EGOV-CeDEM-ePart

Data Science Methods and Techniques for Goods and Services Trading Taxation: a Systematic Mapping Study

Douglas Silva*, Sergio Carvalho**

**Federal University of Goias, Goiania-GO, Brazil, douglas.bernardes@inf.ufg.br*
***Federal University of Goias, Goiania-GO, Brazil, sergiocarvalho@ufg.br*

Abstract: Taxation on goods and services trading operations is the main revenue source for States and Provinces around the world. Collecting such taxes, however, constantly faces a series of challenges, ranging from the incorrect filling of tax documents involved (which leads to the incorrect calculation of the due tax) to attempts of tax fraud. As this context involves analyzing a very large amount of data, data science techniques appear as an interesting alternative to provide effective solutions to the problems that arise. This article describes a systematic mapping of the literature aimed to identify how data science methods and techniques have been applied to this context and how the problems inherent in this domain are being handled. Results show that there are very well-defined categories of problems being researched in this area, and that data science can efficiently be used to improve the collection of these types of taxes.

Keywords: value-added tax, goods and services tax, sales tax, data science, systematic mapping study.

1. Introduction

Taxes are compulsory financial charges imposed on an individual or entity taxpayer by Government in order to fund public expenditures (Mathews et al. 2018, Rad et al. 2015). They are regulated by specific laws that describe their composition, their collection and compliance processes and even resulting revenue application — if needed.

Taxation on the sale of goods and provision of services is the main source of revenue for most states and provinces around the world and therefore a relevant kind of indirect taxes. It is a tax applied to each link in the consumption chain, and which generates, at each transaction, one or more tax documents with a complete record of the items and parties involved in the transaction (Yu et al. 2019), including their tax classification and due tax rate. This information is usually registered in an electronic *invoice*.

Trading operations taxation has been implemented in different ways, but usually as a non-cumulative tax due proportionally to each taxpayer that compose the consumption chain. Some taxes of this nature are *Sales Tax* in the United States (Buxton et al. 2019), *ICMS* (that stands for *Tax on Circulation of Goods and Services*) in Brazil, *GST (Goods and Services Tax)* in countries like Australia, Canada, Singapore and recently India (Mathews et al. 2018, Mehta et al. 2018), and variations of *VAT (Value-Added Tax)* — that is used in most countries, like China (Yu et al. 2019) and European Union.

Tax law applied to goods and services trading, however, in addition to being complex, constantly changes, and the taxpayer is not always up to date on the tax rules applicable to each product he sells, or to each service he is willing to provide (Lahann et al. 2019). Tax benefits and exemptions are also granted seasonally and for a specific period of time to specific segments of taxpayers, and all of these possibilities directly impact the tax bookkeeping declared by all of them.

These situations allow taxpayers, intentionally or unintentionally, to generate damage to public treasury and consequently undermine provision of public services to the citizen. It becomes then necessary to not only collect the taxes, but to verify if taxpayers done it properly and to proceed with debt collection when necessary (Abe et al. 2010).

The analysis of tax compliance information is currently a tax auditor's responsibility. The limited number of human resources, associated with the volume of generated information, however, makes conventional procedures ineffective and inefficient (Wang 2012). It is necessary to direct auditor's focus, so that he acts less in formalities and more in signs of anomalies or fraud (Basta et al. 2009).

Although technological development has enabled the automation of operational processes, analysis of massive amounts of data aimed at identifying anomalies, inconsistencies and behavior patterns for detecting evidence of fraud and tax non-compliance is still a challenge.

Methods traditionally used to solve aforementioned problems are time-consuming, costly and imprecise, and in *big data* scenario it is impractical.

Although government have been analyzing tax data for ever, and analytics, AI and modern technology help them do better, big data in this domain is recent. GST itself has been implemented in India, e.g., only on 2017 (Das et al. 2017), and electronic invoicing was made mandatory for Italian companies just in January 2019 (Bardelli et al. 2020). Problems related to data characteristics — as volume, inconsistency and incompleteness — are hence also recent, and mapping how that computing areas deals with this domain becomes needed.

Data science models and strategies are, in general, useful to a context analogous to this. However, their applicability varies according to the characteristics of available data. Therefore, it is necessary to identify which techniques could be used and for which reasons, and also to identify aspects of these models and strategies that have not yet been addressed.

