IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 January 2025, accepted 16 February 2025, date of publication 19 February 2025, date of current version 28 February 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3543714

RESEARCH ARTICLE

Use of Survival Analysis for Shell Companies Analyses in a Brazilian State

JÔNATAS DELL DUCAS[1,2,3] AND RODRIGO COLNAGO CONTRERAS[3,4,5]

[1]Americana Tax Office, SEFAZ/SP, Americana 13465-009, Brazil
[2]Engineering Department, Americana College, Americana 13477-360, Brazil
[3]Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos 13566-590, Brazil
[4]Department of Science and Technology, Institute of Science and Technology, Federal University of São Paulo (UNIFESP), São José dos Campos 12247-014, Brazil
[5]Department of Computer Science and Statistics, Institute of Biosciences, Letters and Exact Sciences, São Paulo State University (UNESP), São José do Rio Preto 15054-000, Brazil

Corresponding author: Rodrigo Colnago Contreras (contreras@unifesp.br)

This work was supported in part by the "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)" through CAPES-JDD and RCC under Grant 001, and in part by "The State of São Paulo Research Foundation (FAPESP)" through FAPESP-RCC under Grant 2019/21464-1.

ABSTRACT One of the current problems in tax evasion is the creation of shell companies. Such companies operate in the provision of spurious credits, harming the market and the State. The Tax Authority needs to act as soon as possible to prevent such companies from generating the credit, as well as within the constitutional deadline for recovery of the amounts. The present work aims to analyze how the SEFAZ of the State of São Paulo has acted on this issue. Through the use of survival analysis in Python, it is sought to verify how the performance was in the period from 2015 to 2023. The time to detect shell companies as well as the time required for administrative processes are analyzed. First, an analysis is made through Kaplan Meier. Subsequently, using computational tools, the possibility of parameterizing the survival curve is verified. Through Cox regression, it is verified if there are relevant covariates that can be studied. The results are expected to provide subsidies for the development of lines of action within the SEFAZ to minimize the survival time of such companies.

INDEX TERMS Survival analysis, shell companies, cox regression, Kaplan-Meier.

I. INTRODUCTION

The issuance of fiscal documents serves to formalize commercial transactions or the provision of services. These documents enable the State to monitor and verify these transactions while also indicating the associated tax obligations. In the context of this study, which is focused on the State of São Paulo in Brazil, these documents are generated from an .xml file provided by the issuer.

Through the legislation of the National Integrated System of Economic-Fiscal Information (SINIEF) called SINIEF Adjustment No. 07/05 the electronic invoice (NF-e) and the electronic invoice auxiliary document (DANFE) were instituted throughout Brazil. Despite being instituted in 2005, the effective use of the NF-e only occurred in meaningful volume in the last decade. Its implementation was an evolution for the Brazilian tax system, considering the more agile dynamics in accountability, as well as the availability of operations to the State.

The evolution of the use of the NF-e allowed for evolution in tax inspection [1]. However, the same dynamic that is applied in the inspection can be seen in the means found by fraudsters to divert the values that should be directed to the public coffers.

One way to defraud the inspection is through the use of shell companies, defined by the Organization for Economic Co-operation and Development OECD [2] as companies that do not have any productive activity of their own. Such companies issue so-called unsuitable documents, that do not represent a transaction, to transfer tax credits. In these cases, neither the company exists, nor the transaction took place, and the documentation is unfit to generate tax credit.

The associate editor coordinating the review of this manuscript and approving it for publication was Sajid Ali.

For the State to recover the tax credit, it is necessary to carry out an administrative process that declares the nullity of the company. After the administrative process, an infraction report for the collection of the tax, together with fines and interest, is necessary. This infraction report is drawn up against the recipient of the fiscal document. All these activities must be carried out before the tax statute of limitations. The statute of limitations is the period in which the law allows for the recovery of the tax credit [3]. Understanding these deadlines is of fundamental importance, mainly due to the limited number of tax auditors available and the relevance of the tax to be recovered.

This work aims to use the statistical technique of survival analysis to verify the efficiency of the State of São Paulo (SP) in the processes that involve the detection of the existence of shell companies, as well as the duration of the administrative process that allows for the recovery of tax credits through penalties. Thus, the analysis aims to verify how long it takes for the State to detect the fraudulent company, as well as the duration of the administrative processes that must be carried out in accordance with current legislation. This is intending to analyze the assertiveness of the execution of the work within the tax statute of limitations.

Understanding how shell companies operate is essential for effective resource management and improvement of work processes, with the outcome serving as a management tool for decision-making. Thus, the aim is to verify whether it is possible to parameterize the relative times concerning the operation of these companies, as well as the work performed by the State Revenue Tax Auditor (AFRE) of SP. Finally, using Cox regression technique, the intention is to determine whether there are relevant and noteworthy parameters within the oversight structure that can be utilized to enhance the process of detecting and declaring the nullity of shell companies.

The study analyzes companies declared null through administrative proceedings between 2015 and September 2023 in SP. The use of survival analysis will allow us to verify whether the results achieved by the State are allowing the recovery of tax credits following the legislation. The result of the work aims to provide state managers with a better understanding of how this entire process has been occurring in the State.

II. PROBLEM DESCRIPTION

To understand the problem, it is necessary to analyze it in two ways. The first is the tax line, in which it is necessary to understand the problem being faced, namely, the issue of unfit companies, also called shell companies or shell companies [4].

The second line is the modeling the survival analysis of the companies in question. For this problem, there is the issue of statistical modeling as well as computational analysis.

A. THE TAX PROBLEM

Taxation is a fundamental factor for the maintenance of a State. Its financial survival depends on the institution and efficient collection of taxes. Taxation is a relationship between a State and society. In countries where democracy is established, such as in Brazil, the people, through their representatives in the legislative power, allow taxes to be instituted. These will be used for state maintenance, as well as to finance eventual public investments [5].

Taxation in Brazil is very complex, to the point that several pejorative expressions have been created, one of them is as a ''legal-tax manic asylum''. The expression, along with others like ''tax carnival'', is evidence of the complexity of taxation in Brazil [6].

The value-added tax (VAT) is the kind of taxation used in most of the world [7]. It is a tax that is charged at all stages of the production and commercialization process. However, the tax charged at one stage of the process becomes a credit for the recipient of the goods, who will use this amount to offset the tax due on the sale, after adding some value to the product. The addition does not necessarily have to come from a production process, but it can be, for example, the result of making the goods available in a certain location. In this case, taxation occurs through the transportation of the goods. The fact that the previously charged tax is deducted from the value of the next stage makes this tax non-cumulative.

Within these taxes is the tax on circulation of goods and services (ICMS). It has a state nature and follows the non-cumulative regime. Non-cumulativity is provided for in Article 155, §2, I, of the Brazilian Federal Constitution of 1988. The ICMS is the main tax among several that make up the Brazilian VAT. Brazil is one of the few countries that VAT is composed of several taxes. In Brazil, there are six taxes that make up the VAT, one state, one municipal, and four federals. The most common worldwide is a single tax that aggregates all [7].

The non-cumulative system operates through the mechanism of debits and credits applied to the taxpayer's fiscal account. The tax value on the purchased goods is recorded as a ''credit'' in the accounting. Upon the sale of the product, there will be a ''debit'' entry. The difference between the ''debits'' and ''credits'' indicates the ICMS to be collected. An important characteristic of non-cumulativity is self-monitoring within the production chain. There is an interest on the part of the purchaser to have the highest possible credit, so that it can be offset against the debits at the time of collection [3], [8]. Thus, the production structure itself is designed in a way that can assist the entire state monitoring structure. Table 1 illustrates the non-cumulative system.

Despite the importance of self-monitoring, Brazilian laws have established two systems that hinder the role of self-monitoring, namely, ''Simples Nacional'' (SN) and Tax Substitution (ST). The SN was established with the purpose of providing a regime of differentiated treatment, favored and single collection of taxes from all Federative Entities for microenterprises and small businesses [9].

TABLE 1. Example of VAT table.

