company operates in wholesale	Binary
If the company operates in retail	Binary
If customers have already purchased from shell companies	Binary
If an accountant works in shell companies	Binary
If the company has an accountant	Binary
If the company issues NFCe	Binary
If the company has a partner who is a partner in other companies	Binary
If the company has a partner who is or was a partner in a shell company	Binary
Book value of sales with the sum of NFe and NFCe	Numeric
Total value of ICMS at exit	Numeric
Book value of goods receipt (purchase) operations	Numeric
Value of ICMS Tax Replacement (ST) highlighted	Numeric
ICMS value highlighted for purchase of merchandise	Numeric
Value of share capital registered in the company's articles of incorporation	Numeric
Number of months between the first and last month with tax issuance	Numeric
Number of company partners	Numeric
Number of invoices issued for goods receipt	Numeric
Ratio between the number of outgoing documents and the total number of documents issued	Numeric
Ratio between sales value and total purchase and sale transactions	Numeric

Source: Elaborated by the authors.

5.3.2 Preparation of the training and testing base

The data were randomly separated into training and testing, as is customary in supervised machine learning experiments, and all precautions were taken to avoid biases due to data leakage between the training and testing base. 75% of the data were used for training and 25% for testing, so that for training, there were 966 examples of invoice issuers and 981 of non-invoice issuers, while for testing, there were 341 shell companies and 309 non-shell companies, constituting a reasonably balanced base.

To correct missing data (missing value), their replacement was carried out using measures of central tendency appropriate to each type of data.

5.3.3 Checking for bias in the data

The study is based on a base classified by tax auditors of the Federal District. The non-shell companies in the study are companies that had their invoices preventively denied and that, after analysis by the tax authorities, were released for issuance. Therefore, there are no companies in the base about which there was doubt whether they are shell companies or not at the end of the analysis. But, at the same time, doubts about their behavior hovered over all the companies in the base classified, at some point, even if for different reasons. Because of this, there remained a doubt whether the base labeled as non-shell company is representative of the universe of taxpayer companies, since the practice of fraud is an exception, or if it already contains a bias for having been previously selected by the auditors.

To examine the possibility of there being an inherent bias in the base of non-shell companies, a new random sample of companies was created from the rest of the companies not analyzed. A statistical analysis was then conducted, through a hypothesis test, which sought to verify whether the subset of non-shell companies from the set of denials presents the same frequency distributions as the new random sample. The test confirmed the null hypothesis that both samples have the same statistical behavior.

5.4 Modeling

5.4.1 Model selection for training

The experiment conducted herein is considered, in the machine learning literature, a supervised learning endeavor, based on a set of labeled examples aiming to find a predictive model that, once trained, is capable of classifying whether the behavior of a company, through its buying and selling transactions, registration data, and relationships with partners, suppliers, and clients, corresponds to that of a shell company.

The algorithms used in the training are all from the Sklearn[2] library for Python, which was chosen due to its ease of use, extensive documentation, free availability, and the diversity of methods it contains.

In projects involving machine learning, it is common to explore a wide range of predictive models and, ultimately, select those that perform best according to pre-established metrics and criteria. Some of these metrics evaluate the accuracy of the predictions, others their interpretability, or even the profile of errors the model typically makes.

In Carvalho et al. (2019), a comprehensive tutorial on metrics and strategies for evaluating predictive models can be found. Our work explored the most cited and utilized classification algorithms in the literature, with special attention to those applied in fraud detection, including those used in the works highlighted in section 2.1. After several rounds of testing and parameter adjustments, the algorithms selected as the most suitable for the problem at hand are presented in Box 2. An overview of how these algorithms function can be obtained in Mahesh (2020).

[2] https://scikit-learn.org/stable/

BOX 2 TRAINING MODELS

Sklearn Function	Algorithm
RandomForestClassifier	Random Forest
GradientBoostingClassifier	Gradient Descent
MLPClassifier	Neural Net
AdaBoostClassifier	Ada Boost
KNeighborsClassifier	Nearest Neighbors
DecisionTreeClassifier	Decision Tress
SVC	Support Vector Machine

Source: Elaborated by the authors.

5.4.2 Training and testing

To achieve the best hyper-parameter configuration of the models, the extended pipeline feature was chosen to process a cross-validation test with various settings of the algorithms (Claesen & Moor, 2015). For this purpose, the GridSearchCV algorithm was used, which performs an exhaustive search over specified parameter values for each classifier.