This article presents a systematic mapping of the literature that intends to comprehend the domain of tax collection in goods and services trading operations and how data science has been used to solve problems that emerge in this context, identifying a possible consensus or good practices in handling these situations. However, it is important to understand that this is a systematic mapping and not a systematic review. Its main objective is to map the domain area, its datasets characteristics and how they influence researchers' choice about which techniques to use, so it can help to clarify the way to future researchers. Techniques itself and how they handle the problems found in this domain here would better be explored in a systematic review, with research questions aimed to this end.

The remainder of the paper is organized as follows. The second section introduces concepts involving tax due to trade of goods and services, and data science techniques. Third section presents the materials and methods used to define the systematic mapping protocol. Fourth section displays the results of the systematic mapping performed, while the fifth section analyzes these results. The sixth section discusses the results found and perceptions over them. Seventh and last section presents the final considerations on the performed procedure.

2. Review Protocol

This systematic mapping of literature followed the procedure described in (Petersen et al. 2008). As part of the process, a research protocol was defined, which is detailed in the following items.

2.1 Research Questions and Search Strategy

This mapping sought to establish the state of the art in scientific research conducted in the field of data science in the domain of tax data from goods and services trading operations. The specific review questions addressed were:

1) What problems in the domain of goods and services trading taxation have been studied in the area of data science?
2) What types of techniques and learning strategies have been applied?
3) Which data sources are used in the analysis?
4) Do the selected attributes vary according to the region / location (where the problem occurs)?
5) Which datasets are used?
6) How big are these datasets?
7) Has the volume of data been a complicating factor for the analysis?
8) How has the problem of volumetry been dealt within this context?

From the main keywords identified in these research questions, an initial *string* was defined and calibrated through a pilot search in digital libraries from IEEE Xplore, ACM and Scopus, in order to reduce likelihood of polarization.

Assessment also took into account that taxes with these characteristics are called *Sales Tax* in the United States, *GST* in India (among other countries) and *VAT* in the European Union and China. By adding these three variations, we apparently reached all (or most of) aimed publications.

The evaluation of pilot search results led to the following search *string*:

("value-added tax" OR "goods and services tax" OR "sales tax") AND
("data science" OR "artificial intelligence" OR "data mining" OR "machine learning"
OR "neural network")

After defining the *string*, we selected most common publications databases to perform systematic mappings and reviews of the literature in the area of Software Engineering (Dyba et al. 2005), namely: ACM Digital, IEEE Xplore, ScienceDirect, SpringerLink and Scopus.

2.2 Criteria for Study Selection

Once primary studies were obtained from the aforementioned research sources, inclusion and exclusion criteria were applied to them in order to select those notably relevant to the systematic mapping objetive.

Thereby studies were considered eligible if they had tax collection in the trade of goods and services as motivation and as scenario for implementation/validation of the proposed method, or whose method had applicability to a context similar to that mentioned.

Selected studies were also evaluated for their relevance (they should bring up data science techniques) and formality, being excluded from the review publications that did not meet the aforementioned eligibility criteria and:

- Papers that do not propose the use of data science methods or techniques to solve a problem found in the mentioned domain;
- Papers that do not present the method proposed to solve the problem;
- Publications that have not been subjected to peer review;
- Publications that are not in English or Portuguese;
- Publications without the full text or unavailable;
- Repeated publications.

The number of excluded papers, as well as the reason for their exclusion, were recorded as the articles were evaluated.

The process for selecting studies followed the one proposed by Meline (2006):

- Step 1 (screening): eligibility criteria were applied to the search results through a preliminary evaluation of their title, abstract and keywords;
- Step 2: studies were then discarded if they meet one or more exclusion criteria, being evaluated the same elements as step 1;
- Step 3 (full text review): eligibility and exclusion criteria were then applied to remaining/accepted studies, now evaluating their full text.

2.3 Methods for Data Extraction and Study Synthesis

After evaluating full text of accepted articles, we filed them using a data extraction form, equalizing the results found in each research and allowing their analysis and summarization.

