INTERNATIONAL TRANSACTIONS IN OPERATIONAL RESEARCH

WILEY

INTERNATIONAL TRANSACTIONS IN OPERATIONAL RESEARCH

Intl. Trans. in Op. Res. 28 (2021) 27–47
DOI: 10.1111/itor.12811

A decision support system for fraud detection in public procurement

Rafael B. Velasco[a], Igor Carpanese[a], Ruben Interian[b,*], Octávio C. G. Paulo Neto[c] and Celso C. Ribeiro[b]

[a]*Consultant, Praia do Flamengo 66B, Rio de Janeiro, RJ, 22210-030, Brazil*
[b]*Institute of Computing, Universidade Federal Fluminense, Niterói, RJ 24210-346, Brazil*
[c]*GAECO – Paraíba Prosecutor's Office and Federal Prosecutor's Office, Brasília, DF 70070-925, Brazil*
E-mail: rafaelbraemvelasco@gmail.com [Velasco]; carpanese@protonmail.com [Carpanese]; rinterian@id.uff.br [Interian]; octpn1@gmail.com [Paulo Neto]; celso@ic.uff.br [Ribeiro]

Received 18 February 2020; received in revised form 5 May 2020; accepted 5 May 2020

Abstract

Over the past few years, investigators in Brazil have been uncovering numerous corruption and money laundering schemes at all levels of government and in the country's largest corporations. It is estimated that between 2% and 5% of the global GDP is lost annually because of such practices, not only directly impacting public services and private sector development but also strengthening organized crime. However, most law enforcement agencies do not have the capability to carry out systematic corruption risk assessment leveraging on the availability of data related to public procurement. The currently prevailing approach employed by Brazilian law enforcement agencies to detect companies involved in potential cases of fraud consists in receiving circumstantial evidence or complaints from whistleblowers. As a result, a large number of companies involved in fraud remain undetected and unprosecuted. The decision support system (DSS) described in this work addresses these existing limitations by providing a tool for systematic analysis of public procurement. It allows the law enforcement agencies to establish priorities concerning the companies to be investigated. This DSS incorporates data mining algorithms for quantifying dozens of corruption risk patterns for all public contractors inside a specific jurisdiction, leading to improvements in the quality of public spending and to the identification of more cases of fraud. These algorithms combine operations research tools such as graph theory, clusterization, and regression analysis with advanced data science methods to allow the identification of the main risk patterns, such as collusion between bidders, conflicts of interest (e.g., a politician who owns a company contracted by the same government body where he or she was elected), and companies owned by a potentially straw person used for disguising its real owner (e.g., beneficiaries of cash conditional transfer programs). The DSS has already led to a detailed analysis of large public procurement datasets, which add up to more than 50 billion dollars. Moreover, the DSS provided strategic inputs to investigations conducted by federal and state agencies.

*Corresponding author.

Keywords: decision support system; data mining; data science; fraud detection; risk patterns; public procurement; corruption risk assessment; COVID-19

1. Introduction

Over the past few years, investigators in Brazil have uncovered numerous corruption and money laundering schemes at all levels of government and in the country's largest corporations. It is estimated that between 2% and 5% of the global GDP is lost annually because of such practices (Kar and Spanjers, 2014; Weeks-Brown, 2018), impacting directly public services and private sector development, but also strengthening organized crime.

The current approach employed by most law enforcement agencies in Brazil for detecting companies involved in potential cases of fraud consists of relying on circumstantial evidence provided by whistleblowers or specific and unsystematic complaints delivered by citizens. Ultimately, it means that detecting corruption, despite a great effort from investigators, still relies on chance. Consequently, there are thousands of unpunished and nonprosecuted cases of fraud across the country. Most state-level and federal law enforcement agencies do not have the capability to carry out systematic corruption risk assessments of all public contractors in their jurisdictions.

A reasonable mechanism for narrowing down the pool of public contractors that should be investigated is by automatically providing agencies with actionable information about the specific risk patterns found for each company. In our approach, we compute dozens of corruption risk patterns for all public contractors inside a specific jurisdiction. They allow law enforcement agencies to establish priorities concerning the companies to be investigated.