The best results were obtained with the following parameters:

- AdaBoostClassifier
    - learning_rate (the model's learning rate during training): 0.15
    - n_estimators (number of estimators in the model): 200
- GradientBoostingClassifier
    - learning_rate: 0.05
    - n_estimators: 200
- KNeighborsClassifier
    - n_neighbors (number of nearest neighbors to be considered): 9
- MLPClassifier:
    - activation (activation function used in the neural network's hidden layers): tanh
    - hidden_layer_sizes (the number and size of the neural network's hidden layers): 150
    - solver (optimization algorithm used to adjust the neural network's weights): adam
- RandomForestClassifier
    - criterion (quality measure used to evaluate the split of each decision tree node): entropy
    - n_estimators: 40
- SVC
    - kernel (boundary estimation method): rbf

After training the models with the aforementioned hyperparameters, testing was conducted on the dataset not used in training. Each instance is presented to each of the models and generates one of the following outcomes:

- True positive – shell companies, identified as shell companies;
- True negative – non-shell companies, identified as non-shell companies;
- False positive – non-shell companies that are identified as shell companies;
- False negative – shell companies that are identified as non-shell companies.

Table 2 summarizes the results based on the conventional metrics used in machine learning.

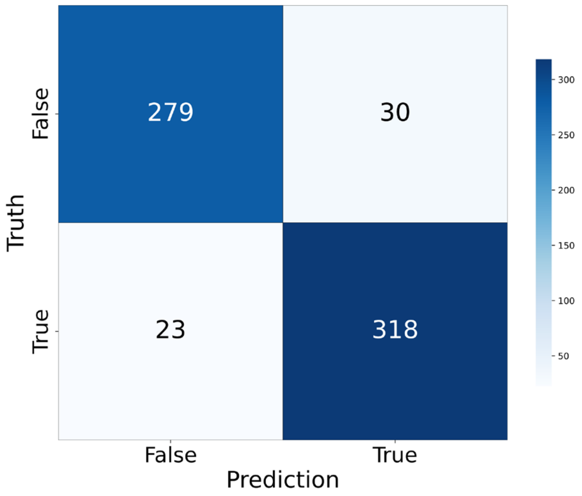
TABLE 2 MODEL EVALUATION RESULTS

Algorithm	Accuracy[1]	Precision[2]	Recall[3]	Specificity[4]	F1Score[5]
MLPC	0,918	0,933	0,914	0,924	0,923
Random Forest	0,915	0,903	0,933	0,897	0,918
Gradient Boosting	0,911	0,912	0,917	0,904	0,915
SVC	0,908	0,912	0,912	0,903	0,912
Ada Boost	0,895	0,9	0,9	0,89	0,9
K Neighbors	0,889	0,909	0,883	0,896	0,896
Decision Tree	0,715	0,745	0,695	0,738	0,719

Source: Elaborated by the authors.

As indicated in Table 4, the model that performed best was the Multi-layer Perceptron Classifier neural network with the confusion matrix in Figure 11.

FIGURE 11 CONFUSION MATRIX WITH MULTI-LAYER PERCEPTRON



Source: Elaborated by the authors.

6. FINAL CONSIDERATIONS

This study aimed to investigate the use of predictive techniques based on machine learning in identifying a specific practice of tax fraud, perpetrated by companies popularly known as "shell companies," formed exclusively to issue undue ICMS (Tax on Circulation of Goods and Services) credits.

Following a consolidated data analysis methodology, the study began by understanding the business context involving the tax issue and its enforcement by state finance secretariats.

Through a survey with tax auditors from various federal units, it was possible to map the vulnerabilities of the ICMS taxation and collection system exploited by shell companies and the strategies used by auditors to combat them.

Subsequently, extensive work in obtaining, processing, loading, and understanding the data typically available to regulatory agencies allowed for the materialization and measurement of tax fraud in its various dimensions. After selecting the most relevant variables derived from this data, based on the business knowledge of tax auditors and the use of appropriate statistical techniques, a database was prepared for the training and testing of machine learning algorithms. The application of a broad spectrum of algorithms, including those most referenced in the literature, demonstrated the feasibility of identifying companies with typical shell company behavior.

The proposed prediction methods based on data science have the potential to improve the selection of taxpayers to be included in the fiscal audit program based on analytical indications and on-site inspections, due to the criterion adopted by the audit to verify the indications.

