Communication Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 281–287
DOI: 10.15439/2023F5207
ISSN 2300-5963 ACSIS, Vol. 37

Data science to identify crimes against public administration

Luan Bruno Souza
Postgraduate Program in Computer Science (PROCC/UFS) – Aracaju, Brazil
Laboratory of Technological Innovation in Health (LAIS) – Natal, Brazil
Prosecution Office of Sergipe (MPSE) – Aracaju, Brazil
luanbrunos@gmail.com

Methanias Colaço Júnior
Postgraduate Program in Computer Science (PROCC/UFS) – Aracaju, Brazil
Laboratory of Technological Innovation in Health (LAIS) – Natal, Brazil
Advanced Center for Technological Innovation (NAVI) – Natal, Brazil
mjrse@hotmail.com

Rodrigo Silva
Ministry of Health – Brasília, Brazil
Laboratory of Technological Innovation in Health (LAIS) – Natal, Brazil
rodrigo.silva@lais.huol.ufrn.br

Raphael Fontes, Caldeira Silva, Jailton Paiva, Ricardo Valentim, Gabriel Lins
Laboratory of Technological Innovation in Health (LAIS) – Natal, Brazil
Advanced Center for Technological Innovation (NAVI) – Natal, Brazil
raphaelf.ti@gmail.com,{caldeira.silva, jailton.paiva, ricardo.valentim, gabriel.lins}@lais.huol.ufrn.br

***Abstract*—Context: The management of public resources is subject to illegal acts and the automatic identification of such acts depends on the analysis of a lot of data. Objective: The object of this work is the analysis of scientific publications through a study based on systematic mapping with the purpose of evaluating them in relation to the use of automated tools to identify crimes against public administration in databases from the perspective of researchers in the data science context. Method: Using PICO strategy (Population, In- tervention, Comparison, and Outcome), a systematic mapping was conducted to find the primary studies in the literature and collect evidence for directing future research. Results: Nineteen works were found that fit the proposed cri- teria. Almost 80% of the studies found seek to identify some type of fraud in bidding processes, obtaining accuracies between 72% and 99%. The research also revealed different techniques for approaching the problem. Considering all the works, the most used databases are bidding bases, lawsuits, public notices and corporate structure of companies, respectively. Conclusions: The work has shown a recent increase in interest in analyzing public data for irregularities. It is expected that this analysis will help control bodies elucidating different ways of detecting crimes against the public administration in an automated way.**

***Index Terms*—Crime, Corruption, Public Administration, Data Science**

I. Introduction

PUBLIC resource management in many countries, as well as in Brazil, is unfortunately subject to illicit acts, which aim at the subtraction usage of the same resources for the public benefit. Among the most common crimes against public administration, according to Brazilian law, are Corruption, Embezzlement, Prevarication, and Concussion. In the Brazilian context, a study carried out by the Department of Competitiveness and Technology (Decomtec) of Fiesp (Federation of Industries of São Paulo) revealed that the economic and social damage caused by corruption in the country reaches R$ 69 billion reais per year [8]. At the same time, the Anti-Corruption Capacity Index (CCC), which is prepared by the American business entity Americas Society/Council of the Americas (AS/COA) and the British consultancy Control Risks, indicates that, since 2019, Brazil has been falling in the ranking that measures each nation ability to fight corruption [24]. In addition, Brazil ranks 96th in the Corruption Perceptions Index, organized by Transparency International, which order countries' ranks according to the degree to which corruption is perceived to exist among public and political officials, in a total of 180 countries [10].

This difficulty in combating crimes against the public administration involves the difficulty in analyzing a large data volume referring to the public asset movement, often dispersed in different databases. As a result, a good part of the investigative processes about damages to public funds originated in complaints made by the citizens themselves [25]. However, despite the difficulty imposed by the large information volume, it is precise that a good part of government services are stored (and to some extent available) in digital format that makes their analysis through Data Science and Data Analytics usage.

Given this scenario, it is necessary to use and improve techniques and tools which aim to detect, identify or predict the potential crime existence against the public administration. In many cases, these deductions can only be extracted from the unified analysis of distinct databases. The information collected in heterogeneous databases, in order to assist in the decision-making process, is already widely used in the private sector worldwide, for example, in training Credit Score - an index that determines how safe it is to provide credit to a given consumer [11].