Other authors have already studied how to reduce fraud and corruption cases by identifying inefficiency in public procurement and corruption risk patterns. A basic introduction to the challenge of overcoming corruption in the field of public procurement is presented by Transparency International (2006). Dávid-Barrett and Fazekas (2020) observed the heterogeneous effects of anticorruption reforms in more than 100 developing countries, showing that technical interventions might not represent the best way to tackle systemic corruption. They also showed how data analytics can be used to monitor public procurement at the system level to inform more adaptive and effective anticorruption patterns.

Broms et al. (2019) studied how low political competition is associated with more restricted public procurement processes. They showed that when one party dominates local politics, noncompetitive outcomes from public procurement processes are more common. Fazekas and King (2019) explored the impact of development funds on corruption. They considered high-level corruption in the Czech Republic and Hungary in 2009–2012 using data from over 100,000 public procurement contracts. They showed that European Union funds increased the corruption risk by up to 34%. These negative effects are largely attributable to overly formalistic compliance requirements, instead of a systematic analysis of the use of these funds. A composite contract-level and organization-level corruption indicator was proposed by Fazekas et al. (2014, 2016) by identifying "red flags" in the public procurement process and linking them to restricted competition and recurrent contracts awarded to the same companies. They also showed that companies with higher corruption risk scores had a relatively higher profitability, a higher ratio of contract value to initial estimated price,

and a greater likelihood of politicians managing or owning them, with respect to companies with lower scores for the proposed indicator.

Fraud leaves traces that can be systematically identified using data science tools. According to a study published by KPMG (2009), the predominant cause for the perpetration of fraudulent acts is the insufficiency of internal control systems. There is a consensus among organizations that improving internal controls is essential to fraud prevention. The decision support system (DSS) described in this work is an attempt to solve this problem. It provides the systematic analysis of public procurement that allows law enforcement agencies to establish objective criteria to decide which physical persons or companies will be investigated in a specific jurisdiction, contrary to the prevalent approach for selecting cases based on specific complaints (such as whistleblowing and anecdotal evidence).

Implementing data mining algorithms that allow quantifying dozens of corruption risk patterns for all public contractors inside a specific jurisdiction will result in improvements in the quality of public spending and in the identification of more cases of fraud. It is important to note that many patterns do not represent by themselves individually a significant risk of the existence of fraud or corruption. However, the risk of fraud increases when combinations of these risk patterns are identified. A more detailed discussion about the legal decisions that allow a consistent set of risk patterns to be used as evidence in investigations in Brazil can be found in Santos and de Souza (2016).

For this purpose, we take advantage of existing and previously unused public datasets to develop data mining algorithms that automatically extract, for each public contractor, dozens of corruption risk patterns. Before the development and implementation of the algorithms, we developed extensive ETL (extract, transform, and load) work. In this sense, dozens of open datasets have been captured and cleansed. These datasets include transactional data related to more than 2 million contracts executed in the following Brazilian states: Acre, Amapá, Bahia, Ceará, Distrito Federal, Espírito Santo, Mato Grosso do Sul, Minas Gerais, Paraiba, and São Paulo, as well as in the Federal government. These databases account for more than 50 billion dollars in contract expenditures (the Brazilian Real was quoted at US$ 0.3023 on January 1, 2018). We also created a data lake composed of microlevel data about all Brazilian companies, campaign donations, public servants, and beneficiaries of cash conditional transfer programs, among other information. The algorithms allow the detailed identification of conflicts of interest, when a politician owns a company contracted by the same government body where the politician was elected; companies owned by straw persons, such as beneficiaries of cash conditional transfer programs; bidders with common partners; identical and concerted bids; unjustified cost overruns; companies with the same publicly declared address participating in the same bidding process; companies whose registration date is very close to the date of its first contract or bid with the public sector; and top losers, companies that participate in a large number of bidding processes, but never (or rarely) wins.

The methodology developed in this work and the associated DSS are capable of improving public spending quality by identifying fraudulent cases in databases of bidding processes and public expenditures.

This article is organized as follows. Section 2 introduces the data sources used for corruption disruption and shows how they can be effectively accessed and used. It also describes procedures used in the process of creation of the data lake used as input by the DSS. The section concludes with the general scheme of the analytical approach. Section 3 describes the main fraud risk patterns that

the DSS is capable of detecting at its current stage, which can be further expanded in the future. Some computational algorithms developed to perform automatic or semiautomatic analysis of large quantities of public spending data to extract collusion risk patterns are described in Section 4. Quantitative and operational results obtained with the practical experience of using the DSS are reported in Section 5. Concluding remarks, together with the discussion of some future extensions, are drawn in the last section. A short glossary is presented at the end, with the English description of some specific terms used in Portuguese along the text.