The following information was extracted from the selected articles: (i) title, authors and year of publication; (ii) research problem; (iii) proposed data analysis technique; (iv) learning paradigm and strategy, if it fits; (v) data sources used; (vi) datasets used; (vii) datasets volumetry (and inherent problems); and (vii) gaps observed by the researchers.

3. Results

Following the process described in Section 2.2, we carried out a literature search on December 29, 2020, which initially returned 867 papers. Of these, 24 papers came from the IEEE Xplore digital library, 66 from the ACM database, 258 from Scopus, 218 from the ScienceDirect digital database and 301 from the SpringerLink database.

After the initial reading of abstract, keywords and title, 71 duplicate articles were found and discarded, and 747 articles were also rejected because they did not meet the eligibility criteria. In these 747, 742 were excluded for not having as motivation and as validation scenario problems related to tax collection in the trade of goods and services, not even by similarity, and another 5 were rejected for proposing computational techniques not related to data science (such as blockchain or ontologies) aiming some other aspect of the mentioned tax domain.

It is important to highlight that, according to our view, the string contains only the terms necessary to direct the results: three variations of how this consumption tax is called around the world, and the name of techniques or areas that could indicate the application of Data Science to this domain. Aiming avoid false positives, even (known) abbreviations of these taxes were removed from the search string. However, several articles mention, often only once, the tax itself, or how useful it would be to use data science to deal with it. Their application, however, or the domain itself, were not the focus of these articles — and whenever that happened, they were discarded.

Thus, from 747 articles, 49 remained for full text evaluation. These 49 articles were obtained and evaluated as full text, and we found that 4 (four) of them should be rejected because they did not have, as their main motivation, the improvement of tax collection in goods and services trading operations (Hoglund 2017, Kong et al. 2014, Krzikallová 2020, Meservy 1992), and three of them were discarded by the exclusion criterion related to the non-use of data science methods or techniques to solve a problem found in the mentioned domain (Akinboade et al. 2009, Bogdanov et al. 2015, Cai et al. 2011). In addition, two of them were rejected for not been written in Portuguese or English (Cadena et al. 2019, Hasanli et al. 2014). Finally, three of them were not even accessible (Loan et al. 2018, Mathews et al. 2018, Vicente et al. 2016).

After examining the full texts 37 articles remained. We applied then the data extraction form defined in Section 2.3 and carried out the analyzes show bellow.

4. Results Analysis

The evaluated publications start, chronologically, with the proposal of Voorhees (2006) to carry out a forecast of goods and services trading revenue through neural networks. He mentions that using a neural network for this purpose is better than performing a regressive analysis, since it is limited to the extent that independent variables cannot be correlated, residues must be independent and errors must be equally distributed.

Defa and Jing (2010) and Buxton et al. (2019) also present approaches to forecasting revenue from this tax. Defa and Jing combines three prediction models: a regression equation model, a time series model and gray model, maximizing their combined accuracy and reaching less than 5% error. Buxton et al., with a more recent work, also combine two models — Auto-regressive MultiLayer Perceptron and LSTM — and are effective in forecasting the collection of different product categories, such as fuels, construction and medicines.

The expected tax, however, does not always match collected one. The process of verifying — and seeking — the correctness of the tax declared by the taxpayer is known as *tax compliance*. In this sense, Lahann et al. (2019) presents an anomaly detection approach, in order to identify obvious transactions that have a high degree of probability of being associated to a false tax code (and, consequently, lead the taxpayer to pay an undue tax and, in most cases, a smallest one). In the same line, Fjeldstad et al. (2020) proposes a model based on a decision tree that verifies whether the expected behavior and the taxpayer documents correspond to the tax operation planned for him. Mehta et al. (2019), to increase compliance levels, propose a regression model to identify defaulting debtors and friendly Android apps to assist auditors in collecting tax. However, they also deal with another aspect in the quest to guarantee the correct collection: the verification of tax evasion. To do this, they explore the detection and analysis of a tax evasion mechanism, known as *circular trading*, using advanced social network and algorithmic analytical techniques.

Mehta et al. (2019) have published a series of surveys involving the analysis of tax data and the detection of tax fraud and tax evasion behavior by the taxpayer. Only from their work group (apparently) 8 (eight) other articles were selected for full-text review in this systematic mapping.

