Big Data Applied to Tax Evasion Detection

A Systematic Review

Paulo Cesar Abrantes, Felipe Ferraz
CESAR – Recife Center for Advanced Studies and Systems
Recife, Brazil
e-mail: pc.abrantes@gmail.com, fsf@cesar.org.br

Paulo Cesar Abrantes
Indra Company
João Pessoa, Brazil
e-mail: pcoliveira@indracompany.com

***Abstract*— Tax evasion fraud is an issue faced by all governments in the world and one way to improve its detection is the application of big data technologies. This work performs a systematic literature review with the objective of identifying primary studies that address the fraud tax detection by the use of big data. This review resulted in the finding of 56 works of which 5 were identified as primary study. An overview of the results is presented categorized by the studies that address the problem by the use of pattern recognition methodologies, natural language processing and data analytics in auditing. The results also present algorithms and models used in each solution.**

Keywords—tax fraud detection; tax evasion; financial fraud; big data

I. INTRODUCTION

Due to the high tax burden, business and individuals look for ways to pay less taxes and this can be done legally or illegally. The legal way is called tax avoidance, while illegal is known as tax evasion [1].

Governments losses with tax evasion are very high. The estimate of tax losses to fraud in the UK public sector between 2012 and 2013 is 15.4 billions of British Pounds [2], as long as Brazilian government estimates its losses in 415.1 billions of Reais in the year of 2015 [3].

According to the study published by Allingham and Sandmo [4], the probability of practicing tax evasion is directly connected to the probability of its detection, so, governments are constantly looking for more efficient and effective ways to face this problem.

The job of a fiscal auditor involves the analysis of lots amount of data and fiscal reports and this could be made easier by the use of big data technologies.

Although there isn't a consensus despite what is and what isn't big data, most of the authors usually associate it with three characteristics: volume, velocity and variety. Volume is associated to the magnitude of data; variety refers to the use of various types of structured, semi-structured and unstructured data; and velocity refers to the rate at which data are generated and the speed at which it should be analyzed [5].

In summary, as defined by Gartner [6], "Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation".

In the literature is reported the application of big data in a vast variety of areas such as medicine, finances, social media, smart cities, etc. In this study, a Systematic Review was conducted to map how people is taking advantage of this technology to detect and prevent tax evasion, as well as discover yet unknown methods that are being used to commit fraud.

The remainder of the paper is structured as follows: Section 2 presents the applied protocol, describing the objective of the review, research questions, inclusion and exclusion criteria, research strategies and the studies selection process. Sections 3 and 4 is presents the results of the research and some discussions. Section 5 presents conclusion and register for future work.

II. APPLIED PROTOCOL

This review was written based on the guidelines for performing systematic literature reviews in software engineering by Kitchenham et al. [7] in its structure, containing the objectives of the review, the research question, the study selection and quality assessment, the data extraction and synthesis.

A. Objective

The objective of this review is to identify primary studies that focus on the detection of tax evasion by the use of big data technologies. The result of this study can provide the discovery of scientific know-how and challenges in this field.

B. Research Questions

The main research question that guides this review is:

> *How is it possible to use big data technologies to detect tax evasion?*

From this main question, the following secondary questions were developed:

- *Which solution models can be applied to the detection of tax evasion?*
- *How these solutions can be optimized for the use of governments and which benefits it would obtain by the use of these technologies?*
- *What are the challenges of using big data for tax evasion detection?*

978-1-5090-5510-4/16 $31.00 © 2016 IEEE
DOI 10.1109/CSCI.2016.88

- *Which algorithms are being used to detect tax evasion?*

C. Inclusion and Exclusion Criteria

For this review were considered studies that presents solutions and/or discussions about the analysis of big quantities of financial data, aiming to detect tax evasion.

Since this field of research is recent, the results were limited to scientific productions published since 2010, and also the following exclusion criteria:

- Studies not published in the English language;
- Studies that were unavailable online;
- Studies not focused on tax evasion;
- Call for works, prefaces, conference annals, handouts, summaries, panels, interviews and news reports.

D. Research Strategies

To find the studies for this review, there were considered the following databases:

- ACM Digital Library;
- IEEE Xplore;
- ScienceDirect – Elsevier;
- SpringerLink.

The sources were found by using search strings formed by logical combinations of keywords presented as follows:

1. "big data" AND "tax evasion";
2. "big data" AND "fiscal fraud";
3. "big data" AND "tax fraud";
4. "big data" AND "tax sheltering".