The present work objective, therefore, is, through a systematic mapping accomplishment, to carry out a survey on the studies that aim at the development and improvement of crime detection techniques against the public administration in databases, which techniques are most used, which crime

Thematic track: International Workshop on AI in Digital Humanities, Computational Social Sciences and Economics Research

types are most addressed, and which databases are most used. The work also aims to observe the countries with the greatest interest in exploring the problem and whether it is possible to establish a correlation between this interest and the corruption perception indices, according to [10], as well as the interest evolution over time.

The rest of the work is organized as follows: section 2 describes the work methodology, the research questions raised, and the search strategies. Section 3 presents the results obtained after the search, as well as the answers to the research questions. A work narrative synthesis is described in section 4. Section 5 looks at threats to the work's validity. Conclusions and final considerations are presented in section 6.

II. Methodology

The following section describes the methodology used to carry out the work. To guide the research question formulation and the bibliographic search, the PICO strategy was used [23]. The PICO strategy guides the research question construction and the bibliographic search, and allows the researcher, when having doubt or question, to locate, accurately and quickly, the best scientific information available. It presents four fundamental research elements: Population, Intervention, Control, and Outcome, which the authors used to describe all components related to the identified problem and structure the research questions.

A. Research questions

Following are the research questions:

QP1. What crime types against public administration are most commonly identified in these works?

QP2. What are the most widely used data science approaches to detect them?

QP3. What are the approach performance metrics?

QP4. What are the most used databases for the approach application?

QP5. What are the main journals and conferences on the topic?

QP6. In which years were more articles published in this area?

QP7. Which countries have the most publications in this area?.

B. Search Strategy

The research was designed according to the PICO strategy [23], and the result is illustrated in Table 1. Therefore, keywords were established for each category. The resulting set is described in Table 2. The first keywords were selected from some control articles, similar to solution sought in this work. In addition, other keywords were included based on criteria such as related works, similarity and synonyms. The keyword set was then refined, removing redundant words and identifying word stems. The process result is illustrated in Table 3.

Table 4 shows the string used for searches in the databases. The population keywords were subdivided into three blocks, the first being related to the action (detection and its correspondences), the second related to the object sought (crime, corruption, and its correspondences), and the third block related to where to find the objects sought

Table I. PICO Strategy Categories

Category	Description
Population	Publications that directly address the crime identification against the public administration.
Intervention	Context of applications that use automated approaches to identify crimes against public administration.
Control	Applications that do not use automated approaches to identify crimes against the public administration.
Result	Automated approaches to identify crimes against the public administration through computing usage.

Table II. Keywords by category

Category	Description
Population	crime detection against public administration, corruption detection, collusion detection, fraud detection, corruption in public sector, fraud detection in public procurement, risk pattern in public sector, cartel detection, corruption risk assessment, offences against public administration, public ghost employee, public ghost payroll, organized crime, prevarication, public treasury, public procurement, public bidding, government purchasing, bid rigging, public fund, money laundering
Intervention	data mining, data science, text mining, artificial intelligence, a.i, data crossing, crossing technologies, data combination, data manipulation, machine learning, neural network, deep learning, cluster analysis, algorithm
Control	-
Result	decision support system, dss, knowledge discovery, automated system, automated information system, prototype, online analytical processing, olap, intelligent agent, corruption indicator, predictive, model, predictive analytics, model

Table III. Keywords Refined by Category

Population	**crime detect*, collusion detect* corruption detect*, fraud detect* offences detect*, cartel detect* prevarication detect*, ghost payroll detect*, ghost employee detect*, bid rigging, money laundering, corruption risk, public administration, public sector, public procurement public treasury, public bidding, public employ*, government* purchas* government* treasury, public fund, risk pattern**
Intervention	data mining, data science, text mining, data crossing, artificial intelligence, crossing technologies, data combination, data manipulation, machine learning, neural network, deep learning, cluster analysis, algorithm
Result	decision support system, dss, knowledge discovery, automated system, prototype, automated information system, online analytical processing, olap, intelligent agent, corruption indicator, approach, predictive model, predictive analytics, model