		VAT non-cumulative				VAT cumulative	
	Sales value (A)	Tax rate (B)	Debit (C=AxB)	Credit (D)	Tax due (C-D)	Tax rate (E)	Tax due (AxE)
Stage 1	100	5%	5	0	5	5%	5
Stage 2	200	15%	30	5	25	5%	10
Stage 3	300	10%	30	30	0	5%	15
Final product	400	10%	40	30	10	5%	20
Overall taxation				**40**			**50**

The TS system is a controversial mechanism that allows the tax authorities to directly monitor the collection of taxes throughout the entire supply chain. However, this system eliminates the self-policing function of the tax that is collected under the non-cumulativity principle. Additionally, companies question the loss of profitability when using this methodology [10].

Within state revenue, composed mainly of three taxes, the ICMS is the main means of collection. Regarding state revenue, in 2021, it represented 83.5% of the state total, 22.39% of the national total collection, and 7.59% of the gross domestic product [11].

With the purpose of preventing fraud, the State has a supervisory force aimed at inhibiting fraud or punishing any deviations. For the audit activity, the State of São Paulo has the career of AFRE, who are the public authorities that carry out such work. However, as can be seen from Table 2, the number of employees in this career is not sufficient to supervise all establishments in the State.

TABLE 2. Fiscalization indicator - SP [12].

Fiscalization indicator - SP		
2023	**Total of AFRE**	**Total numbers of taxpayers registered / Total number of Tax Inspectors**
JAN	2,587	1,760
FEB	2,583	1,771
MAR	2,618	1,758
AVARAGE 1º QTR	2,596	1,763
APR	2,574	1,796
MAY	2,566	1,813
JUN	2,565	1,824
AVARAGE 2º QTR	2,568	1,811

In Table 2, you can find the total number of AFRE available and active in the state, however, many of them do not work directly in the direct taxpayer supervision. There are management, planning, and leadership positions held by these professionals, in addition to the motor vehicle property tax (IPVA) and the inheritance and donation tax (ITCMD), which are also state taxes, and need to be audited. Thus, the ratio presented in the table between the number of taxpayers and AFRE is much higher than calculated.

The cost-benefit ratio should also be evaluated when planning which taxpayers should be supervised. Taxpayers subject to periodic assessment regime (PAR) are those whose revenue is much higher, resulting in greater collection. If the Pareto principle is applied, most of the supervisory staff should be allocated to the supervision of companies subject to PAR. This results in a greater possibility of collection, either through self-regularization (foreseen in Complementary Law No. 1,320/2016) or even through the issuance of infraction notices. Self-regularization, in fact, is a new system that has been implemented in SEFAZ/SP since the creation of the ''Nos Conformes'' program. The program aims to improve the relationship between taxpayers and the tax authorities.

Still, even if all the supervisory force were applied to taxpayers under the PAR, the need for accuracy in supervisory tasks is fundamental. Supervision should aim at recovering tax credits, meaning it should be applied precisely to taxpayers who are deviating from compliance. Self-regularization, on the other hand, works on correcting any potential errors that may have occurred.

Before the implementation of the Electronic Invoice (NF-e), there were several misconducts that are no longer possible or have been ''modernized''. One of the practices that gained significant traction with the creation of the NF-e were the fictitious operations. In these cases, shell companies issue a large number of NF-e with high ICMS credit values for third parties, who take advantage of this illegitimate credit to offset their owed amounts.

B. SHELL COMPANIES AND TAX FRAUD

Defining a shell company is no easy task. The concept encompasses various aspects and can be interpreted in different ways. However, it is crucial to understand that the primary purpose of a shell company is to facilitate illegal activities.

The OECD defines ''shell entities'' or shell companies as those that lack any productive activity of their own [2]. Pacini et al. further emphasize that the purpose of shell companies is to create the appearance of a legitimate business, either domestic or foreign, with the sole intention of exploiting this facade of legality to violate the law. They point out that shell companies can be used to commit a wide range of crimes, such as currency evasion, money laundering, financial fraud, insurance fraud, corruption, and bribery [13].

International transparency presents shell company or corporation as a limited liability entity that has no physical presence in the jurisdiction, no employees and no commercial activity. Its main or sole purpose is to insulate the real

beneficial owner from taxes, disclosure or both. They also may be known as 'mailbox' or 'letterbox' companies [14].

Salazar et al. provide another definition of a shell company as a legal entity that, despite being formally registered, lacks operations, assets, employees, or any commercial activity. In other words, it does not engage in any operational activities. The authors also highlight that many countries offer simplified procedures for opening a company, requiring only a legal representative or even an address. This ease of incorporation facilitates the creation and operation of shell companies for illicit purposes [15].

Cruz Pérez provides a definition of the Spanish term ''Empresas de papel'' that is particularly relevant to this study. These companies are created fraudulently with the sole purpose of issuing false fiscal documents, which do not correspond to any actual sale of goods or provision of services. Their primary objective is to generate cold invoices, which are fraudulent documents used to evade taxes and launder money [16].

Pacini et al. [13] highlight that one of the main aspects of creating shell companies is to conceal the true beneficiary of the fraud. This results in three main points: lack of transparency in identifying beneficial owners, gatekeepers and trust company service providers, and layering and chaining of shell entities.

In India, for example, there are studies on the creation of shell companies for future incorporation, with the aim of committing fraud. Singh [17] also reports that not all cases involve fraud being planned at the time of the company's formation; however, after incorporation, the possibility of fraudulent use arises, and there is exploitation of the situation.

The case that occurs in the State of São Paulo, which is the subject of this study, involves the creation of companies to circumvent the tax authorities by issuing invoices that do not support a real commercial transaction. The sole purpose of the documentation is to provide ICMS credit indicated in the fraudulent documentation, so that the recipient can use it to reduce the tax owed, in other words, defrauding the state [18], [19]. In Brazil, these cases are identified as ''noteiras,'' as they issue fiscal documents solely to generate tax credits for the recipient. In other words, the purpose of the shell company is only to issue fiscal documents for the generation of spurious tax credits, and causes a fraud.

Fraud can be defined as a deceitful or dishonest act with the intention of deceiving a third party and gaining an unfair advantage over them. It can also involve the failure to fulfill an obligation. Salazar et al. present estimates from studies on VAT fraud. According to these studies, the United Kingdom experiences a loss of 10 to 12% of potential VAT revenue due to fraud, while in Germany, this figure is estimated at 11%. This means that a significant portion of taxes goes uncollected, depriving public coffers of much-needed revenue [15]. Another studies estimates that in 2013 over EUR 168 billion was lost across the European Union [20].

In Europe, the type of operation created to provide irregular tax credits for companies is described within the illicit activities of cross invoicing. An European document reports seven main types of fraud involving VAT, with the creation of tax documents by shell companies solely for the purpose of providing credit being one of the investigated cases of fraud in the area of cross invoicing [21].

The cases of companies issuing tax documents solely to generate credits (cross invoicing) are not a privilege limited to Brazil and Europe. Examples of this type of fraud can be found in other countries, such as Mexico [22], Ecuador [23], and Indonesia [24].

One way to analyze tax losses is through the VAT-Gap. The VAT-Gap is the percentage of VAT or other tax that should have been received by the government but was not. Reducing this Gap is important for several reasons: the revenue loss due to the VAT-Gap can have a negative impact on government spending, and monitoring the VAT-Gap can help develop better targets for future policies [25]. Furthermore, the VAT-Gap serves as a tool for evaluating the efficiency of tax activities aimed at preventing tax evasion. But the metric is not exclusively used to evaluate tax losses, as in cases like Brazil, potential tax benefits influence the metric and must be taken into account.

Thus, the State's objective is always to reduce the VAT-Gap, demonstrating better efficiency in VAT collection. Therefore, the work of AFRE is closely linked to the reduction of the VAT-Gap. The result of this work aims to provide managerial tools with the ultimate goal of assisting decisions that reduce this gap.

It is important to highlight that tax fraud not only harms the state but also has broader economic consequences. By evading taxes, fraudsters gain an unfair advantage over businesses that comply with the law. This allows them to offer their products at lower prices, distorting the market and creating an uneven playing field. In this way, the reduction of fraud is not only important for state revenue but also allows for greater competitive balance among companies [21].

C. DETECTION OF FRAUDULENT TAX CREDITS

As previously described, there are several reasons for the establishment of shell companies. A large part of the studies focuses on the use of tools for detecting suspicious activities, mainly for money laundering, as described, for example, by Singh and Best [26]. Unlike cases of money laundering, the creation of shell companies for generating bogus credits has different characteristics and modes of operation, which leads to the use of different approaches to seek the detection of such situations.