The model presented in this work is being applied in practice at the Finance Secretariat of the Federal District (Sefaz-DF) in the search for indications of tax fraud, with the aim of increasing effectiveness in combating shell companies.

As future work, collecting results over a statistically significant period will allow for comparing the models' accuracy estimates with field reality, as well as proposing improvements to the proposed system. The methodology used is transparent and consistent with the practice of most state regulatory bodies, and similarly, the data sources necessary for its application are, in principle, also available in most federal units, so that the work carried out can be useful or serve as a starting point for other similar initiatives.

The main weaknesses of the present study are centered on external factors, such as the infeasibility of using databases not accessible by tax enforcement, among which the General Register of Employed and Unemployed (Caged) can be mentioned, which could bring better adequacy to the registration of partners.

REFERENCES

Abrantes, P. C., & Ferraz, F. (2016). Big data applied to tax evasion detection: a systematic review. In *International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA.

Allingham, M. G., & Sandmo, A. (1972). Income tax evasion: a theoretical analysis. *Journal of Public Economics*, *1*(3-4), 323-38. http://www.sciencedirect.com/science/article/pii/0047-2727(72)90010-2

Andrade, J. P. A., Paulucio, L. S., Paixão, T. M., Berriel, R. F., Carneiro, T. C. J., Carneiro, R. V., Souza, A. F., Badue, C., & Oliveira-Santos, T. (2021). A machine learning-based system for financial fraud detection. In *Anais do 18º Encontro Nacional de Inteligência Artificial e Computacional*, Porto Alegre, RS, Brasil. https://doi.org/10.5753/eniac.2021.18250

Azevedo, R. R., Silva, J. M., & Gatsios, R. C. (2015). Comparação de modelos de previsão de série temporal com base no ICMS estadual. Contabilidade e controladoria no século XXI. In *Anais do 4º Congresso Controladoria e Contabilidade*, São Paulo, SP, Brasil. https://congressousp.fipecafi.org/anais/artigos152015/35.pdf

Bittencourt, S. A. P., Neto. (2018). *Análise de "outliers" para o controle do risco de evasão tributária do ICMS* (Dissertação de Mestrado). Universidade de Brasília, Brasília, DF, Brasil.

Carvalho, L. (2018, 07 de dezembro). Mato Grosso integra operação nacional de combate às empresas noteiras. *Sefaz Notícias*. http://www5.sefaz.mt.gov.br/-/10942780-mato-grosso-integra-operacao-nacional-de-combate-as-empresas-noteiras

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: a survey on methods and metrics. *Electronics*, *8*(8), 832. https://doi.org/10.3390/electronics8080832

Castellón-González, P., & Velásquez, J. D. (2013). Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications*, *40*(5), 1427-36. https://doi.org/10.1016/j.eswa.2012.08.051

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: step-by-step data mining guide*. SPSS Inc.

Claesen, M., & Moor, B. (2015). *Hyperparameter search in machine learning*. https://doi.org/10.48550/arXiv.1502.02127

Ferreira, R. P. (2018). *Reconhecimento de cenários baseado nas localizações dos fornecedores do governo federal* (Dissertação de Mestrado). Universidade de Brasília, Brasília, DF, Brasil.

Ippolito, A., & Lozano, A. (2020). Tax crime prediction with machine learning: a case study in the municipality of São Paulo. In *Proceeding of the 22º International Conference on Enterprise Information Systems*. https://www.scitepress.org/Papers/2020/95647/95647.pdf

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., & Saez, E. (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica*, *79*(3), 651-92. https://eml.berkeley.edu/~saez/kleven-knudsen-kreiner-pedersen-saezEMA11taxaudit.pdf

Lima, S. L. M. (2007). O acompanhamento tributário - um novo paradigma em fiscalização para a Receita Federal do Brasil. In Ministério da Fazenda, & Secretaria da Receita Federal (Orgs.), *Administração pública: prêmio de criatividade e inovação auditor fiscal da Receita Federal José Antônio Schöntag: 6º prêmio Schöntag: monografias premiadas* (pp. 877-917). SRF. https://repositorio.enap.gov.br/handle/1/4575

Mahesh, B. (2020). Machine learning algorithms: a review. *International Journal of Science and Research*, 9, 381-386. https://www.ijsr.net/archive/v9i1/ART20203995.pdf