TABLE IV. GENERIC SEARCH STRING

(“detect*” OR “search*” OR “find*” OR “look* for” OR “predict*”) **AND (“crime” OR “corruption” OR “clue” OR “fraud*” OR “collusion” OR “offence” OR “cartel” OR “malfeasance” OR “prevarication” OR “ghost payroll” OR “ghost employee” OR “bid rigging” OR “irregularity” OR ”money laundering”OR “anomaly” OR “suspicious”)** **AND (“public administration” OR “public sector” OR “public procurement” OR “government* procurement” OR “public treasury” OR (“bidding” AND (“public” OR “government*”)) OR “public employ*”OR “government* purchas*” OR “government* treasury” OR “public fund”)**	Population
AND (“data mining” OR “data science” OR “text mining” OR “artificial intelligence” OR “data crossing” OR “crossing technologies” OR “data combination” OR “data manipulation” OR “machine learning” OR “neural network” OR “deep learning” OR “cluster analysis” OR “algorithm”)	Intervention
AND (“decision support system” OR “dss” OR “knowledge discovery” OR “automated system” OR “automated information system” OR “prototype” OR “online analytical processing” OR “olap” OR “intelligent agent” OR “predictive model*” OR “predictive analytics” OR “model” OR “corruption indicator” OR “approach*”)	Result

(public sector, bids, and their correspondence). Searches in titles, abstracts, and keywords were used in the Scopus, IEEE Xplore Digital Library, Web of Science, and ACM Digital Library search bases.

Following are the Inclusion Criteria: (1) Short and complete works published and available in full in scientific databases, with title, abstract, and keywords available in the English language; (2) Recent works (published from 2010 onwards), however, they have already been approved by the scientific community. (3) Works that propose a method, tool, or application for the detection, selection, or fraud or crime prediction against public administration in databases through Data Science usage. The 2010 limit year was determined to be immediately prior to the Law implementation on Information Access [1], which regulated the citizens' constitutional right to access public information.

The following are Exclusion Criteria: (1) Duplicate works; (2) Restricted works; (3) Revision works; (4) Works that do not seek to detect crimes; (5) Works that seek to detect or predict other crimes outside the context of this work.

C. Information Extraction Strategy

To assess the work quality and answer the exposed research questions in section 2.1, a form was designed to be answered for each article read completely. According to [12], data extraction forms should be designed to collect all the information necessary to address the issues and quality criteria of the study. Table 5 presents the extraction form used in this research. For the attributes Crime Types, Approaches, Performance Metrics, and Databases, the results are multivalued, that is, there is the possibility of more than one answer of the same attribute for each article.

TABLE V. EXTRACTION FORM

Attribute	Description
Crime Type	Identification of the crime type against the public administration which the work aims to identify. Part of this task was already carried out in exclusion criterion 5, which sought to remove crime identification work outside the public administration context.
Approach	The Data Science identification approach used in the crime identification
Performance Metric	The evaluation criteria identification of the approach according to the authors' experiment, if there is any.
Database	The databases identification, structured or not, analyzed by the approaches.

III. RESULTS AND DISCUSSION

The following subsections describe the search process and discuss the results. In subsection 3.1 the resulting treatment and the exclusion criteria application until the analysis base formation is described. Subsection 3.2 runs briefly over each selected job. From subsection 3.3 onwards, the research questions are answered based on the results.

A. Results

Once the works were searched in the specified databases using the keywords, the first step was the duplicate work removal since they were found in more than one database. Figure 1 presents a flow describing the article extraction process from this phase to analysis. The search sum in the databases returned a total of 251 works, a number that was reduced to 223 after the duplicate article removal.

Then the other exclusion criteria were applied. Two articles were removed for being of a restricted domain. Afterward, the work title was read to identify review articles. Along with the title, the work abstract was also observed, which allowed us to remove those that were not intended to detect fraud and crimes. These three criteria allowed us to reduce the number of works to a total of 100 articles.

After the exclusion according to these criteria, we were left with 100 works that aimed to search for techniques and tools to detect crimes or fraud automatically. Yet, many of these works did not aim at crime identification in the public administration sphere. Among the events sought by these articles were common crimes, hacking invasions, health insurance fraud, and even illegal immigration. Frequently reading the title and abstract were sufficient for this discernment, but often the article introduction needed to be read for more precision. Finally, after the last step in applying the removal criteria, we reached 19 articles. All have been read completely and a brief commentary is described in the following sections.