With the large amount of existing data and operations, it becomes increasingly important to use data analysis and statistical tools for fraud detection. With the use of computational tools, there is an expansion in the monitoring and detection of tax fraud.

Zheng et al. [27] presents an overview of the various data mining techniques that have been used for detecting tax risk behavior. They categorize the techniques into non-relationship-based and relationship-based, highlighting 14 techniques that have been used as essential tools for detecting tax fraud. However, it is noted that despite the existence of various techniques in use, they still have limitations. In addition to the inherent limitations of the tools, there is the difficulty arising from the dynamic changes in methods by fraudsters, which prevents a static solution to the problem. In Brazil, for example, the implementation of the digital tool SPED has not proven to be efficient in reducing the VAT gap without the use of other complementary tools [28].

The detection of shell companies has been carried out using various tools, which may or may not work complementarily. Segovia-Vargas et al. [15], for example, presents the use of dynamic social networks in such detection. They managed to achieve a low false positive rate, which is favorable in reducing time, costs, and the need for human resources for detection. Krysovatyy et al. [29] utilized logistic regression, achieving an accuracy of 0.99. In a case study in India, Aggarwal and Dharni [30] employed Benford's Law with satisfactory results.

For detecting VAT fraud, various techniques have been used. Zumaya et al. [31] presents the use of deep neural networks, random forests, and complex network approaches, concluding that they are useful tools for fraud detection; however, they cannot be considered a final solution. In the Brazilian state of Bahia, there was also a simulation using neural networks, which yielded better results compared to the tools currently used by the state entity [32].

The detection of ICMS fraud, specifically in Brazil, has had various approaches in different states. In Sergipe, a hierarchical classifier was used to assist in the detection of fraud in medications [33]. This classifier achieved an accuracy of 0.99. Aguilo and Quadrelli [34] presents the use of fuzzy logic for detecting fraud arising from ''empresas noteiras'' in Rio de Janeiro. The technique used by Rio de Janeiro is based on set theory. The new tool has emerged as an alternative with the potential for significant gains.

In Brasília, several fraud detection tools are being studied. Neto et al. [35] conducted studies in the field of data envelopment analysis as a monitoring tool for activities aimed at detecting fraud. On the other hand, Di Oliveira et al. [36] utilized machine learning to assess risk analysis in taxation. In the study, it was concluded that the full discretization methodology provided a good solution for preparing the data for modeling. Finally, Gomes and Balaniuk [37] also conducted work with ''note companies,'' using 7 tools, with the best detection result being achieved by the MLPClassifier.

D. SURVIVAL ANALYSIS

According to Colosimo and Giolo [38], survival analysis is one of the areas of statistics that experienced the most growth in the late 20th century. Much of this growth was due to advances in computing, which have allowed for faster and more accurate analysis.

A good definition for survival analysis would be a time-to-event analysis until the occurrence of a certain event of interest. The time between the start of the count and the occurrence of the event is called the failure time. It is clear from the definition that this is an analysis of an event that has a binomial response, i.e., 1 if it has occurred and 0 while the event has not occurred.

Harrell [39] highlights three reasons why survival analysis requires specialized analysis methods. The first is that the failure time does not follow a usual distribution. Unlike the vast majority that follows a normal distribution, survival analysis has a skewed distribution. The second is that after a certain time, the probability of survival becomes more relevant than the expected survival time. And third, the fact that within survival analysis there is the hazard function, which helps in understanding the mechanisms related to failure.

The author also discusses the importance of survival analysis experiments, stating that even a simple model allows for several comparisons. Namely:

- Testing and describing interactions for different decision models;
- Understanding prognostic factors;
- Modeling the effects of treatments;
- Understanding the effects of actions on the time to an event;
- Testing the effects of actions;
- Adjusting for imbalances in actions.

In terms of its application fields, the technique had its initial development linked to medicine, where it is still widely applied. However, there are several areas where the technique has been applied, such as credit recovery analysis [40], employability-related analyses [41], materials [42], and retirement issues [43], among others.

In the case of using survival analysis in issues related to fraud or taxes, some instances of its application can be observed. Kuzmenko et al. [44] used it to evaluate the electronic fraud detection system in Ukraine. A similar evaluation (of the survival period of fraudulent companies) is what is intended in this work, meaning that the result will be used as a tool to assess current techniques.

Meanwhile, Tonetto et al. [45] applied survival analysis in the Brazilian state of Rio Grande do Sul, evaluating the survival time of companies after being fined. This is a different approach from this work, but using the same technique. This work seeks to verify the survival of a shell company, while that one aims to understand how long a certain company survives in the market after being penalized by the state.

For survival analysis, mainly two functions are used to represent events, namely the survival function and the hazard function. The main variable associated with the survival model is represented by T and is a continuous, non-negative

variable representing the time to failure. That is, it measures the time between the start of the analysis until the moment the response transitions from state 0 to state 1 [38].

The survival function can be defined as the probability of no failure occurring until a certain time t. From another perspective, it is the probability that the observation survives beyond time t. Thus, the survival function can be defined in Equation (1), where $S(t)$ denotes the survival function and $F(t)$ represents the cumulative distribution function of T [46].

$$S(t) = P(T \geq t) = 1 - F(t). \tag{1}$$

The cumulative distribution function, which can be considered as complementary to the survival function, is given by Equation (2).

$$F(t) = P(T \leq t) = \int_0^t f(s)\mathrm{d}s, \tag{2}$$

where $f(s)$ is the probability density function.

It is possible to approach survival analysis based on risk. That is, the probability of failure occurring within a time interval $[t1, t2)$. If the situation is expressed in terms of the survival function, the result is given by Equation (3):

$$S(t_1) - S(t_2). \tag{3}$$

The failure rate in the interval presented above is defined as the probability that failure will occur within this interval, given that failure has not occurred before t_1. If the difference between the two times is very small, we can define it as Δt.

Thus, if the value of Δt is very small, it will tend to 0. From these considerations, it is possible to define the hazard function of T, which is expressed by Equation (4), and the cumulative hazard function, given by Equation (5).

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{4}$$

$$\Lambda(t) = \int_0^t \lambda(t)\mathrm{d}s. \tag{5}$$

Regarding the computational analysis of survival analysis problems, Wang et al. [47] emphasize that survival analysis methods can be classified into two main groups: those that use statistics and machine learning methods. It is highlighted that both groups have the common objective of being able to make predictions related to survival probability and, consequently, estimate life expectancy.

Within this taxonomy, statistical methods focus on using the distribution of events over time along with statistical properties that allow parameter estimation to make predictions. On the other hand, machine learning methods focus on predicting the occurrence of an event at a specific time by incorporating traditional survival analysis methods with machine learning techniques.

Regarding practical application, machine learning techniques are traditionally applied to problems with high-dimensional data, whereas statistics is more applied to cases with lower-dimensional data.

The use of statistical techniques for survival analysis is interesting in the aim to identify factors that may influence the survival time of a particular event, and how each factor contributes to the final outcome. However complex the study and associated factors may be, answers to the questions of interest are derived from survival data. The first step is typically statistical analysis, starting with a descriptive analysis of the data [48].

III. MATERIALS AND METHODS

The study is based on taxpayers whose State Registrations (SR) were declared null by the State of São Paulo from 2015 until mid-2023. The declaration of nullity is an event that signifies the existence of a shell company. The data pertains to Brazilian companies, all located in the state of São Paulo.

A. DATA STRUCTURE

For proper data modeling, it is necessary to obtain and treat the data correctly. The information used in this case can be obtained in two ways. Data can be obtained through manual extraction, individually, from publications in the Official Gazette of the State (DOE) and the Sintegra/Cadesp system, or through the use of Data Warehouse (DW) tools. Sintegra is the Brazilian state database where information about companies provided by the states is available.

The data was obtained from two distinct sources. The first source is the DW of SEFAZ-SP, which contains several Data Marts (DM) with various information about taxpayers in São Paulo. One of these DM is responsible for storing taxpayer registration data, known as the Cadesp universe. The second database is a collaborative environment. In this environment, data related to Administrative Cancellation Processes (PAC) and Administrative Cancellation and Nullity Processes (PCN) are registered.