Mathews et al. (2018) had already started exploring the *circular trade* problem. In this type of transaction, a group of merchants "manufactures" sales and (or) purchases between themselves, which results in the flow of goods in a circular manner without any added value: for the collecting entity, the taxpayer (or the group) is entitled to an abatement of the tax to be paid, since the nature of the tax indicates that it must pay only the tax on the value it added to the product. However, as there was no acquisition initially, this "credit" is free, and in fact the taxpayer is only withholding what would be due to him for selling the goods.

To solve this problem, the entire series of articles published by the group seeks to model the relationships between taxpayers, as well as the commercial transactions that take place between them, in the form of a graph (where the contributors are the vertices and their relations, the edges), and so that machine learning models can identify patterns and outliers in these relationships.

In another paper by Mathews et al. (2018), the classification of suspected contributors is given in three steps. In the first, taxpayers are clustered based on 7 correlations between variables such as tax paid, the total amount of sales, the amount of tax paid in cash and the amount of tax-free sales. They then use an application of Benford's law to classify taxpayers in each cluster as "trusted" and "suspect". Finally, it uses data from trusted taxpayers to create a linear regression model, which is then applied to suspect taxpayers to predict the amount of tax each tends to evade in the next period.

Mehta et al. (2018) try to predict whether a taxpayer tends to declare the tax appropriately in the next reference. They are based on the behavior of statements of each company in previous years, on the turnover of the current month, on the value of interactions with other taxpayers and on the average absolute deviation obtained by the law of Benford, when applied to taxpayer sales transactions. It also uses information from transport communications to carry out associations (all transport of products demands this auxiliar document).

5. Discussion

Table 1 shows a comparison of accepted papers. As can be seen, there is a preponderance on using machine learning unsupervised techniques in two major categories of tax problem, which are even related: fraud and tax evasion.

Table 1. Papers grouped by the learning paradigm used to solve each tax problem found

Tax Collection Problem	**Learning Paradigm**
Revenue Forecasting	Supervised *(Buxton et al. 2019, Defa and Jing 2010, Voorhees 2006)*
Compliance	Supervised *(Fjeldstad et al. 2020, Lahann et al. 2019, Mehta et al. 2019)*
Debt Collection	Reinforcement *(Abe et al. 2010)* Statistical learning *(Holkova and Falat 2017)*
Fraud	Supervised *(Basta et al. 2009, Castellón Gonzaléz et al. 2013, Rad and Shahbahrami 2015, Yu et al. 2019)* Unsupervised *(Castellón Gonzaléz et al. 2013, Junqué de Fortuny et al. 2014, Mathews et al. 2018, Mehta et al. 2019, Mehta et al. 2019, Mittal et al. 2018, Priya et al. 2019, Vanhoeyveld et al. 2020, Zha 2020)* Reinforcement *(He et al. 2020)*
Evasion	Supervised *(Didimo et al. 2020, Mathews et al. 2018, Mathews et al. 2021, Rahimikia et al. 2017, Wu et al. 2012)* Unsupervised *(Assylbekov et al. 2016, González-Martel et al. 2020, Mathews et al. 2018, Mehta et al. 2019, Mehta et al. 2018, Mehta et al. 2019, Mehta et al. 2020, Wang 2012, Wu et al. 2020)* Semi-supervised *(Kleanthous and Chatzis 2020)* Positive learning *(Wu et al. 2019)* Transfering learning *(Zhu et al. 2018)*

Evasion occurs when any action by the taxpayer leads to the non-collection by the Public Administration of the taxes due to it. It can occur intentionally or not, but regardless it causes damage to the treasury, and for this reason it is combated. Fraud is a more specific case of evasion, in which the taxpayer (or a group of taxpayers) intentionally uses techniques or subterfuge to avoid being held responsible for the purchase and sale of goods they carry out. The most prominent of these, according to the results of systematic mapping, is *Circular Trading* (Mathews et al. 2018, Mathews et al. 2018, Mehta et al. 2019, Mehta et al. 2019, Priya et al. 2019, Mathews et al. 2021). However, there are other actions, such as the indication of a false operating address to get rid of tax obligations — known as *Residence Fraud* (Junqué de Fortuny et al. 2014) — and clandestine transportation of goods without a tax document.

As we analyze the results of the mapping, it is clear that the techniques and learning paradigms vary widely, but in general are associated with the characteristics of the data available in each context.