2. Public spending data and data sources

Broadly considering, the main public expenditure datasets in Brazil come from:

- relational databases containing details of all phases of the public expenditure cycle;
- electronic invoices containing disaggregated product-level data (e.g., unit of measure, specific product, product quantity); and
- banking records containing specific details of each transaction. A relevant share of these data is centralized, since the vast majority of government bodies in Brazil use the state-owned bank *Banco do Brasil* for processing payments.

These public expenditure datasets are decentralized, locally managed, and not monitored in real time. The federal and state governments are not able to use this vast amount of public spending data, let alone to develop algorithms and DSSs. Among the reasons, there is almost no collaboration with academia and little competitive advantage at public agencies to attract data scientists.

In the state of São Paulo alone, more than 48,000 companies bid in procurement procedures between 2016 and 2018. Law enforcement agencies are not capable of analyzing more than a few dozen public contractors. The goal of this work is to use relational databases containing public expenditure data in order to provide law enforcement with the systems and analytical tools to generate corruption risk reports in seconds for any company. Therefore, a task that would require thousands of working hours by the traditional approach will be fully automated and considerably optimized.

2.1. Data sources

We present below the data sources used in our analytical approach:

- *Auditoria Eletrônica de Órgãos Públicos (AUDESP)*: public spending data of the state of São Paulo for the period from 2016 to 2018, containing R$ 71 billion in contracts, R$ 89 trillion in proposals, and 55,000 companies who bid a contract or made a proposal.
- *Sistema de Acompanhamento da Gestão dos Recursos da Sociedade (SAGRES/PB)*: public spending data of the state of Paraíba for the period from 2009 to 2016, containing R$ 24 billion in contracts, R$ 3 trillion in proposals, and 24,000 companies.
- State-level lists of public suppliers: lists with 166,000 different public suppliers of the states of Acre, Amapá, Bahia, Ceará, Mato Grosso do Sul, Minas Gerais, Distrito Federal, and Espírito Santo for the period from 2009 to 2019.

- Federal conditional cash transfer programs: data related to *Bolsa Família* and other social programs for the period from 2013 to 2019, with approximately 7,000,000 social beneficiaries of programs *Garantia Safra*, *Seguro Defeso*, *Benefício de Prestação Continuada*, and *Programa de Erradicação do Trabalho Infantil* and 13,500,000 families of *Bolsa Família* itself.
- Blacklisted companies: blacklists of companies punished by a Brazilian state-level or Federal court, including different types of punishment for the period from 2003 to 2019, containing more than 9000 blacklisted companies and 10,000 blacklisted people.
- *Cadastro Nacional de Empresas (CNE)*: historical corporate data of Brazilian companies and its partners until 2016, containing 17,000,000 companies and their 20,000,000 partners.
- *Receita Federal*: corporate data of Brazilian companies for the year 2017, containing 35,000,000 companies and their 12,500,000 partners.
- *Tribunal Superior Eleitoral (TSE)*: dataset with personal data of politicians and candidates, electoral spendings and donations, and afiliates of political parties for the period from 2006 to 2018, containing data about 1,000,000 people (including 133,000 elected politicians) and details about R$ 15 billion in electoral spendings and R$ 14 billion in donations.
- Financial sector companies: list of banks and other financial institutions according to the Brazilian Central Bank for the year 2019, containing data of almost 50,000 companies related to the financial sector.
- Politically exposed persons: list of politicians, ministers, and directors of public agencies for the period from 2011 to 2019, containing data of almost 70,000 people.
- 2010 census data, including several variables such as household income and education level for each of the 300,000 census sectors in Brazil.

2.2. Scalable data unification

The ETL process (Vassiliadis, 2009; Wibowo, 2015) consisted mostly of three phases: data collection, scalable data unification (SDU), and migration of the data sources into our data lake.

A global schema was built to allow the unification of different data stores into one single relational database. The SDU process is summarized in Fig. 1. The unified database schema contains all relevant information contained in the local schemes of the data sources mentioned in the previous section. It contains information on the entire public procurement process, starting from the bidding process and ending with the contracts and payments.