B. Work Abstracts

The works of [14] and [20] present an approach to crime detection based on users' perceptions. The first is based on the post content on the social network Twitter, while the authors of the second created a survey to be applied by public

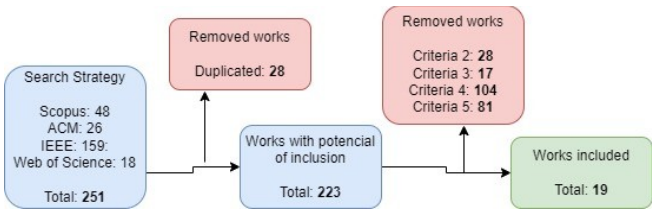

Figure 1. Prism chart with data extraction

service users. The article by [13] seeks to analyze financial transactions in order to find suspicious transactions that lead to money laundering, while [3] propose an ontology to, applied to a data warehouse, identify inconsistencies in payroll.

From here, the articles focus on fraud detection in the public purchase sphere. [7], [16], and [21] seek to identify potential signs of fraud already in terms of the bids using opening, among other devices, text mining techniques. The work of [5] proposes to use Bayesian networks to identify fractional purchases, where the bidding process is suppressed if each purchase value does not exceed a maximum value defined by Brazilian legislation.

The article by [17] added, to the bidding database, a bid list against those for which there are legal proceedings or formal complaints. The objective is to detect patterns of attributes of problematic processes in order to identify problems in new bidding processes using random forest. [9] used a similar approach, in addition to using other available process data, such as budget, duration, delays, time before electoral processes, and geographic patterns.

Other works seek to detect bidding processes with potential collusion through the association network analysis of other purchases involving the same buyers or suppliers. They are [22], [6], [19] and [4]. For this, they use techniques such as association rules and random forest. Articles such as [2], [18], and [26] use clustering algorithms to group competitive and non-competitive bidding processes based on data such as the ratio between the bid values offered by companies and initial value of the contract.

Finally, [25] and [15] propose the veracious data analysis suite creation of bidding processes, precisely with the addition of information available in other databases. It allows the fraudulent schemes detection that could only be elucidated from this distributed information joining. Auxiliary databases include corporate structure data of companies, income transfer programs, and electoral data

C. QP1: What crime types against public administration are most commonly identified in these works?

The vast majority (78.9%) of the work authors focused their efforts on automated techniques to detect fraud in bids, as illustrated in Figure 2. However, the works differ on when the detection attempt is performed. Some works, such as [7], [16] and [21], seek to identify potential fraud signs in terms of the bid opening. Other works, such as [2] and [18], use variables found during the bidding process to find collusions, such as bid values and time intervals. Finally, works such as [22] and [4] based on the compilation of different bidding processes already carried out in search of participation patterns and winners. There are still other works, such as [25] and [15], which use multiple approaches to detection.

Two other works ([14] and [20]) did not define a specific crime type but were concerned with detecting fraud in a more comprehensive way through opinion collection and user perception. There are also works aimed at finding fraud in the government employee payroll [3] and money laundering [13].

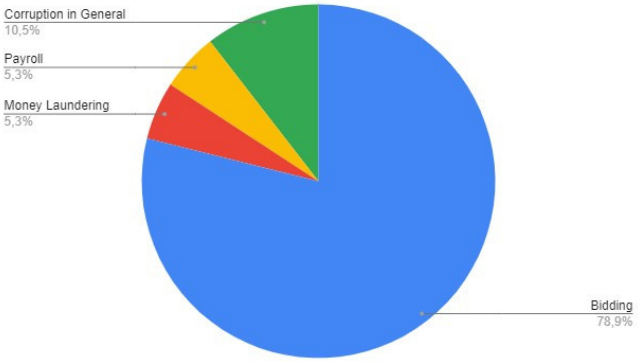

Figure 2. Crimes or Frauds identified in the approaches

D. QP2: What are the most widely used data science approaches to detect them?

As seen in the previous item, most works focus on fraud detection in government purchases through bidding processes. Some works, such as [7], [16] and [21], seek to identify signs already in terms of opening the process. For this, text mining tools are used to analyze specific term elements. Once found, they apply logistic regression or deep learning algorithms to detect a competitiveness lack in bidding terms, which could point to a possible collusion between the bidder and interested companies. On the other hand, works aimed at identifying fraud in the same processes using data generated during the bidding process, such as [2], [26] and [18], using data as the ratio between bid value and initial bid value, through clustering algorithms to differentiate competitive and non-competitive processes. Finally, clustering algorithms, as well as association rules used by works such as [22] and [4] to identify collusions between companies and suppliers through several bidding process analysis. Other works, such as [25] and [15], combine other techniques for this detection, in addition to the assigning score possibility to certain companies that participate in bidding processes. For this, they use other data sources in addition to the basis of contracts and public bids generally used in other approaches, as detailed in the following section.