Matos, R. T. B. R. (2019). *Feature selection with low correlated binary features for potential tax fraudsters classification* (Tese de Doutorado). Universidade Federal do Ceará, Fortaleza, CE, Brasil. https://repositorio.ufc.br/handle/riufc/43348

Ortiz, J. (2022, 01 de fevereiro). Fazenda: Receita estadual identificou 844 empresas falsas nos últimos cinco anos. *Agência Estadual de Notícias*. https://rrmais.com.br/noticia/noticias/parana/fazenda-receita-estadual-identificou-844-empresas-falsas-nos-ultimos-cinco-anos

Redação DM Anápolis. (2021, 11 de novembro). *Operação desarticula esquema de notas fiscais frias*. https://www.dmanapolis.com.br/noticia/16207/

operacao-desarticula-esquema-de-notas-fiscais-frias

Ruzgas, T., Kižauskienė, L., Lukauskas, M., Sinkevičius, E., Frolovaitė, M., & Arnastauskaitė, J. (2023). Tax fraud reduction using analytics in an East European country. *Axioms*, *12*(3), 288. https://doi.org/10.3390/axioms12030288

Santos, A. H. S., Rocha, K. de L., Toldo, L. de A., & Fabel, V. H. B. (2022). *Estimativa da carga tributária bruta do governo geral*. Tesouro Nacional. https://www.anfip.org.br/wp-content/uploads/2023/03/RT-CARGA-TRIBUTARIA-ANUAL-2021-1.pdf

Shimakura, S. E. (2006, 30 de agosto). *Interpretação do coeficiente de correlação, CE003 – Estatística II*. Universidade Federal do Paraná. http://leg.ufpr.br/~silvia/CE003/node74.html

Souza, J. M. (2018). Tributos sobre consumo: novo modelo para um Brasil mais justo. In Fagnani, E. (Org.), *A reforma tributária necessária: diagnóstico e premissas* (804 p). Anfip, Fenafisco e Plataforma Política Social.

Tadat. (2019, 01 de novembro). *Tadat Subnational Field Guide*. https://www.tadat.org/assets/files/TADAT%20Subnational%20Field%20Guide%20-%20November%202019.pdf

Xavier, O., Pires, S., Marques, T., & Soares, A. (2022). Identificação de evasão fiscal utilizando dados abertos e inteligência artificial. *Revista de Administração Pública*, *56*(3), 426-40. https://doi.org/10.1590/0034-761220210256

Zumaya, M., Guerrero, R., Islas, E., Pineda, O., Gershenson, C., Iñiguez, G., & Pineda, C. (2021). Identifying tax evasion in Mexico with tools from network science and machine learning. In O. M. Granados, & J. R. Nicolás-Carlock (Eds.), *Corruption networks: understanding complex systems*. Springer. https://doi.org/10.1007/978-3-030-81484-7_6

Gunther Siqueira Lemos Gomes

Master in Governance, Technology and Innovation from the Catholic University of Brasília (UCB); Tax Auditor of the Federal District Revenue Service. E-mail: gunther.gomes@economia.df.gov.br

Remis Balaniuk

Ph.D. in Computer Science from the Institut National Polytechnique de Grenoble; Professor and researcher at the Catholic University of Brasília (UCB). E-mail: remis@puc.br

AUTHORS' CONTRIBUTION

Gunther Siqueira Lemos Gomes: Conceptualization (Supporting); Data curation (Lead); Formal analysis (Equal); Investigation (Equal); Methodology (Equal); Project administration (Supporting); Resources (Lead); Software (Lead); Validation (Equal); Visualization (Lead); Writing - original draft (Lead); Writing - review & editing (Equal).

Remis Balaniuk: Conceptualization (Lead); Project administration (Lead); Supervision (Lead); Validation (Equal); Visualization (Suporte); Writing - original draft (Supporting); Writing - review & editing (Equal).

DATA AVAILABILITY

The dataset supporting the results of this study is not publicly available because it contains sensitive information protected by tax secrecy.

FUNDING

Fundação de Apoio à Pesquisa do Distrito Federal (FAP-DF).

Annex

QUESTIONNAIRE FOR STATE TAX AUDITORS

Dear colleague from the state tax authority, thank you very much for your help!

This questionnaire is part of a research study on the behavior of shell companies and understanding which predictive variables are most effective in identifying this behavior in the universe of companies registered in each state. Due to its national characteristic, we ask tax auditors who work directly with combating the practice of shell companies in finance secretariats or equivalent bodies to answer the questions below.