When it comes to a problem that involves historic of carried out operations, such as audits already carried out or collection from previous months, the paradigm is usually supervised, since the data tend to be labeled. This is also the case for tax compliance, as it is inherent to it to know the expected tax classification for each item and to check if proper rate has been assigned to it.

Tax fraud or evasion cases, on the other hand, can be dealt under both points of view. If data analysis makes use of information from audits already carried out, with proofs that a certain behavior was actually due to a "fraudulent" contributor, learning will be supervised and the algorithm will use the characteristics associated with the given label to rank the next contributors.

This is a rarer case, however, as the volume of audits performed and recorded is still small compared to the volume of tax documents issued. Therefore, the trend observed in systematic mapping is that the algorithms and learning techniques use the relationships between the taxpayers, and the commercial transactions carried out by them, to identify patterns and outliers that indicate suspicious behavior in an effective and efficient way.

It is also worth noting that the use of machine learning in this domain is recent. According to the mapping, 75% of the elected works carried out in this area were published in the last 5 years.

This is due, in part, to the fact that the tax documents processed in the operations of trade in goods and services have only recently become electronic. In the state of Goias, e.g., they are 100% electronic since 2018, only.

Finally, it is necessary to highlight that the volume of tax data to be processed during the learning process was not mentioned as a problem. However, this may be due to a fact mentioned in several studies: fiscal secrecy prevents researchers outside Revenue agencies from having access to data from commercial transactions, limiting the scope of the proposals.

This, however, could be a new opportunity when it comes to evaluating new learning techniques, if access to tax data is granted.

6. Threats to Validity

Despite the mapping's systematic character, some aspects are threats to its validity. The main one is due to a characteristic inherent to a mapping or systematic review: when addressing specific research questions, and for this purpose choosing the most appropriate terms for the search string, search may fail to return interest results to the purpose of the review or mapping — just by not matching the chosen terms (Kitchenham et al. 2007).

For this work's matter, we defined that one of the mandatory expressions would be *value-added tax* (with its syntactic variations), due to its recurrence as a tax on operations in the trade of goods and services in different parts of the world. However, its acronym (VAT) was not included, as well known as, but associated with the most diverse expressions (such as *Visceral Adipose Tissue*, in medical articles). In contrast, articles of interest in this research that use only the known acronyms of the surveyed taxes (VAT, GST), without naming them in full, were not returned by this review.

Another threat to validity is due to the fact that the mapping was carried out by a single reviewer, which may have biased in some way papers' interpretation.

7. Concluding Remarks

The systematic mapping study presented here showed, within the scope of the main digital libraries used to index studies published in the field of Computing, the state of the art of the proposed approaches to deal with aspects related to tax collection in operations of trade in goods and services through data science methods and techniques.

Mapping showed that there are five major problems researched by the scientific community in this context, with a greater focus on identifying and predicting of tax evasion behaviors by the taxpayers, whether due to incorrect filling of tax documents or intentional attempts at tax fraud and evasion.

The mapping also showed that each of these problems requires specific data analysis methods and techniques, and that the nature of these data leads to the choice of the appropriate learning technique for each case. To address tax compliance (verifying if proper rate is being applied to each product), for example, characteristics related to each tax class are labeled and a supervised learning algorithm is needed to classify products and taxpayers. In order to detect tax evasion or fraud attempts, such as *circular trading*, not only purchase and sale operations are analyzed, but also the relationships between taxpayers, in order to identify outliers in their behavior. For this, an unsupervised learning technique for clustering these taxpayers seems to be more suitable.

Regarding the datasets used, there are two considerations. Unlike the initial suspicion, the volume of data was not mentioned — in general — as an issue to be handled. On the other hand, this may be due to the fact that most returned papers found it difficult to access tax data, due to confidentiality involved, which limited the amount and variability of data used in the validation of the proposed methods. It also guided — and maybe biased — the choice of the learning technique to be used in some cases.