The works [14] and [20], which did not define a specific crime type because they are concerned with detecting fraud in a more comprehensive way, making use of reports and impressions of public service users. While [14] use machine learning techniques to detect fraud evidence in public services through posts on Twitter, [20] applied forms to users of different services in order to search for inefficiency signs based on the responses to these forms using clustering algorithms.

To detect money laundering crimes, [13] used a Bayesian classifier based on a bank operation set. As for looking for inconsistencies in payrolls (not necessarily fraud) [3] de-

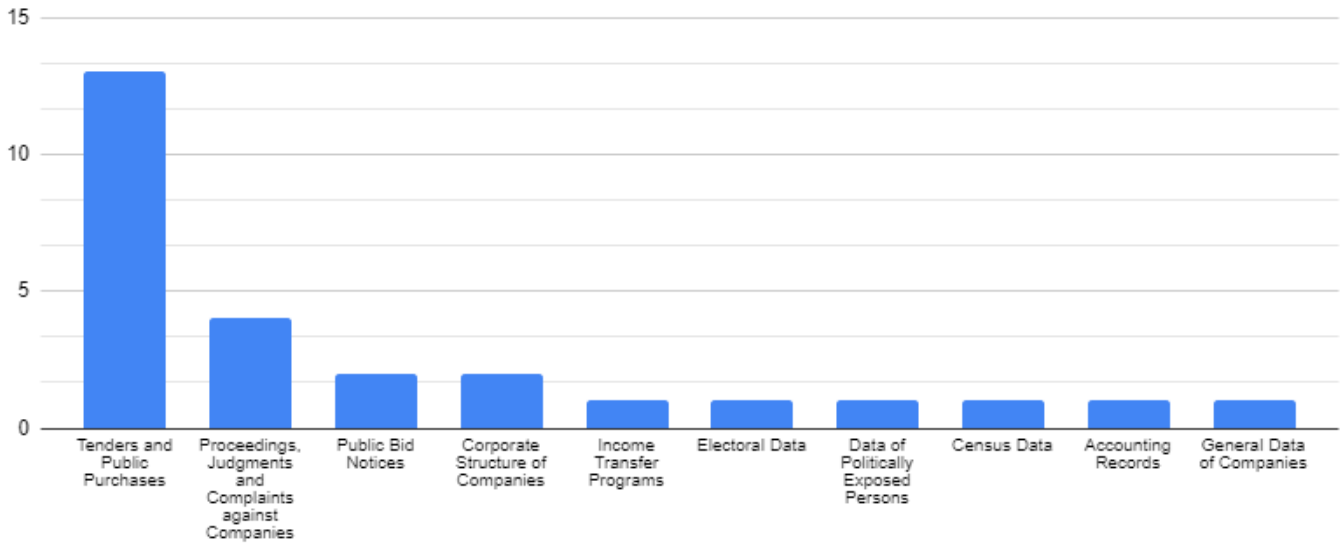

Figure 3. Databases used to detect bid fraud

fined an ontology indexing process through concept maps and audit indicators as a tool for documenting evidence.

E. QP3: What are the approach performance metrics?

In general, the authors used accuracy as the predominant form of statistical evaluation of the proposed models, with the exception of [22] who obtained an assertiveness of 90% according to its own evaluation index, called RQ, when trying to identify cartel formation and [7] who obtained a Mean Square Error (MSE) of approximately 0.0013 when trying to predict fraudulent bids from their opening terms.

Also, when analyzing the bid opening terms [16] obtained an accuracy of 76% using SVM, while [21] obtained accuracies between 72% and 85%, depending on the product group of the used bidding process utilizing Logistical Regression and Bayesian Networks. [6] obtained an accuracy of 67% in identifying cartels. Through the attribute analysis of the bidding process, [26] reached an accuracy of 99%, while in the work of [17] the same rate was 90% using similar data, including data from known problematic bids. The work of [5] reached an accuracy of 99.9% analyzing fractional bids where the global value is divided into bids with lower values to circumvent some legal requirements.

Outside the bidding process context, [20] obtained an accuracy of 87.5% in the irregularity discovery when applying a questionnaire to public service users. [13] reached an accuracy of 81% when searching for suspicious financial transactions in order to find money laundering evidence. The other works found proposed data analysis models without presenting statistical validations regarding these models' assertiveness.