A first set of questions aims to verify the situation of combating shell companies in each state. Then, we will present topics that seek to identify the indicator variables of behaviors associated with tax fraud committed by shell companies. Each variable suggestion will be described, and we ask you to evaluate it on a scale from 0 (Irrelevant) to 5 (Very relevant) for the detection of shell companies.

In open fields, colleagues can propose new predictive variables, which will be analyzed in the study and may be tested against the databases of the Finance Secretariat of the Federal District.

IDENTIFICATION OF THE TAX AUDITOR COLLEAGUE

a. Name:
b. Email:
c. Effective position:
d. Length of service in the position:

STATE TAX AUTHORITY

e. Federative unit:
f. Do you have a COI or equivalent body to combat shell companies?
g. Do you have any procedure supported by software for the detection of shell companies?

   i. Yes – supported by artificial intelligence.
   ii. Yes – supported by a BI dashboard.
   iii. No.

h. Do you use the denial of the issuance of electronic fiscal documents as a way to prevent fraud with shell companies?

ON COMBATING SHELL COMPANIES IN YOUR STATE

i. Shell companies are a problem that affects:

   i. Only the issuer's tax authority;
   ii. Only the recipient's tax authority;
   iii. Both the issuer's and the recipient's tax authority;
   iv. The entire tax enforcement.

j. Shell companies are a problem that requires:
  i. Tax-related operations;
  ii. Police operations;
  iii. Filing of charges for crimes against the tax order;
  iv. None of the above;
  v. All of the above.

k. How important is the participation of the Civil Police in combating shell companies?
  i. From 0 to 5 (Not important – Very important).

l. Does the Civil Police play an effective role in combating shell companies in your state?
  i. Yes or no.

m. How important is the participation of the Public Prosecutor's Office in combating shell companies?
  i. From 0 to 5 (Not important – Very important).

n. Does the Public Prosecutor's Office play an effective role in combating shell companies in your state?
  i. Yes or no.

o. How important is the participation of the Judiciary in combating shell companies?
  i. From 0 to 5 (Not important – Very important).

p. Does the Judiciary play an effective role in combating shell companies in your state?
  i. Yes or no.

BINARY TYPE PREDICTIVE VARIABLES (YES OR NO)

Evaluate each variable suggestion based on your experience in combating shell companies. Variables that you think do not influence the detection of shell companies should be marked with zero, and variables that have a high influence, 5, with intermediate values indicating relative importance.

1. Did the company purchase from a company already classified as a shell company?

In this case, the company under study, whether a shell company or not, purchased from a known shell company.

2. Has the company ever sold to companies that buy from companies already classified as shell companies?

In this case, the company under study, whether a shell company or not, sold to a company that knowingly buys from a shell company.

3. Is the partner/accountant of the company also a partner/accountant of a company classified as a shell company?

In this case, the company under study, whether a shell company or not, has a partner or accountant who has an accounting responsibility or partnership with a company that has already been classified as a shell company.

4. Has the company ever purchased from companies that have a partner/accountant who has companies that have already been classified as shell companies?

In this case, the company under study, whether a shell company or not, has a partner or accountant who purchased from a company that has a partner or accountant who has companies classified as shell companies, but the company from which it purchased was not classified as a shell company.

5. Does the partner have formal employment outside the federative unit for which they have a company?

   This case assesses the possibility of the partner being a front (strawman) due to having formal work in another state.

6. Does the partner have formal employment with a salary of up to 2 thousand reais?

   This case assesses the possibility of the partner being a front (strawman) due to having formal work with a low salary.

7. Is the company under the Simei regime but issued a note with ICMS highlighted?
8. Is the predominant NCM at entry the same as the predominant NCM at exit?
9. Delivered EFD (if normal) or PGDAS (if Simple National)?

NUMERICAL TYPE PREDICTIVE VARIABLES

1. Net revenue.
2. Total entries/total exits.
3. Share capital/total exits.
4. Number of fiscal exit documents.
5. Number of customers.
6. Number of suppliers.
7. Payroll/total exits.
8. Number of employees/total exits.
9. ICMS to be collected/total exits.

VARIABLE SUGGESTIONS

- Binary type (text field – free).
- Numerical type (text field – free).
- Other suggestions (text field – free).