This search strings were performed separately on each database between August and October 2016 and the results were grouped by database, as shown in Table 1.

TABLE I. AMOUNT OF STUDIES FOUND ON EACH DATABASE

Database	Number of studies
ACM Digital Library	0
IEEE Xplore	2
ScienceDirect – Elsevier	16
SpringerLink	38

E. Studies Selection Process

Three steps composed the studies selection process: databases search, title analysis and abstract analysis.

At the first step, the search string was executed on the databases and the results were catalogued in an excel sheet. This resulted in the finding of 56 non-duplicated citations.

In the second step, were analyzed all the titles to match with the inclusion and exclusion criteria. In this analysis, there were discarded 27 articles that weren't relevant to this review and the unclear or vague titles were put aside to be analyzed on the next step. At the end of this step, remained 27 for further analysis.

Finally, at the third step, were analyzed the abstract of the remaining articles and those that weren't aligned with the scope of the review were removed. In some articles the conformity with the inclusion and exclusion criteria was unclear in the abstract, so they were included to be analyzed in the quality assessment phase. In the end of this step, there were 15 articles.

Table 2 shows the amount of studies filtered on each step of the selection process.

TABLE II. AMOUNT OF STUDIES FILTERED ON EACH STEP OF THE SELECTION PROCESS

Phase of Selection Process	Number of Studies
1. Databases Search	56
2. Title Analysis	27
3. Abstract Analysis	15

F. Quality Assessment

The quality assessment stage is used to provide more detailed inclusion/exclusion criteria. In this stage, all the selected papers were entirely analyzed and filtered only the ones that contributes to the objective of this review.

After analyzing the 15 remaining studies, 10 were discarded because it wasn't considered relevant to this review and 5 passed to the Data Extraction and Synthesis phase.

Then, the studies were confronted to eight questions, created based on [7] and **Error! Reference source not found.**, and each one received a relevance grade based on its compliance to the questions. The questions were:

1. Does the study examine big data analysis as a way to detect tax evasion?
2. Does the study present a model to detect tax evasion?
3. Are the objectives of the study clearly stated?
4. Is the context of the study adequately described?
5. Were the methods for data gathering correctly used and described?
6. Was the research project adequate to reach the research objectives?
7. Were the research results adequately validated?
8. Does the study contribute to research or to the improvement of tax evasion detection by governments?

The process of quality assessment will be described at the results Section, where will be presented the grade of each primary study.

III. RESULTS

In this section, 5 primary studies were identified [9]-[13]. Four of them presents models for tax evasion detection and

prevention while the other one shows an overview of the tax auditing scenario with some opportunities and challenges in this field.

A. Quantitative Analysis

From the 56 studies returned from the database search, only 5 could be included. They were published between 2014 and 2016 by 16 authors, representing institutions based in 5 different countries (China/Taiwan, United Kingdom, Ireland, Lithuania and USA).

There were identified 18 keywords, which the most relevant with its frequency were: tax evasion (3), big data (2) and Hadoop/MapReduce (1). All the other keywords were used only once.

From the 4 studies that presented big data solutions for the detection of tax evasion, 3 are based on pattern recognition/discovery methods, while 1 is based on Natural Language Processing (NLP). The remaining study presented some trends and discussions about data analytics in auditing.

B. Quality Analysis

In the quality analysis, each primary study was assessed based on the eight questions described at the previous Section. The grade obtained from this analysis measures how much the study is relevant to this review. The classification for each question used a scale of positives and negatives.

Table 3 presents the results of the assessment. Each row represents a primary study and the columns 'Q1'to 'Q8' represent the eight criteria defined by the questions described previously. For each criterion, '1' represent he positive answer and '0' the negative one.

TABLE III. QUALITY ANALYSIS OF PRIMARY STUDIES

Study	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Total
[9]	1	1	1	1	1	1	1	1	8
[10]	1	1	1	1	1	1	1	1	8
[11]	1	1	1	1	1	1	1	1	8
[12]	1	1	1	1	0	1	0	1	6
[13]	1	0	1	1	0	1	0	1	5
Total	5	4	5	5	3	5	3	5	

All the studies analyzed in this step had positive answers for the questions 1, 3, 4, 6 and 8. One study did not present a model for tax evasion detection and two of them neither described methods for data collection (Q5) nor described the validation process.

Three works ([9]-[11]) obtained the maximum score in the quality analysis and the study [13] obtained the maximum number of negative answers (2).