F. QP4: What are the most used databases for the approach application?

The answer to this question must take into account the fraud or crime type that the work aims to detect. [13], for example, used bank transaction databases to look for fraud evidence. [3] used a payroll database to build a data warehouse and define its ontology. In turn, [14] and [20] used posts on the social network Twitter and data from an applied survey, respectively, to identify fraud in the public sector through the perception of users.

Figure 3 counts the databases used to help detect fraud in bidding processes. Note that one approach can make use of more than one database simultaneously. Altogether, 13 of the 15 studies found that proposed to detect anomalies in bidding processes utilizing public bidding and procurement bases, while the other two analyzed only opening documents and the process definition. In order to negatively consider processes involving companies against which there was a history of lawsuits, some works made use of procedural bases, judicial sentences, and complaints. Other databases used were those that included the corporate structure of companies, income transfer programs, electoral data, data on politically exposed persons, census data, accounting records, and company registration data.

G. QP5: What are the main journals and conferences on the topic?

Among the results found, all of them were published in different Magazines, Journals, or Conferences. Thus there isn't a periodical or conference that stood out from the others.

H. In which years were more articles published in this area?

As shown in Figure 4, it is possible to notice an increase in the publication of works that address the researched topic from 2019, with four papers published. The year 2020 was,

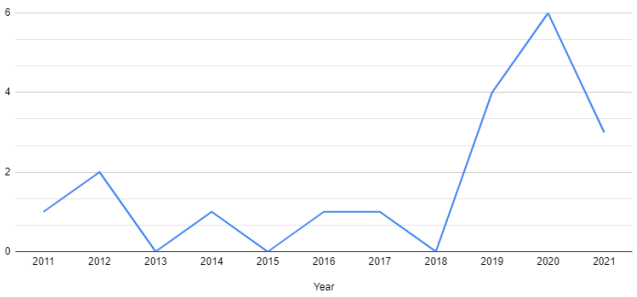

Figure 4. Databases used to detect bid fraud

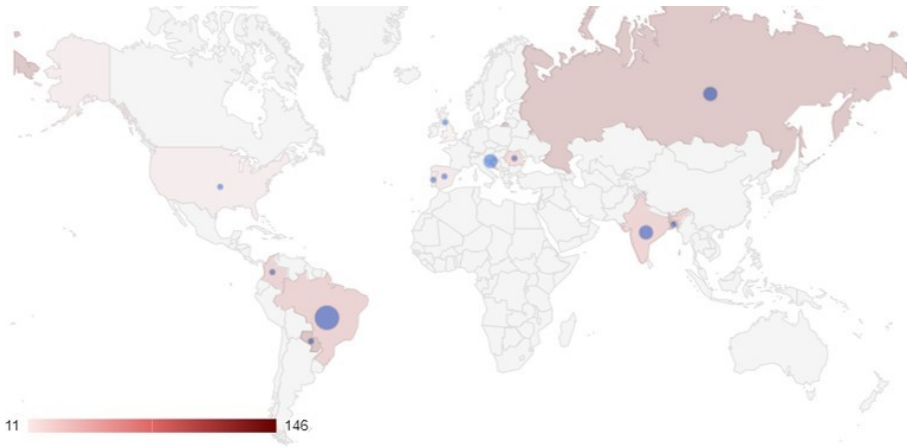

Figure 5. Publication number by country and position in the IPC-2020

until then, the one with the highest number of publications, accounting for six papers.

I. QP7: Which countries have the most publications in this area?

Figure 5 shows the distribution of published works around the world, where the size of the blue circle represents the publication number. It is possible to notice that the country with the largest publication number is Brazil (5), followed by Russia (2), India (2), and Croatia (2). Colombia, Spain, Bangladesh, the United Kingdom, the United States, Romania, Paraguay, and Portugal complete the list with one work each. Figure 5 also plots, in red, the position of that country (only where works was found) in the Corruption Perceptions Index prepared by International Transparency for the year 2020 [10], where the darker the red color, the greater the corruption perception. Through the results, it was not possible to establish a relationship between the number of published works and the corruption perception in the country.

IV. Narrative Synthesis

Quantitatively, the result observation allowed us to observe that the search for automated ways to detect fraud and crimes is relatively recent in the scientific context. For many authors, this is often due to the late digitization process of governments in relation to the private sector, especially in underdeveloped countries. With no government data available in digital format, there is no means to perform such a task.