The specific interest in companies where the PAC/PCN process occurred arises from the fact that the process is initiated with the purpose of recovering tax credits. Without the administrative procedures, it is not possible to recover the credit through an infraction notice. When the amounts to be recovered are very low, or the recovery would be inefficient, no administrative procedures are carried out.

The Cadesp universe consists of all companies that have been registered in the State of São Paulo. In this system, there are registration information along with their respective histories (and alterations). All information related to registered companies in the State is located in the Cadesp system. There, it is possible to find information regarding address, partners, the sector to which it belongs, economic activities carried out, as well as the entire history of changes in the company's registration. It is within the Cadesp system that one can find the dates related to the suspension and null companies. These dates are what matter and need to be extracted.

The dates of cadastral changes for each situation are recorded in Cadesp, as published in the DOE. In this way,

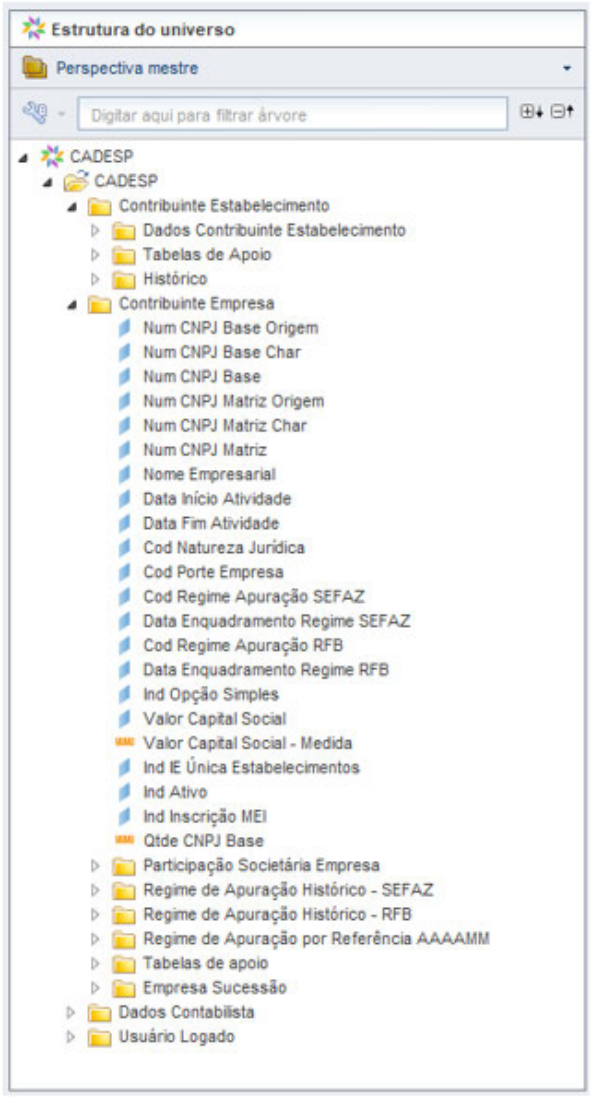


FIGURE 1. Structure of the Cadesp universe.

since there are two locations where this information is available, data extraction can be performed in two ways: through a DW that contains all the information from Cadesp, or through the manual extraction of information directly from the DOE. This last option, although it exists, is technically complex and inefficient.

The SEFAZ has a DW where various information is available. Within this DW, there is the ''Cadesp universe,'' where all registration information of São Paulo companies is stored in a structured manner. Figure 1 presents a representation of the structure of the ''Cadesp universe'' within the DW. The universe is updated daily through the ETL process and contains approximately 11 million taxpayers. The main purpose of the figure is merely to illustrate how it is structured the DW. The language used is Portuguese, given that the DW was developed exclusively for SEFAZ-SP.

To extract information from the DW, the Business Intelligence tool available at SEFAZ was used through the Rich Client web intelligence, developed by SAP. The interest lies in companies that have been declared null, which is one of the five existing registration statuses (active, canceled, suspended, null, and ineligible). Thus, companies declared null from January 1, 2015, until September 6, 2023, can be selected.

However, there are seven situations in which a company can be declared null (simulation of the existence of the establishment or the company, simulation of the company's corporate structure, non-existence of the establishment for which the registration was made, indication of false registration data, annulment due to non-confirmation, annulment due to defect or judicial decision, and annulled by the Federal Revenue of Brazil). Although there is not complete uniformity in the selection made by SEFAZ officials, the preference is to choose the two main events used. The choice was for ''Simulation of existence of the establishment or the company'' and ''Non-existence of establishment for which registration was made or incorrect indication of its location'', as they best characterized the times to be calculated. The total number of state registrations that met the situation was 5,369.

With the desired registration characteristics in hand, the selection of other information related to the SR that will be extracted begins. The selected data were as follows: start date of the SR, start date of validity of the current situation, date of change of the SR (the date it was declared null), date of change to suspension, Regional Tax Office (DRT) to which the establishment is subordinated (there are 18 DRT), and its sectorial (companies are separated by the industrial sectors to which they belong).

Regarding the collaborative data, they were manually extracted from the tax database. Two files were extracted in .csv format containing the information available in the environment. The data frames were concatenated, totaling 9,006 occurrences. From the data available in the data frame, the only ones containing relevant information are the CNPJ and whether the company has a PAC/PCN prepared.

Therefore, with the integration of the data from the DW with those from the collaborative spreadsheet, there will be two databases that will be analyzed. The first consists of all companies with nullity in effect from January 1, 2015 (data extracted from the DW), and the second, which is a subset of the first, contains those that underwent an administrative procedure aimed at declaring the nullity of tax documents (cross-referencing the data extracted from the DW with the collaborative spreadsheet). The latter is the database of companies that went through the PAC/PCN process.

There is no censoring applied in the cases in question, as the consulted values already contain the necessary information. It is not possible to conduct a partial survey or analyze those that are in progress.

B. TIMES OF INTEREST

With the data properly prepared, we proceed to define which times will be calculated. That is, the definition of the representations of 0 and 1 and of T. There are two distinct

survivals that we wish to verify, as they represent different time analyses and are of real importance.

The first period to be measured is the period between the start date of the nullification of the SR and the date of suspension of the registration. Suspension is an event carried out by the Tax Authority preventively, aimed at preventing the company from conducting commercial operations. From the moment of the taxpayer's suspension, it is not possible to issue NF-e. There is no possibility of invoicing. In cases of ineligibility, there is no possibility of issuing fraudulent documents for the transfer of spurious credit. The nullification date represents the moment when the company is considered null by the Tax Authority. Although the nullification declaration may occur later, it always refers to a date prior to suspension.

The interest in this time period arises from the fact that it shows the survival of an ineligible company until it is prevented from issuing fiscal documents. This time represents the period during which the company was able to commit fraud.

The second interval to be measured is the time between the nullification date of the SR and the date of the nullification declaration, in other words, the definitive cancellation of the SR. The nullification declaration is important because from this moment onwards, the Tax Authority is authorized to initiate credit recovery through an infraction notice and imposition of fines.

This time period also holds significant importance due to the provision of Article 150, §4 of the CTN (National Tax Code). In this article, the deadline of five years is established for the Tax Authority to contest self-assessed tax assessments. Therefore, if nullification is not declared before this period, documents issued more than 5 years ago cannot have their credits contested.

The third interval to be measured is the time between the suspension of the SR and the date when nullification is declared. The importance of this time is related to the Tax Authority's ability to gather evidence and carry out procedures that allow the continuation of its work. Certainly, this cannot be regarded as an efficiency evaluation, considering that processes are not always the same, and in some cases, there are deadline suspensions by the Judiciary. However, it demonstrates the time that the Tax Authority has been spending on administrative procedures for nullification declaration.

Table 3 summarizes the definition of the times to be calculated, and Figure 2 presents the definition of the time intervals, as well as a graphical representation of them.

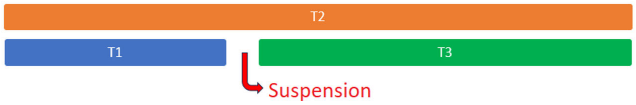


FIGURE 2. Graphical representation of the measured times.

TABLE 3. Definition of the times to be analyzed.