IV. OVERVIEW OF PRIMARY STUDIES

This Section is dedicated to present an overview of the models and discussions addressed on the primary studies. After the analysis and data extraction it was possible to identify the approach used by each study and what kinds of tax evasion method they try to combat.

A. Pattern Discovery/Recognition Models

According to [14], pattern recognition is a machine learning paradigm that aims to develop methods that are able to (partially) imitate the human capacity of perception and intelligence. The algorithms built in this paradigm are commonly designed to train on labeled data to solve the pattern recognition task.

As mentioned before, three papers provided solutions based on pattern discovery/recognition. The solution proposed by Tian et. al [9] describes a graph-based method to discover cases of tax evasion by Interest Affiliated Transactions (IAT).

IATs are transactions between companies that have people with a certain degree of kinship and hold positions with decision-making power, such as Chairman of the board, Chief Executive Officer or Director. There are evidences of tax evasion when in these transactions are offered products/services at much cheaper prices than similar ones offered to other companies.

The method adopts a heterogeneous information network called Taxpayer Interest Interacted Network (TPIIN) to characterize economic behaviors, social relationships and IATs, using a trail-based pattern recognition to select suspicious groups. An example of a TPIIN is illustrated in Fig. 1.

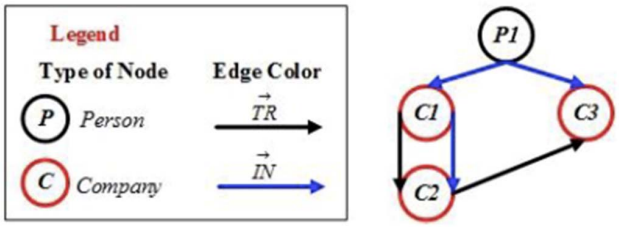

Fig. 1. Example of a TPIIN

The next method, proposed by Cheng, Lu and Hsu [10], uses a visualized data analysis process to detect potential tax evasion caused by bogus business entities.

One of the criteria used to identify bogus business entities was the ratio between the input and output amount of Value Added Tax (VAT) that tends to be higher for bogus entities than for normal enterprises.

It was used supervised and unsupervised learning technics, like Relational Perspective Map (RPM), Newton-Raphson method and K-means, with a tool called Orange Canvas for data collection, data pre-processing, preliminary analysis, clustering analysis and data visualization. The process is illustrated in Fig 2.

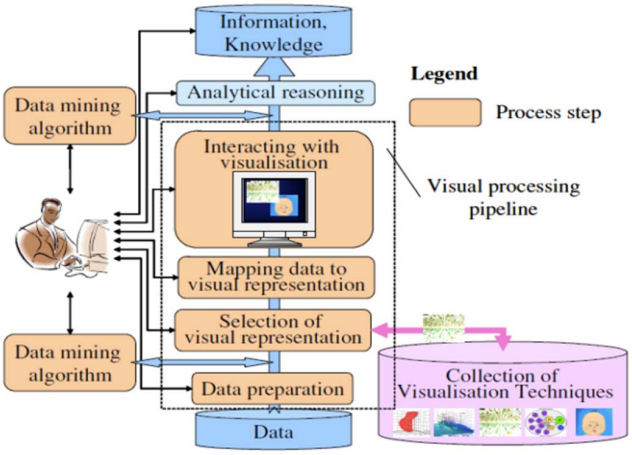

Fig. 2. Visual Data Mining Process

Finally, Stankevicius and Leonas [12] proposes a hybrid approach model for prevention of tax evasion and frauds created based on individual's and his environment's social analysis, business rules, anomaly characteristics and typical fraud monitoring models. The study is theoretical and suggests the integration of different models that are better suitable for different patterns as shown in Table 4.

TABLE IV. HYBRID APPROACH TO FRAUD DETECTION

Pattern type	Model
Associative link patterns	Social Network Analysis Knowledge discovery through associative links
Known patterns	Rule-based learning
Unknown patterns	Anomaly Detection
Complex patterns	Predictive Models

The study also proposes a tax evasion process composed of the following main stages:

1. Exploratory data analysis and transformation;
2. Connection of business rules;
3. Application of advanced analytics like anomaly detection, predictive modeling and social network analysis;
4. Generation of alerts.

The flow diagram of this process is illustrated in Fig. 3.