The quantitative analysis also placed Brazil as a major contributor to this approach type, despite being in an intermediate position in the Corruption Perception Index (CPI) in 2020 provided by International Transparency. It was observed that the publication number per country cannot be directly related to the countries' perceptions of corruption, according to the same index.

The bidding processes, as analyzed, are the main target of automated fraud detection processes in the public context. In general, the authors justified the interest in this collusion type due to the large financial volumes involved in government purchases that carry out the bidding processes. Furthermore, the amount of money involved in these transactions inevitably ends up arousing malicious people's interest.

For such detections, the works take turns using predictive and deductive models. Deductive models are generally based on local legislation and prior knowledge about the fraudulent scheme typologies, which is often a disadvantage because this approach type is not able to predict new scheme formats. On the other hand, predictive models are more difficult to apply due to the absence of training bases, considering that the number of proven frauds is often insufficient for modeling this approach type.

V. Threats to Validity

The great difficulty of the current work concerns the keyword selection to search in the databases. As much as the search context is well defined, the expressions used to describe crimes, frauds, or anomalies are diverse and are subject to different regionalities and descriptions depending on the country where the laws are written making it difficult to select terms used as population keywords, according to the PICO model. This characteristic threatens above all the excessive volume of works from Brazil. Another similar difficulty is the wide term variety used to describe the methods used for detection, described in the intervention keywords. An incomplete keyword selection can considerably limit the number of results returned.

As for the exclusion criteria, the heterogeneity of different laws and policies in different countries can compromise the researcher's interpretation, regarding the often subjective analysis of these criteria. For example, in the current work, fraud against health plans was not considered, given that Brazil has a single public health system that is not very intertwined with the private system so financial fraud against health plans in Brazil generally does not involve public administration. But it is not possible to infer that this does not occur in other countries.

VI. Final Consideration

All over the world, to a greater or lesser extent, public money management deprives the population of the right to fully take advantage of the resources provided by them through taxes. This mismanagement is often intentional, resulting from criminal actions that seek to subtract or misuse public goods for their own benefit. Fortunately, the recent governmental service digitization, allied to the principle that

part of this information load is in the collective domain, allows the organizational or popular initiative emergence, aimed at these illicit act identification. The information sheer volume, however, requires an automated process.

The current work described a systematic mapping with the objective of elucidating scientific works aimed at the automated tools development or improvement for the fraud detection or crimes against public administration. The research questions were raised and, based on the PICO strategy, the search keywords were selected. Once searched, the works were selected based on pre-defined inclusion and exclusion criteria.

The result analysis shows that this concern, fortunately, is growing. Several studies were found with this objective in mind, and they do so by approaching different strategies. Due to the financial resource volume involved, bidding fraud is the main target of this initiative type. Some works even look for the association of different databases, seeking the fragmented information discovery. The computational resources for this range from text mining to machine learning algorithms.

It is hoped that this work can provide guidance to entities that seek to develop initiatives and develop tools that allow a better public expenditure monitoring. As noted in this work, part of the information available for this task is in the public domain, allowing non-governmental entities to participate directly in these initiatives. However, it is the control bodies that have exclusive control over part of the data identified as a source for the detecting crime work, in addition to having civil liability for such.

It is recommended the existence of periodical works in this sense, in order to maintain the population and control institutions always updated on the best practices to achieve the final objective, which is the fight against fraud in public administration. In future works, it is recommended a better understanding of the terms used to define illegal or suspicious acts which will be used in the search string, in order to avoid the existence of false negatives in the process. In addition, a more in-depth analysis of the results offered by the applications found is also recommended, comparing them and indicating the best approach for each situation.

REFERENCES

[1] Brasil (2011). Lei nº 12.527, de 18 de novembro de 2011. Dia´rio Oficial da Repu´blica Federa- tiva do Brasil.

[2] Busu, M. and Busu, C. (2021). Detecting bid-rigging in public procurement. a cluster analysis approach. Administrative Sciences, 11(1):13.

[3] Campos, S. R., Fernandes, A. A., De Souza, R. T., De Freitas, E. P., da Costa, J. P. C. L., Serrano, A. M. R., and Rodrigues, D. d. C. (2012). Ontologic audit trails mapping for detection of irregularities in payrolls. In 2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN), pages 339-344. IEEE>

[4] Carneiro, D., Veloso, P., Ventura, A., Palumbo, G., and Costa, J. (2020). Network analysis for fraud detection in portuguese public procurement. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 390–401. Springer.