The advantage of using a global schema instead of a set of several local schemes consists of decreasing the marginal cost of implementing the risk detection patterns for each of the public expenditure datasets that are being unified. The algorithm that obtains each risk pattern is implemented only once, for the unified database, as illustrated in Fig. 2. The demand for obtaining the risk patterns for a new public expenditure dataset is faced reducing the original dataset to its unified version and then executing the already existing algorithms. The effort required to implement this procedure is several times smaller than that of implementing all risk detection patterns for each dataset separately.

Cleaning and transformation procedures analyze the consistency of the data that are being migrated. A collection of rules for matching and merging the data was also developed.

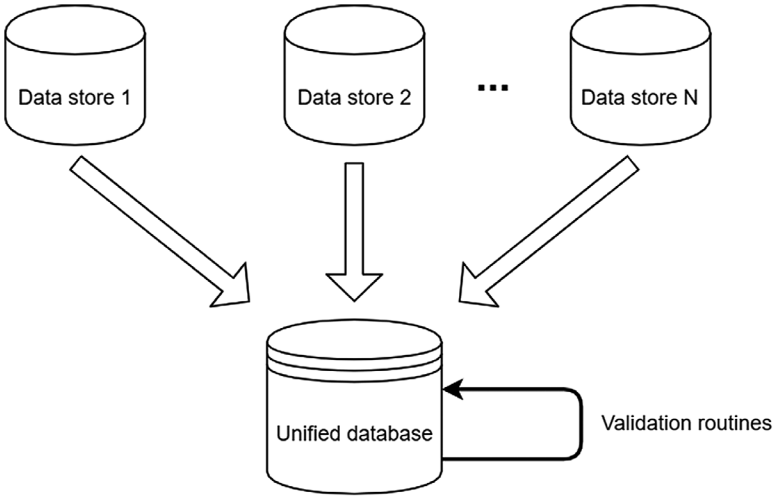


Fig. 1. Scalable data unification: different data stores with their respective schemes are unified into a single unified database. This process is performed over public procurement, electoral, and cash transfer program data.

We also designed validation routines for ensuring current and future data migrations to satisfy a set of specific assertions, such as

- filling columns with critical data (e.g., bidder and contractor identification, completeness of public entities;
- verification of invariants relating original and migrated data (for each public contractor, total values spent by month and by year); and
- correctness of check digits in document numbers.

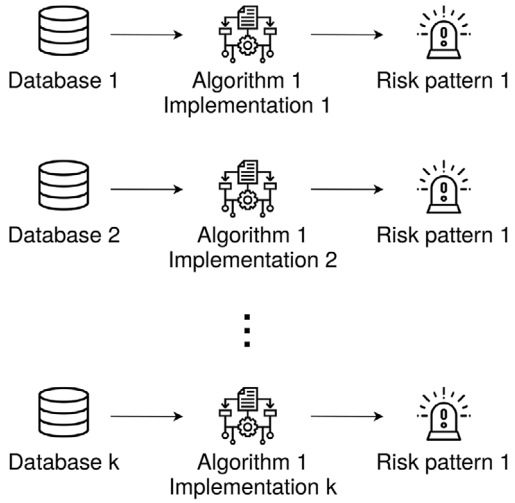
Similarly, the data unification process was performed for electoral data and cash transfer program datasets. However, in these cases, the dimensionality of the data was much lower. Consequently, the complexity of the unification process was much simpler when compared to the unification of public expenditure data sources.

2.3. General scheme of the analytical approach

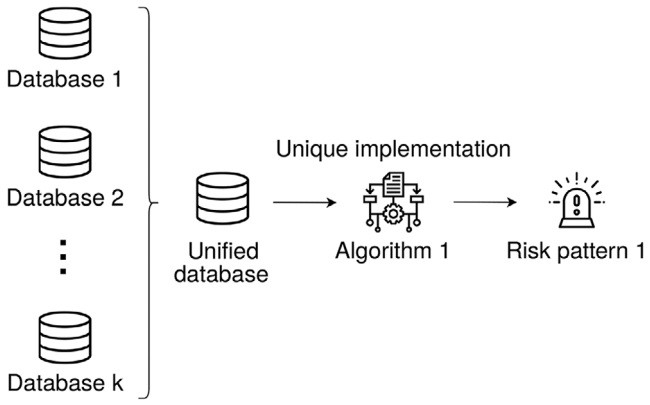
The general scheme of the analytical approach is described in Fig. 3.