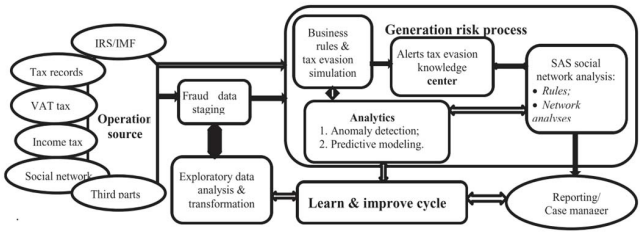

Fig. 3. Tax evasion process flow diagram.

B. Natural Language Processing

Liddy [15] defines NLP as "a range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications".

Banerveld, Le-Khac and Kechadi [11] presents a tool called LES that uses NLP techniques in order to help criminal investigators handle large amounts of textual financial information and improve the evidence finding.

LES tool was built on an Apache Hadoop platform that processes the files containing the financial data of the investigation and makes it available in a web based interface with advanced search capabilities. Fig. 4 shows LES high level design.

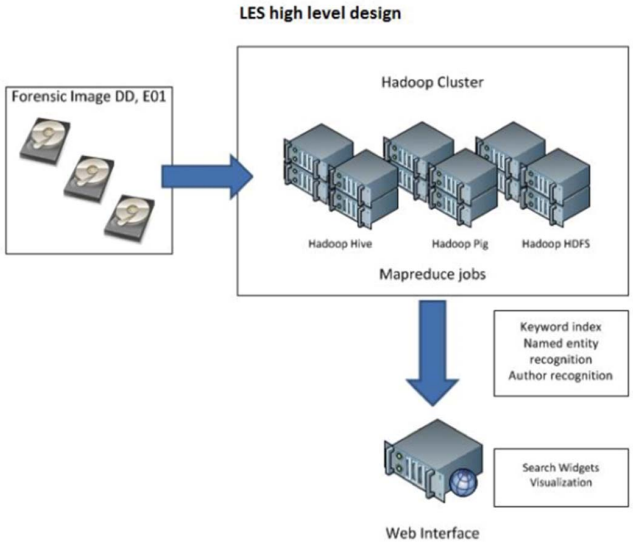

Fig. 4. LES high level design

The process of investigation starts with the evidence files import to the LES system and goes on by running the following MapReduce jobs:

1. Prepare Evidence;
2. Extraction phase;
3. Indexing phase;
4. Named Entity Recognition (NER) extraction phase;
5. Relationship Extraction;
6. Analyzing.

C. Data Analytics in Auditing

Earley [13] presents in her article a background regarding how Data Analytics (DA) and big data can change the public accounting profession, its application in auditing process and some challenges in the area.

The paper points out that taking advantage of DA in public accounting is important based on a survey made in 2014 with public accounting firms where 85% of their managers noted that one of the biggest challenges in the area was how to best analyze the data they have collected. Its explained that the application of DA in this area can enable companies to reduce

tax errors, find tax-saving opportunities and cut administrative costs.

According to the author, there are four primary benefits of the adoption of DA:

- Auditors can test a greater number of transactions than they do now;
- Audit quality can be increased by providing greater insights into client's process;
- Fraud will be easier to detect because auditors can leverage tools and technology that they already use;
- Auditors can provide services and solve problems for their clients that are beyond current capabilities by utilizing external data to inform audits.

The author also makes an analysis on how big data can impact the auditing area. This analysis is shown in Table 5.

TABLE V. TYPES OF BIG DATA AND THEIR IMPACT ON AUDIT APPROACH

Type of Data	Current Practice	Potential Future Practice
Non-financial Data (NFD) or Non-financial Measures (NFM)	Used only marginally on audits, or used with significant auditor judgment required to interpret.	Tools developed to run models or predictive analytics to aid auditors in identifying business risks and areas of focus during planning; aid in fraud detection, and help evaluate and assess going concern.
Financial Data (FD)	Auditors collect and test a sample of transactions and use judgment on those areas that are difficult to test (such as management estimates).	Tools can test 100% of transactions. Will identify anomalies/unexpected patterns in client-provided transaction data. This will guide additional testwork, possibly uncovering fraudulent transactions. Judgment used in assessing next steps after anomalies are uncovered.

The three major challenges listed in the article are the training and expertise of auditors; data availability, relevance and integrity; and expectations of the regulators and financial statement users.

V. CONCLUSION

This paper presented a systematic literature review of big data applied to tax evasion detection. The objective of this review was to identify primary studies that focus on the detection of tax evasion by the use of big data technologies. The database search returned 56 papers which 5 were classified as primary studies after selection and quality criteria were applied. Table 6 summarizes the four studies that proposed a solution for tax evasion detection along with the type of tax evasion that it is applied and the technologies used.