Caption	Definition
T1	Time of irregular operation
T2	The difference between the date the registration was changed to null and the date it was deemed null
T3	Time of execution of the administrative process for declaring nullity

C. DATA

As previously described, the information extracted from Cadesp was: start date of the SR, start date of validity of the current situation, date of change of the SR (the date it was declared null), date of change to suspension, DRT to which the establishment is subordinated, and its sectorial.

Each piece of information serves a purpose in the final result. The DRT and sectoral data were extracted for use in the Cox analysis. Through this data, it will be possible to see if regionality (DRT) or the company's area of economic activity has any influence on the times presented in the previous subsection. Understanding the influence of territorial or economic regionality can allow the manager to better understand who has been operating with greater efficiency and, thus, try to understand what causes the difference.

DRT is used only to determine in which region of the state the company is located, that is, it is originally a string format data. To use it in the Cox analysis, they were converted into integers, from 1 to 18, considering that the value itself will not influence the survival times. Each value represents a specific region of São Paulo.

The sectoral data has characteristics similar to the DRT; however, it identifies the economic area of activity of the company in question. The main difference is that there are 21 economic sectors, not just 18.

The date information is essential to obtain the times T1, T2, and T3 presented earlier. The information is in date format in Cadesp, and the times are obtained by calculating the difference in days between the dates.

T1 is the result of the difference between the date of change to suspension and the start date of the SR, which reveals the time between the opening of the company and the state's detection of irregular operation. T2 is the difference between the start date of validity of the current situation and the start date of the SR. This interval reveals the time difference between the end of the administrative process and the creation of the company. T3 is the difference between the date of change of the SR and the date of change to suspension, revealing the time that the state took to declare the company null.

D. ANALYSING

Based on the model descriptions presented by Wang et al. [47], models within the statistical methods are selected. Analyses are carried out using the Kaplan-Meier (KM) method to obtain a non-parametric result. Subsequently,

it is checked which parametric modeling can better fit the obtained results. Subsequently, we will use Cox regression to check if there is any covariate influencing the times.

The analyses are performed in Python, using the lifelines library. KM curves are plotted, as well as analyses of parametric methods. The seaborn library will be used to create boxplots to assist in temporal analyses. The boxplots will help verify if the deadlines for the tasks have been satisfactory. Figure 3 illustrates the sequence of events that will be developed.

IV. RESULTS

Before presenting the results, it is important to highlight the relevant aspect of each of the times for the work of state inspection. These times bring different information and relevance to the problem itself.

T1 is responsible for measuring the period during which the company operated irregularly. In other words, the period during which it could issue fiscal documents, consequently causing losses to the public treasury. The intention is for this time to always be as short as possible.

T2 is related to the legal deadline for recovering tax credits. If an NF-e was issued on a certain date with spurious credit, and the nullification declaration occurs more than 5 years later, this credit cannot be recovered according to Article 173 of the CTN.

T3 provides an analysis of administrative efficiency in proving nullification. After the suspension of the state registration, there is a need to conduct a process to declare nullification. These are administrative procedures aimed at proving the irregularity of operations.

A. SURVIVAL ANALYSIS OF THE COMPLETE DATABASE

From the DW data, survival analysis was conducted using the Kaplan-Meier method for the 5,369 companies. Using the lifelines tool in Python, the three Kaplan-Meier curves were plotted. The result is presented in Figure 4.

A brief analysis of the graph shows a much steeper decline in T1 compared to T3. This demonstrates that there is greater agility in detecting such companies than in the bureaucratic processes for declaring nullity.

Through an analysis of T1, it is found that just over 20% of the companies reach 500 days of operation. Furthermore, 50% of irregularity cases are detected within a period of 233 days.

Regarding T2, an important number to note is that only 8.5% survived beyond 1,825 days, meaning less than 10% had their period for tax recovery initiation after 5 years from the nullification date.

Analyzing the data for T3, it is possible to determine that 50% of administrative processes are completed within 492 days, and 90% of cases are finalized within 1,114 days. In other words, approximately 90% of cases have the administrative process completed within 3 years, allowing for a significant amount of time to take actions for tax credit recovery.

In order to verify the possibility of using a parametric curve, an analysis of six regression methods was conducted. Regressions were performed and the Akaike Information Criterion (AIC) values were extracted to evaluate the possibility of parameterizing the data. According to Amin and Khan [49], the AIC aims to estimate the information loss. The higher its value, the greater the loss. Table 4 presents the obtained AIC values.

The high AIC values demonstrate the difficulty in finding a parametric method that can be used to represent the times in question. This difficulty in parameterization may stem from the quantity of variables involved in the tasks. One notable aspect is the human factor. Both the possibility of detection and the administrative procedures involve human actions, both on the part of the perpetrator and on the part of the state's officials.

To enhance the visualization of the information, a boxplot of the three survival times was created. Figure 5 presents the boxplots.

Through the boxplot, it is possible to mainly observe the volume of outliers. Although not only outliers exceed the desired period of procedures, they represent the most problematic cases.

Subsequently, the semiparametric Cox analysis was used to verify possible relationships between covariates and survival analysis. The covariates of DRT and Sectorial were analyzed. The first covariate DRT is related to the territorial division of the offices responsible for tax inspection in the state. The second one is linked to the division between the most representative commercial and industrial sectors in the state.

The idea was to check if there was any influence at T1, T2, and T3 from any of the covariates. Since the information was non-numeric, a numerical encoding was performed so that the Cox analysis could be carried out. Figure 6 shows the Cox and KM graphs together for comparison. In Table 5, the main results associated with the DRT and sectorial covariates at each of the times are presented.

Based on the results observed in Table 5, it is possible to interpret that the sectorial variable is significant for the value of T1. It is noticeable that the p-value for the sectorial code is less than 0.005, indicating its statistical significance for this regression. As for DRT, such an assertion cannot be made, given its much higher p-value.

Meanwhile, the value of exp(coef) is very close to 1. This value indicates a 1% decrease in the variable's effect, meaning that although significant, the sectorial variable has a very low decay rate.

Despite being significant for T1, the sectorial covariate is not significant for T2, and for T3 it does not show any significant relevance. Looking at Table 7, it is noticeable that the p-value for sectorial at T2 is higher than at T1, but still shows some relevance. However, for T3, the value is even higher than that of DRT. The fact that the sectorial covariate is not significant for T3 may be one of the reasons why its p-value at T2 is much higher than at T1. This is due to T2

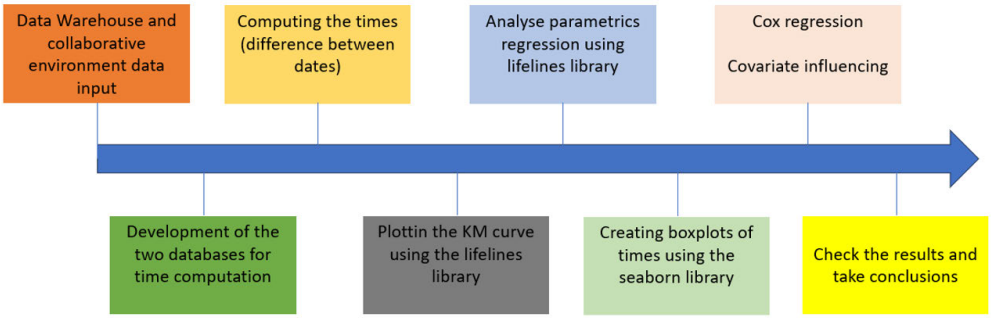


FIGURE 3. Sequence of steps.

TABLE 4. AIC values.

	Parametric regression					
	Weibull	Exponencial	LogNormal	LogLogistica	PieceWise	Gamma
T1	74567.23	74585.67	74794.61	74940.64	74593.31	74502.56
T2	82128.96	84541.79	82230.90	82387.57	83924.82	82010.18
T3	77850.22	79133.84	79458.66	78239.28	78386.60	77827.26

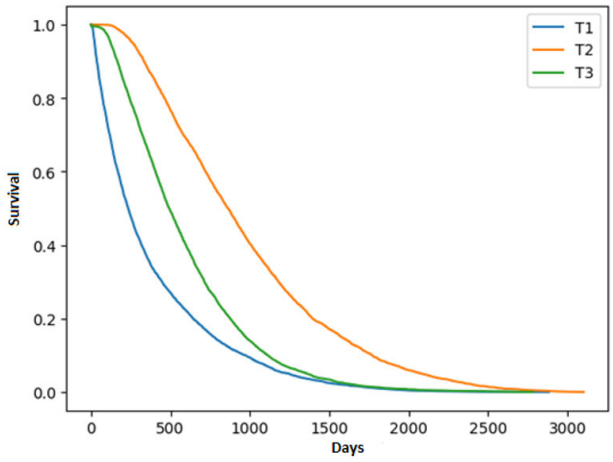


FIGURE 4. Analysis of survival times by KM.