[5] Carvalho, R. N., Sales, L., Da Rocha, H. A., and Mendes, G. L. (2014). Using bayesian networks to identify and prevent split purchases in brazil. In *BMA@ UAI*, pages 70–78.

[6] Domashova, J. and Kripak, E. (2021). Application of machine learning methods for risk analysis of unfavorable outcome of government procurement procedure in building and grounds maintenance domain. *Procedia Computer Science*, 190:171–177.

[7] Domingos, S. L., Carvalho, R. N., Carvalho, R. S., and Ramos, G. N. (2016). Iden- tifying it purchases anomalies in the brazilian government procurement system using deep learning. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 722–727. IEEE.

[8] FIESP (2010). Relato´rio corrupc¸a˜o: custos econoˆmicos e propostas de combate. In *DECOM- TEC*. FIESP - Federac¸a˜o das Indu´strias do Estado de Sa˜o Paulo.

[9] Gallego, J., Rivero, G., and Mart´ınez, J. (2021). Preventing rather than punishing: An early warning model of malfeasance in public procurement. *International journal of forecasting*, 37(1):360–377.

[10] International, T. (2021). Corruption perceptions index.

[11] Investopedia (2021). Credit score. https://www.investopedia.com/terms/ c/credit_score.asp. Accessed: 2021-10-16.

[12] Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

[13] Kumar, A., Das, S., and Tyagi, V. (2020). Anti money laundering detection using na¨ıve bayes classifier. In *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)*, pages 568–572. IEEE.

[14] Li, J., Chen, W.-H., Xu, Q., Shah, N., and Mackey, T. (2019). Leveraging big data to identify corruption as an sdg goal 16 humanitarian technology. In *2019 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 1–4. IEEE.

[15] Mart´ınez-Plumed, F., Casamayor, J. C., Ferri, C., Go´mez, J. A., and Vidal, E. V. (2018). Saler: a data science solution to detect and prevent corruption in public administration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 103–117. Springer.

[16] Modrusan, N., Rabuzin, K., and Mrsic, L. (2020). Improving public sector efficiency using advanced text mining in the procurement process. In *DATA*, pages 200–206.

[17] Niessen, M. E. K., Paciello, J. M., and Fernandez, J. I. P. (2020). Anomaly detection in public procurements using the open contracting data standard. In *2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG)*, pages 127–134. IEEE.

[18] Padhi, S. S. and Mohapatra, P. K. (2011). Detection of collusion in government procurement auctions. *Journal of Purchasing and Supply Management*, 17(4):207–221.

[19] Popa, M. (2019). Uncovering the structure of public procurement transactions. *Business and Politics*, 21(3):351–384.

[20] Pramanik, A., Sarker, A., Islam, Z., and Hashem, M. (2020). Public sector corruption analysis with modified k-means algorithm using perception data. In *2020 11th International Conference on Electrical and Computer Engineering (ICECE)*, pages 198–201. IEEE.

[21] Rabuzin, K. and Modrusan, N. (2019). Prediction of public procurement cor- ruption indices using machine learning methods. In *KMIS*, pages 333–340.

[22] Ralha, C. G. and Silva, C. V. S. (2012). A multi-agent data mining system for cartel detection in brazilian government procurement. *Expert Systems with Applications*, 39(14):11642–11656.

[23] Santos, C. M. d. C., Pimenta, C. A. d. M., and Nobre, M. R. C. (2007). A estrate´gia pico para a construc¸a˜o da pergunta de pesquisa e busca de evideˆncias. *Revista Latino-Americana de Enfermagem*, 15:508–511.

[24] Simon, R. and Aalbers, G. (2019). The capacity to combat corruption (ccc) index. *AS/COA*, page 15.

[25] Velasco, R. B., Carpanese, I., Interian, R., Paulo Neto, O. C., and Ribeiro, C. C. (2021). A decision support system for fraud detection in public procurement. *International Transactions in Operational Research*, 28(1):27–47.

[26] V.I., D., N.V., M., and M.I., B. (2017). Adaptation of cluster analysis methods in respect to vector space of social network analysis indicators for revealing suspicious government contracts. In *2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pages 57–62.