TABLE VI. RESULTS SUMMARIZATION

Study	Tax evasion type	Technologies
[9]	IATs-based tax evasion	Heterogeneous information networks, trail-based pattern recognition
[10]	Bogus Entity tax evasion	Orange Canvas tool, supervised and unsupervised learning techniques
[11]	Financial fraud in general	NLP, Apache Hadoop/MapReduce
[12]	Tax evasion in general	Associative link analysis, rule-based learning, anomaly detection, predictive models

Governments could benefit of taking advantage of big data technologies to face the tax evasion issue by many ways like:

- Increase of the volume of audits;
- Improvements in tax evasion detection;
- Improvements in complex fraud investigations;
- Improvements in tax revenue collections;
- Possibility to analyze all the company/individual data available.

The authors of the papers that proposed a solution model didn't reported the challenges and difficulties of its appliance. Earley [13] points out some challenges like training users to handle the systems and maintaining relevant data available and its integrity.

As a future work is pretended to use the findings obtained from this review to propose a model of tax evasion detection using big data adapted to Brazilian regulations on fiscal area and evaluate it in a revenue agency.

ACKNOWLEDGMENT

To the CESAR.EDU teaching staff that contributed with methodological guidance for the development of this paper. We also appreciate the patience and dedication of our families that unconditionally supported the work that has been done so far.

REFERENCES

[1] J. C. Zanluca, "The price of tax evasion,". [Online]. Available: http://www.portaltributario.com.br/artigos/precodasonegacao.htm. Accessed: Aug. 07, 2016.

[2] University of Portsmouth, "Annual fraud indicator 2016," 2016. [Online]. Available: http://www.port.ac.uk/media/contacts-and-departments/icjs/ccfs/Annual-Fraud-Indicator-2016.pdf. Accessed: Aug. 08, 2016.

[3] SINPROFAZ, "Tax evasion in Brazil – A deviation of the estimate of 2015 exercise collection," 2016. [Online]. Available: http://www.quantocustaobrasil.com.br/artigos-pdf/sinprofaz_indicador_sonegacao-28-06-2016.pdf. Accessed: Aug. 06, 2016.

[4] M. G. Allingham and A. Sandmo. "Income tax evasion: A theoretical analysis", *Journal of Public Economics*, vol. 1, no. 3-4, pp. 323-338, Nov. 1972.

[5] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137-144, Apr. 2015.

[6] Gartner, "What is big data? – Gartner IT glossary – big data," in *All Definitions*, Gartner IT Glossary, 2012. [Online]. Available: http://www.gartner.com/it-glossary/big-data/. Accessed: Oct. 10, 2016.

[7] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering", Technical Report EBSE-2007-01, 2007.

[8] F. Ribeiro, F. Ferraz, M. Torres, and G. Alexandre, "Big data solutions for urban environments a systematic review," ALLDATA 2015, The First International Conference on Big Data, Small Data, Linked Data and Open Data, pp. 22-28, Apr. 2015.

[9] F. Tian *et al.*, "Mining suspicious tax evasion groups in big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2651-2664, Oct. 2016.

[10] H. Cheng, Y.-C. Lu, and C.-C. Hsu, "A visualized data analysis for bogus business entity detection," *2014 IEEE International Conference on Big Data*, pp. 9-15, Oct. 2014.

[11] M. Banerveld, N.-A. Le-Khac, and M.-T. Kechadi, "Performance evaluation of a natural language processing approach applied in white collar crime investigation," in *Future Data and Security Engineering*, T. K. Dang, R. Wagner, E. Neuhold, M. Takizawa, J. Küng, and N. Thoai, Eds. Switzerland: Springer International Publishing, 2014, pp. 29-43.

[12] E. Stankevicius and L. Leonas, "Hybrid approach model for prevention of tax evasion and fraud," *Procedia – Social and Behavioral Sciences*, vol. 213, pp. 383-389, Dec. 2015.

[13] C. E. Earley, "Data analytics in auditing: Opportunities and chalenges," *Business Horizons*, vol. 58, no. 5, pp. 493-500, Sep. 2015.

[14] K. Riesen, *Structural pattern recognition with edit distance: Approximation algorithms and applications*. Switzerland: Springer International Publishing, 2015.

[15] E. D. Liddy, "Natural Language Processing," in *Encyclopedia of Library and Information Science*, 2nd ed. New York: Marcel Deccker, Inc, 2001.