Major implications for future research include a need for more taxpayers' behavior analysis variations. As data is limited — in amount and depth, by confidentiality — only some aspects of taxpayer behavior, as amount of sales and related tax, are usually investigated. Some works have been done around fraud techniques as circular trading and residence fraud, mas it is still limited. Taxpayers use regulation gaps in tax domain to apply fraud without breaking out tax procedures, and therefore not being seen as an anomaly. Tax benefits and exemptions, granted seasonally and for a specific period of time to specific segments of taxpayers, are also a huge opportunity for tax evaders. This exceptions and unusual behaviors must be taken into account and be added to current models for improvement and performance analysis.

Furthermore, it would be interesting to systematically evaluate techniques current proposed to handle tax evasion, how they arrange to adapt incomplete and inconsistent tax data and if a consensus emerge of it. This could be proper done with a systematic literature review focused on data science methods and techniques specifically proposed for tax evasion and fraud behavior.

Finally, it lacks an evaluation of efficiency loss due to incomplete tax data, by the confidentiality issue, and a definition of how to definitely deal with this problem. It could be achieved throw a comparison of performance and effectiveness between a complete and incomplete data scenarios.

References

Abe, N., & Melville, P., & Pendus, C., & Reddy, C.K., & Jensen, D.L., & Thomas, V.P., & Bennett, J.J., & Anderson, G.F., & Cooley, B.R., & Kowalczyk, M., & Domick, M., & Gardinier, T. (2010). *Optimizing debt collections using constrained reinforcement learning*. Proceedings of the 16th ACM SIGKDD. p. 75–84.

Akinboade, O.A., & Kinfack, E.C., & Mokwena, M.P., & Kumo, W.L. (2009). *Benchmarking tax compliance efficiency among south african retail firms using stochastic frontier approach.* 32(13), 1124–1146.

Assylbekov, Z., & Melnykov, I., & Bekishev, R., & Baltabayeva, A., & Bissengaliyeva, D., & Mamlin, E., & Czarnowski, I., & Caballero, A.M., & Howlett, R.J., & Jain, L.C. (2016). *Detecting Value-Added Tax Evasion by Business Entities of Kazakhstan*, pp. 37–49.

Bardelli, C., & Rondinelli, A., & Vecchio, R., & Figini, S. (2020). *Automatic electronic invoice classification using machine learning models.* Machine Learning and Knowledge Extraction 2(4), 617–629.

Basta, S., & Fassetti, F., & Guarascio, M., & Manco, G., & Giannotti, F., & Pedreschi, D., & Spinsanti, L., & Papi, G., & Pisani, S. (2009). *High quality true-positive prediction for fiscal fraud detection*. pp. 7–12.

Bogdanov, D., & Jõemets, M., & Siim, S., & Vaht, M., T., O., R., B. (2015). *How the estonian tax and customs board evaluated a tax fraud detection system based on secure multi-party computation*. vol. 8975, pp. 227–234.

Buxton, E., & Kriz, K., & Cremeens, M., & Jay, K. (2019). *An auto regressive deep learning model for sales tax forecasting from multiple short time series*. Intern. Conf. on Machine Learning and Applications. 1359-1364.

Cadena, M., & Morán, E. (2019). *Analysis for possible tax evasions from the value added tax in ecuador using an stochastic model with a non-parametric technique*. pp. 428–438.

Cai, D., & Zhang, A., & Cai, J. (2011). *The improvement on china's regional standard value added tax revenue estimate method - the construction, application and verification of standard rate model*. pp. 783–786.

Castellón González, P., & Velásquez, J.D. (2013). *Characterization and detection of taxpayers with false invoices using data mining techniques.* 40(5), 1427–1436.

Das, S., & Kolya, A.K. (2017). *Sense gst: Text mining & sentiment analysis of gst tweets by naive bayes algorithm*. pp. 239–244.

Defa, C., & Jing, C. (2010). *Construction of combination forecasting model and related validation – based on combined forecast of sales tax and enterprise income tax in heilongjiang province.* pp. 328–331.

Didimo, W., & Grilli, L., & Liotta, G., & Menconi, L., M., F., P., D. (2020). *Combining network visualization and data mining for tax risk assessment.* pp. 16073–16086.

Dyba,T., & Kitchenham,B.A., & Jorgensen,M. (2005). *Evidence-based software engineering for practitioners*. 58-65.

Fjeldstad, O.H., & Kagoma, C., & Mdee, E., & Sjursen, I.H., & Somville, V. (2020). *The customer is king: Evidence on vat compliance in tanzania*. 128, 104841.