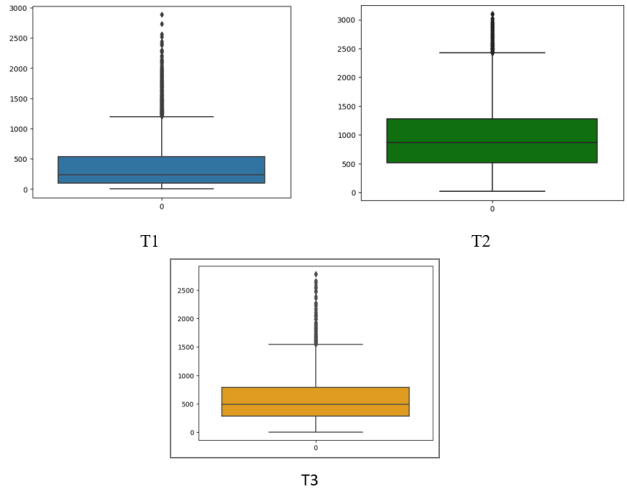


FIGURE 5. Boxplots for T1, T2 and T3.

TABLE 5. Main results of Cox for the database.

Time	Covariate	coef	exp(coef)	p
T1	Sectorial	-0.01	0.99	<0.005
	DRT	0.00	1.00	0.27
T2	Sectorial	-0.01	0.99	0.01
	DRT	0.00	1.00	0.99
T3	Sectorial	0.00	1.00	0.38
	DRT	0.00	1.00	0.31

being a composition of T1 and T3, as previously described. Therefore, the T2 interval ends up being the result of the sum of the other time intervals.

B. SURVIVAL ANALYSIS OF THE DATA WITH PAC/PCN

After cross-referencing the data obtained from the DW with the collaborative database, it was possible to identify the number of null companies that underwent PAC/PCN administrative procedures, totaling 2,633 cases. The subsequent procedures were the same as those in the total database. Figure 7 presents the graph obtained for the three times within the new database.

The cases with more likelihood of credit recovery ended up having shorter survival times. T1 took 211 days to reach 50% survival, and T2 now only has 4.5% of cases exceeding 1,825 days (five years), compared to the 8.5% in the previous database.

The time to complete 50% of the processes (T3) is 427 days, and 90% of cases are completed in 936 days. Both are lower numbers compared to those found in the non-specific database.

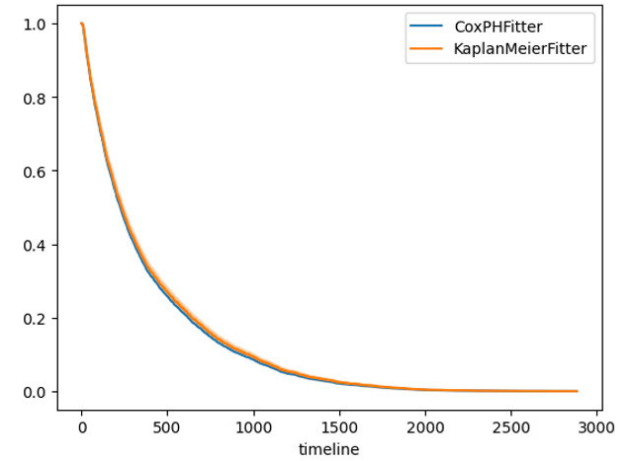


FIGURE 6. Cox regression and KM plots for T1.

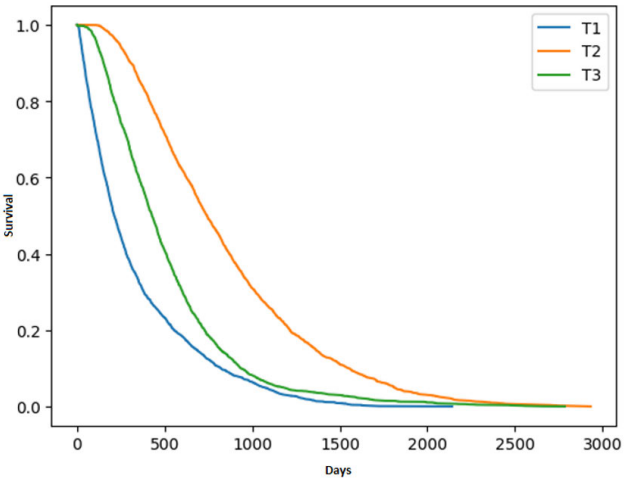


FIGURE 7. KM for T1, T2 and T3 in the new database.

TABLE 6. Summary of the efficiency increase.

Measure variable	Complete database	PAC/PCN database	Reduction
S(T1) = 0.5	233	211	9.4%
T2 >1,825 days	8.5%	4.5%	-
S(T3) = 0.5	492	427	13.2%
S(T3) = 0.9	1,114	936	16.0%

Just like in the complete database, boxplots were created for the database where administrative deepening was carried out. The result is presented in Figure 8.

A comparative look between Figures 5 and 8 reveals that there is greater efficiency on the part of SEFAZ-SP in cases that require the execution of PAC/PCN. In other words, when it is necessary to implement a more in-depth administrative process aimed at tax credit recovery, the state agency is more efficient.

Table 6 presents a summary of process improvement when working with the database of PAC/PCN cases.

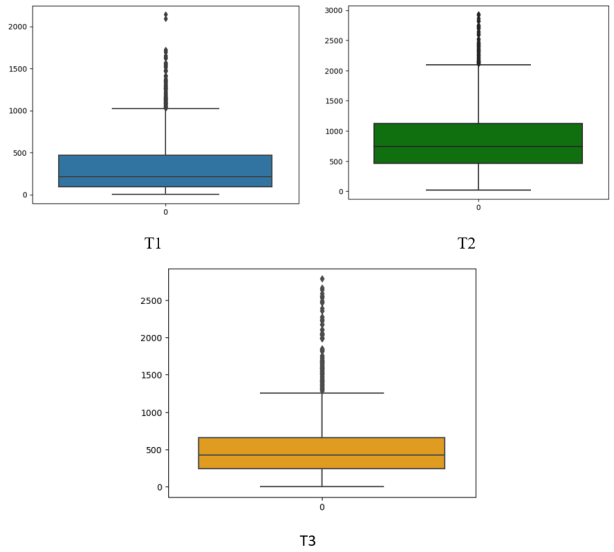


FIGURE 8. Boxplots for T1, T2 and T3 in the new database.

In the table is possible to see interesting results. The survival of half of the cases was reduced by 9.4%, demonstrating a greater ability to detect fraudulent companies. However, what is more striking is the steeper decline in the survival curve of T3. Administrative procedures in cases where tax credit recovery is required have a faster completion rate.

Cox regression analysis was performed for the variables T1, T2, and T3 regarding the DRT and sectorial covariates. In relation to the Cox regression for T1, a similar p-value result can be observed for the sectorial covariate. This indicates that the sectorial covariate continues to have significant relevance and maintains the decay rate of 1%. This fact can be observed in Table 7.

TABLE 7. Main results of Cox in new database.

Time	Covariate	coef	exp(coef)	p
T1	Sectorial	-0.01	0.99	<0.005
	DRT	0.00	1.00	0.74
T2	Sectorial	-0.01	0.99	0.07
	DRT	0.02	1.02	<0.005
T3	Sectorial	0.00	1.00	0.40
	DRT	0.03	1.03	<0.005

In Table 7, we have the Cox analysis for the T2 results with the new database. It is important to note that the p-value for the sectorial variable is seven times higher than in the previous database. However, in this new database, the DRT covariate becomes significant, unlike in the previous database. This result will also be noticeable for T3.

In addition to the relevance of DRT for T2 and T3, it shows an increase rate of 2% and 3% respectively. These data are based on the coefficient. The big issue is that, despite being significant, the rate of change is still low.