Junqué de Fortuny, E., & Stankova, M., & Moeyersoms, J., & Minnaert, B., & Provost, F., & Martens, D. (2014*). Corporate residence fraud detection*. p. 1650–1659. KDD '14.

González-Martel, C., & Hernández, J.M., & Manrique-de Lara-Penãte, C. (2020). *Identifying business misreporting in vat using network analysis.* p. 113464.

Hasanli, Y., & Agayev, S. (2014). *Assessment of tax evasion risks for vat payers.* 153(3), 487–495.

He, Y., & Wang, C., & Li, N., & Zeng, Z. (2020). *Attention and memory-augmented networks for dual-view sequential learning*. Proceedings of the 26th ACM SIGKDD. p. 125–134. KDD '20.

Hoglund, H. (2017). *Tax payment default prediction using genetic algorithm-based variable selection.* 88, 368–375.

Holkova, B., & Falat, L. (2017). *Statistical learning as a tool for optimizing the level of excise tax of mineral oils in slovakia.* 192, 318–323

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering.*

Kleanthous, C., & Chatzis, S. (2020). *Gated mixture variational autoencoders for value added tax audit case selection.* 188, 105048.

Kong, D., & Saar-Tsechansky, M. (2014). *Collaborative information acquisition for data-driven decisions.* 95, 71-86.

Krzikallová, K., & Tosenovsk`y, F. (2020). *Is the value added tax system sustainable? the case of the czech and slovak republics*. 12(12).

Lahann, J., & Scheid, M., & Fettke, P. (2019). *Utilizing machine learning techniques to reveal vat compliance violations in accounting data*. IEEE 21st Conference on Business Informatics (CBI). vol. 01, pp. 1–10.

Loan, N.T., & Hac, L.D., & Anh, N.V.H., & Anh, L.H., & Dong, L.S., & Kreinovich, V., & Thach, N.N. (2018). *Application of Statistical Methods for Tax Inspection of Enterprises: A Case Study in Vietnam*. pp. 648–655.

Mathews, J., & Mehta, P., & Babu, C.S., & Kasi V. Rao, S.V. (2018). *An algorithmic approach to handle circular trading in commercial taxation system*. pp. 67–75.

Mathews, J., & Mehta, P., & Babu, C.S., & Kasi V. Rao, S.V. (2018). *Clustering collusive dealers in commercial taxation system.* Advances in Intelligent Systems and Computing, vol. 869, pp. 703–717.

Mathews, J., & Mehta, P., & Kuchibhotla, S., & Bisht, D., & Chintapalli, S.B., & Rao, S.V.K.V. (2018). *Regression analysis towards estimating tax evasion in goods and services tax.* IEEE/WIC/ACM WI. 758-761.

Mathews, J., & Mehta, P., & Suryamukhi, K., & Babu, S. (2021). *Link prediction techniques to handle tax evasion*. 8th ACM IKDD CODS and 26th COMAD. pp. 307–315.

Mehta, P., & Mathews, J., & Bisht, D., & Suryamukhi, K., & Kumar, S., & Babu, C.S., & W., A., G., K. (2020). *Detecting tax evaders using trustrank and spectral clustering*. vol. 389 LNBIP, pp. 169–183.

Mehta, P., & Mathews, J., & Kasi V. Rao, S.V., & Kumar, K.S., & Suryamukhi, K., & Babu, C.S. (2019). *Identifying malicious dealers in goods and services tax*. pp. 312–316.

Mehta, P., & Mathews, J., & Kumar, S., & Suryamukhi, K., & Babu, C.S. (2019). *Curtailing the tax leakages by nabbing return defaulters in taxation system.* vol. 1127 CCIS, pp. 183–195.

Mehta, P., & Mathews, J., & Kumar, S., & Suryamukhi, K., & Babu, C.S., & Rao, S.V.K.V., & Shivapujimath, V., & Bisht, D. (2019). *Big data analytics for tax administration.* vol. 11709 LNCS, pp. 47-57.

Mehta, P., & Mathews, J., & Kumar, S., & Suryamukhi, K., & Sobhan Babu, C., & Kasi Visweswara Rao, S.V. (2019). *Big data analytics for nabbing fraudulent transactions in taxation system.* vol. 11514 LNCS, pp. 95-109.

Mehta, P., & Mathews, J., & Suryamukhi, K., & Kumar, K.S., & Babu, C.S. (2018). *Predictive modeling for identifying return defaulters in goods and services tax*. pp. 631–637.

Meline, T. (2006). *Selecting studies for systemic review: inclusion and exclusion criteria.* 33, 21–27.

Meservy, R.D., & Denna, E.L., & Hansen, J.V. (1992). *Application of artificial intelligence to accounting, tax, and audit services.* 4(2), 213–218.

Mittal, S., & Reich, O., & Mahajan, A. (2018). *Who is bogus? using one-sided labels to identify fraudulent firms from tax returns*. In: Proceedings of. COMPASS ’18.

Petersen, K., & Feldt, R., & Mujtaba, S., & Mattsson, M. (2008). *Systematic mapping studies in software engineering*. In: 12th EASE. pp. 1–10.

Priya, & Mathews, J., & Kumar, K.S., & Babu, C.S., & Rao, S.V.K.V. (2019). *A collusion set detection in value added tax using benford’s analysis*. vol. 858, pp. 909–921.

Rad, M.S., & Shahbahrami, A. (2015). *High performance implementation of tax fraud detection algorithm.* pp. 6–9.

Rahimikia, E., & Mohammadi, S., & Rahmani, T., & Ghazanfari, M. (2017). *Detecting corporate tax evasion using a hybrid intelligent system: A case study of iran.* 25, pp. 1–17.

Vanhoeyveld, J., & Martens, D., & Peeters, B. (2020). *Value-added tax fraud detection with scalable anomaly detection techniques.*

Vicente, E., & Mateos, A., & Jiménez-Martín, A., & Torra, V., & Narukawa, Y., & Navarro-Arribas, G., & Yañez, C. (2016). *Complicity Functions for Detecting Organized Crime Rings*. vol. 9880, pp. 205–216.

Voorhees,W.R. (2006). *Neural networks and revenue forecasting: a smarter forecast?* 1(4), 379–388.

Wang, G.L. (2012). *Research on sampling method of tax-checking based on neural network.* pp. 1541–1546.

Wu, R.S., & Ou, C.S., & Lin, H.y., & Chang, S.I., & Yen, D.C. (2012). *Using data mining technique to enhance tax evasion detection performance.* 39(10), 8769–8777.

Wu, Y., & Dong, B., & Zheng, Q., & Wei, R., & Wang, Z., & Li, X. (2020). *A novel tax evasion detection framework via fused transaction network representation.* pp. 235–244.

Wu, Y., & Zheng, Q., & Gao, Y., & Dong, B., & Wei, R., & Z., F., & He, H. (2019). *Tedm-pu: A tax evasion detection method based on positive and unlabeled learning*. pp. 1681–1686.

Yu, J., & Qiao, Y., & Shu, N., & Sun, K., & Zhou, S., & Yang, J. (2019). *Neural network based transaction classification system for chinese transaction behavior analysis.* 2019 IEEE BigData Congress. pp. 64–71.

Zha, Z. (2020). *Taxaa: A reliable tax auditor assistant for exploring suspicious transactions.* WWW '20. p. 240–244.

Zhu, X., & Yan, Z., & Ruan, J., & Zheng, Q., & Dong, B. (2018) *Irted-tl: An inter-region tax evasion detection method based on transfer learning*. pp. 1224–1235.

About the Authors

Douglas Silva

Douglas B. Silva is a PhD student in Computer Science at the Federal University of Goias, Goiania-GO, Brazil. His research interests include Data Science, Artificial Intelligence, Computer Systems and E-Government. He currently works at the Public Treasury of the State of Goias, Brazil, analyzing goods and services trading operations data.

Sergio Carvalho

Sergio T. Carvalho is a full professor at the Informatics Institute of the Federal University of Goias in Goiania-GO, Brazil. He received bachelor's degree in Computer Science from the Federal University of Goias, Master and Doctoral degrees in Computer Science, both from the Fluminense Federal University, Brazil. He has experience in the areas of Distributed Systems and Software Engineering and his main areas of expertise are ubiquitous computing, with a focus on healthcare applications, in addition to adaptive distributed systems and software architecture.