The significance of DRT in T2 results may suggest that cases belonging to this database, where PAC/PCN were executed, could be prioritized by some DRTs, leading to a

more expedited workflow. This priority may be linked to the issue of tax credit recovery, considering the concern with the five-year deadline for recovery. Other factors such as greater availability of tax auditors or simpler cases in certain regions could also be considered.

Delving deeper into this correlation could be quite interesting, as if some places are developing or using tools that can reduce the administrative process time, best practices can be disseminated. Additionally, it may be interesting to understand the aspects that could alter the coefficient value, allowing for a reduction in time.

V. DISCUSSION AND CONCLUSION

The use of survival analysis as a tool for monitoring the work related to null companies in SEFAZ-SP has proven useful. It was possible to perceive that the tasks have been carried out within temporal parameters that can be considered reasonable, given that the majority of them are completed within the timeframe for tax credit recovery.

Despite this, it may be interesting to use benchmarking tools with the purpose of reducing time T1. A target, for example, of $S(T1) = 0.5$ in less than 100 days would be very appealing, as it would significantly limit the survival of null companies and the loss to the state.

The reduction in T3 when PAC/PCN is required should also be highlighted as something productive. A PAC/PCN process is much more labor-intensive and tends to be more time-consuming than the regular nullification process. However, these cases are experiencing shorter survival times, demonstrating greater efficiency in their execution.

Another issue to be addressed is the values of T2 above Q3, as well as the outliers. Despite their reduction in the cases of the reduced database, the upper limit remained above 2,000. This may signify the loss of the possibility of recovering part of the tax credit, considering the timeframe exceeding 1,825 days. Examining whether the cases above Q3 have certain similarities can help in devising solutions to reduce them.

Although there are outlier cases and not all T2 values are within the ideal timeframe, SEFAZ-SP is efficiently addressing the cases in question. It is important to emphasize that these cases result from illicit activities, and fraudsters always seek to evolve their methods of defrauding public funds, making their detection more difficult.

The use of survival analysis has proven useful to present and understand the results being achieved by SEFAZ-SP. Expanding studies with more in-depth tools and analyzing other specific variables may lead to further insights and provide avenues for improving fiscal processes' efficiency.

REFERENCES

[1] P. A. Vieira, D. P. Pimenta, A. F. D. Cruz, and E. M. S. D. Souza, "Efeitos do programa de nota fiscal eletrônica sobre o aumento da arrecadação do estado," *Revista de Administração Pública*, vol. 53, no. 2, pp. 481–491, Apr. 2019.

[2] *Behind the Corporate Veil: Using Corporate Entities for Illicit Purposes*, OECD Publishing, Paris, France, 2001.

[3] R. Alexandre, *Direito Tributário Esquematizado*. São Paulo, Brazil: Método, 2016.

[4] E. A. Akartuna, S. D. Johnson, and A. Thornton, "Preventing the money laundering and terrorist financing risks of emerging technologies: An international policy Delphi study," *Technolog. Forecasting Social Change*, vol. 179, Jun. 2022, Art. no. 121632.

[5] M. Buffon, "A extrafiscalidade e direitos fundamentais: A introdução do interesse humano na tributação," *Revista da FESDT*, vol. 3 pp. 151–170, Jan. 2009.

[6] A. A. Becker, *Teoria Geral do Direito Tributário*. São Paulo, Brazil: Noeses, 2018.

[7] B. Appy, "Por que o sistema tributário brasileiro precisa ser reformado," *Revista Interesse Nacional*, vol. 8, no. 31, pp. 65–81, 2015.

[8] A. M. Moreira, "Não cumulatividade tributária no Brazil e no mundo: Origens, conceito e pressupostos," *Revista do Congresso Mineiro de Direito Tributário e Direito Financeiro*, vol. 1, pp. 49–84, Jan. 2011.

[9] E. R. Gomes and F. Guimarães, "A política de simplificação e renúncia fiscal para as micro e pequenas empresas no Brazil e o pacto federativo: Uma análise do simples nacional," *Teoria e Pesquisa*, vol. 21, no. 2, pp. 34–47, 2012.

[10] R. Biava Júnior and J. C. T. Oyadomari, "Impactos da Substituição Tributária do ICMS na lucratividade e nos custos empresariais," *Revista de Contabilidade da UFBA*, vol. 4, no. 2, pp. 71–91, Aug. 2010.

[11] *Estimativa da Carga Tributária Bruta do Governo Geral*, Secretaria do Tesouro Nacional, STN, Brasília, Brazil, 2022.

[12] J. D. Ducas, *Utilização de análise de Sobrevivência para Inidôneos: Uma Análise dos Tempos de Inaptidão e Cassação dos Casos de 2015 a 2023*. São Carlos, Brazil: Universidade de São Paulo, 2024.

[13] C. Pacini, W. Hopwood, G. Young, and J. Crain, "The role of shell entities in fraud and other financial crimes," *Managerial Auditing J.*, vol. 34, no. 3, pp. 247–267, Mar. 2019.

[14] F. Angélico, "São Paulo: Does corruption live next door: Shell companies and the real estate sector in the largest city in the southern hemisphere," *Transparency Int.*, vol. 1, pp. 1–28, Jan. 2017. [Online]. Available: https://knowledgehub.transparency.org/product/s%C3%A3o-paulo-does-corruption-live-next-door-shell-companies-and-the-real-estate-sector-in-the-largest-city-in-the-southern-hemisphere#: :text=Shell%20companies%20and%20the%20real%20estate%20sector,ownership%20in%20S%C3%A3o%20Paulo%20and%20identifies%20how

[15] J.-D.-J. Rocha-Salazar, M.-J. Segovia-Vargas, and M.-D.-M. Camacho-Miñano, "Detection of shell companies in financial institutions using dynamic social network," *Expert Syst. Appl.*, vol. 207, Nov. 2022, Art. no. 117981.

[16] B. L. C. Pérez, "Mercado de facturas apócrifas en México: Un análisis teórico de las operaciones simuladas como medio de evasión fiscal," *El Semestre de las Especializaciones*, vol. 1, no. 2, pp. 305–357, 2020.

[17] D. Singh, "Incorporating with fraudulent intentions: A study of various differentiating attributes of shell companies in India," *J. Financial Crime*, vol. 17, no. 4, pp. 459–484, Oct. 2010.

[18] C. Andrew and M. K. Baer, *How to Combat Value-Added Tax Refund Fraud*. Washington, DC, USA: International Monetary Fund, 2023.

[19] A. A. Aronowitz, D. C. G. Laagland, and G. Paulides, *Value-Added Tax Fraud in the European Union*. Amsterdam, The Netherlands: Kugler Publications, 1996.

[20] M. Griffioen and E. C. J. M. Van Der Hel-Van Dijk, "Tackling VAT-fraud in Europe: A complicated international puzzle," *Intertax*, vol. 44, no. 4, pp. 290–297, Apr. 2016.

[21] G. Poniatowski, A. Smietanka, and A. Skowronek, *VAT Compliance Gap Due to Missing Trader Intra-Community (MTIC) Fraud Final Report. Phase II*. Luxembourg, Luxembourg: Publications Office of the European Union, 2024.

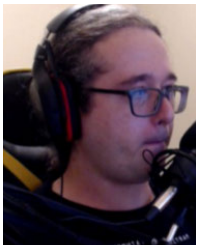
[22] S. Barajas, R. Campos, H. Sobarzo, and A. Zamudio, "Evasión fiscal derivada de los distintos esquemas de facturación," *México: El Colegio de México*, vol. 154, pp. 427–444, Jan. 2011. [Online]. Available: https://dave-donaldson.com/wp-content/uploads/CDPS.pdf

[23] P. Carrillo, D. Donaldson, D. Pomeranz, and M. Singhal, "Ghosting the tax authority: Fake firms and tax fraud in Ecuador," *Amer. Econ. Review: Insights*, vol. 5, no. 4, pp. 427–444, Dec. 2023.

[24] E. S. Subroto, D. D. Sidarta, and N. Soekorini, "Legal consequences for tax crimes with fictitious tax invoices," *Policy Law Notary Regulatory Issues (POLRI)*, vol. 3, no. 4, pp. 472–482, Nov. 2024.

[25] S. Wijaya and S. P. P. Surbakti, "Determinants of the VAT gap in the European union: An empirical evidence with corruption control as a moderating variable," *Educoretax*, vol. 4, no. 1, pp. 91–103, Feb. 2024.
[26] K. Singh and P. Best, "Anti-money laundering: Using data visualization to identify suspicious activity," *Int. J. Accounting Inf. Syst.*, vol. 34, Sep. 2019, Art. no. 100418.
[27] Q. Zheng, Y. Xu, H. Liu, B. Shi, J. Wang, and B. Dong, "A survey of tax risk detection using data mining techniques," *Engineering*, vol. 34, pp. 43–59, Mar. 2024.
[28] L. G. C. de Sousa and A. J. Rezende, "A eficiência das práticas fiscais digitais na redução do vat gap," *Cadernos De Finanças Públicas*, vol. 23, no. 1, pp. 218–227, Jun. 2023. [Online]. Available: https://omega.polsl.pl/info/article/PSLca60b7c6f6fe444e9e0308811314a9e6?r=publication&ps=20&tab=&title=Publication%2B%25E2%2580%2593%2BRecognizing%2Bthe%2Bfictitious%2Bbusiness%2Bentity%2Bon%2Blogistic%2Bregression%2Bbase%2B%25E2%2580%2593%2BSilesian%2BUniversity%2Bof%2BTechnology&lang=en
[29] A. Krysovatyy, K. Lipianina-Honcharenko, S. Sachenko, O. Desyatnyuk, A. Banasik, and I. Lukasevych-Krutnyk, "Recognizing the fictitious business entity on logistic regression base," in *Proc. IntelITSIS*, 2022, pp. 218–227.
[30] V. Aggarwal and K. Dharni, "Deshelling the shell companies using Benford's law: An emerging market study," *Vikalpa: J. Decis. Makers*, vol. 45, no. 3, pp. 160–169, Sep. 2020.
[31] M. Zumaya et al., "Identifying tax evasion in Mexico with tools from network science and machine learning," in *Corruption Networks* (Understanding Complex Systems), O. M. Granados and J. R. Nicolás-Carlock, Eds. Cham, Switzerland: Springer, 2021. [Online]. Available: https://doi.org/10.1007/978-3-030-81484-7_6
[32] F. N. D. Oliveira and L. P. G. D. Santos, "Estratégias para combater a sonegação fiscal: UM modelo para o ICMS baseado EM redes neurais artificiais," *Revista de Gestão, Finanças e Contabilidade*, vol. 10, no. 1, pp. 42–64, Feb. 2021.
[33] R. S. Fontes, M. C. Júnior, F. R. de Oliveira, and R. A. de Medeiros Valentim, "Avaliação experimental de um classificador para apoiar a detecção de fraudes em notas fiscais públicas de medicamentos," *Revista Gestão Tecnologia*, vol. 24, no. 1, pp. 112–148, 2024.
[34] R. R. Aguilo and G. Quadrelli, "Aplicação da lógica nebulosa na detecção de Fraude de ICMS," *Revista de Engenharia da Universidade Católica de Petrópolis*, vol. 16, no. 2, pp. 53–63, 2022.
[35] S. Neto, S. B. S. Monteiro, J. C. F. Souza, and R. M. Chaim, "Using data envelopment analysis to ranking ICMS's taxpayers," *Math. Comput. Sci.*, pp. 1–32, Jun. 2018.
[36] V. Di Oliveira, R. Chaim, L. Weigang, S. Neto, and G. Filho, "Towards a smart identification of tax default risk with machine learning," in *Proc. 17th Int. Conf. Web Inf. Syst. Technol.*, 2021, pp. 422–429.
[37] G. S. L. Gomes and R. Balaniuk, "Identifying tax evasion in shell companies and fraudulent credits of Brazil's state value-added tax (ICMS)," *Revista de Administração Pública*, vol. 58, pp. e2023, Sep. 2024.
[38] E. A. Colosimo and S. R. Giolo, *Análise de Sobrevivência Aplicada*. São Paulo, Brazil: Editora Blucher, 2021.
[39] F. E. Harrell, *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY, USA: Springer, 2001.
[40] E. Borelli and B. D. Lucena, "Recuperação de crédito estimada pela análise de sobrevivência: Credit recovery estimated by survival analysis," *Revista de Gestão e Secretariado*, vol. 13, no. 3, pp. 949–973, Nov. 2022.
[41] B. Wroblevski and M. de Sousa Lucas, "Desemprego e hiato salarial de gênero na reinserção na ocupação: Uma análise para o Brazil no contexto da crise econômica de 2014," *RDE-Revista de Desenvolvimento Econômico*, vol. 3, no. 50, pp. 119–147, 2021.
[42] B. Showkat and D. Singh, "Perceiving moisture damage of asphalt mixes containing RAP using survival analysis based on Kaplan–Meier estimator and cox proportional hazards model," *Construct. Building Mater.*, vol. 320, Feb. 2022, Art. no. 126249.
[43] M. Boissonneault and J. de Beer, "Assessing the capacity to work among older workers: A survival analysis of retirement behavior," *Work, Aging Retirement*, vol. 8, no. 1, pp. 38–50, Jan. 2022.
[44] O. V. Kuzmenko, T. V. Dotsenko, and L. O. Skrynka, "Economic and mathematical modelling of the effectiveness of the national system for combatting cyber fraud and legalisation of criminal proceeds based on survival analysis methods," *Sci. Bull. Mukachevo State Univ. Ser. 'Econ.'*, vol. 8, no. 1, pp. 144–153, Mar. 2021.
[45] J. L. Tonetto, J. M. Pique, A. Fochezatto, and C. Rapetti, "Tax evasion and company survival: A Brazilian case study," *Economies*, vol. 12, no. 11, p. 286, Oct. 2024.
[46] P. C. Austin, D. S. Lee, and J. P. Fine, "Introduction to the analysis of survival data in the presence of competing risks," *Circulation*, vol. 133, no. 6, pp. 601–609, Feb. 2016.
[47] P. Wang, Y. Li, and C. K. Reddy, "Machine learning for survival analysis: A survey," *ACM Comput. Surv. (CSUR)*, vol. 51, no. 6, pp. 1–36, 2019.
[48] R. C. S. Nogueira, *Análise de Sobrevivência e Tempo de Vida: Modelos e Aplicações*. Rio de Janeiro, Brazil: Elsevier Brasil, 2014.
[49] M. T. Amin and F. Khan, "Dynamic process safety assessment using adaptive Bayesian network with loss function," *Ind. Eng. Chem. Res.*, vol. 61, no. 45, pp. 16799–16814, Nov. 2022.

JÔNATAS DELL DUCAS received the master's degree in engineering from Federal Fluminense University, in 2009, the master's degree in pipeline engineering from PUC-RJ, in 2010, the master's degree in public management from UNIFESP, in 2019, and the M.B.A. degree in data science from ICMC, USP, in 2024. He has held numerous roles at SEFAZ-SP. He is currently a Tax Auditor with the Department of Finance, State of São Paulo, Brazil. He is also the Head of the Fiscal Office, Americana. He is also a Professor with the Mechanical Engineering Department, Americana College (FAM).



RODRIGO COLNAGO CONTRERAS received the bachelor's degree in applied and computational mathematics and the master's degree in mathematics from the Institute of Biosciences, Letters, and Exact Sciences (IBILCE), São Paulo State University Jélio de Mesquita Filho (Unesp), in 2013 and 2015, respectively, and the Ph.D. degree in computer science and computational mathematics from the Institute of Mathematics and Computer Science (ICMC), University of São Paulo (USP), in 2020. He was a Postdoctoral Fellow with IBILCE, Unesp, funded by FAPESP, in 2024. He also holds two volunteer postdoctoral positions: one at IBILCE, Unesp, and another at ICMC, USP, which were completed in 2023 and 2022, respectively. He is currently an Adjunct Professor with the Institute of Science and Technology, Federal University of São Paulo (UNIFESP). His experience lies in the fields of signal processing and computational intelligence, with a primary focus on the following topics, such as fraud detection in biometric authentication systems, the development of bio-inspired strategies applied to computational optimization, and analytical and visual pattern recognition. He was the Runner-Up in the 2021 Liveness Detection Competition (LivDet) and received an honorable mention in LivDet 2023, ranking among the top three competitors and achieving the highest accuracy in challenge 2.