The first stage (data fusion) consists in the lower level of the figure. Data fusion is a "multilevel, multifaceted process dealing with the detection, association, correlation, and combination of data and information from multiple sources" (Kalivas, 2019). The unified datasets that resulted from SDU and the raw datasets are combined and stored in a data lake.

In the second stage (analytics), we use several algorithms developed for company-level fraud detection in bidding processes and contracts with public institutions. In addition to techniques based on descriptive statistics and outlier detection methods, we developed specific algorithms for extracting company-level risk patterns that are described in Sections 3 and 4.



(a) Case when scalable data unification is not implemented.



(b) Case when scalable data unification is implemented.

Fig. 2. Scalable data unification: the algorithm that obtains each risk pattern is implemented only once, for the unified database.

3. Risk patterns for fraud detection

We conceptually describe in this section the main risk patterns that the DSS can detect. Each risk pattern allows the evaluation of different variables related to companies, physical persons, or groups of companies. For example, the blacklisting risk pattern for some company allows to compute the number of contracts of this company during the period it was blacklisted, the total value of contracts of this company during this period, and the ratio between this value and the total value of contracts of this company over the period of the investigation. The DSS computes more than 200 variables based on the risk patterns presented below. Due to space limitations, we do not detail all these variables individually.

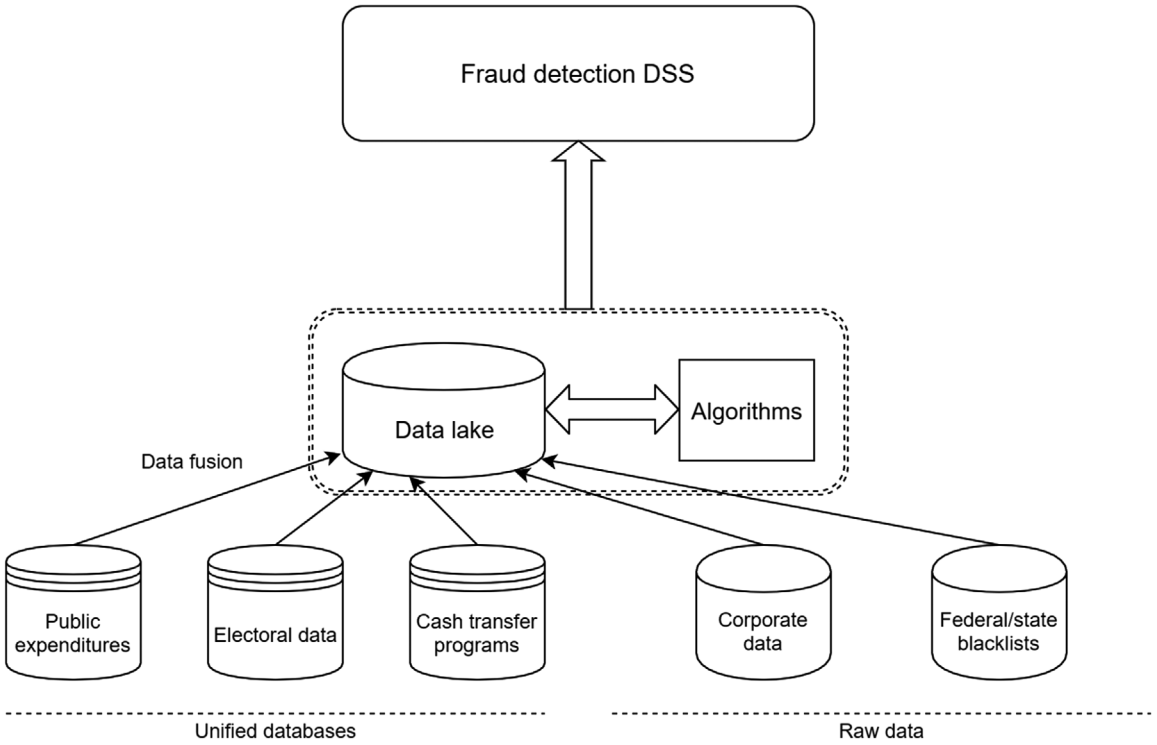


Fig. 3. General scheme of the analytical approach.

3.1. Collusion risk patterns

The Brazilian law (Law 8666/93, Article 90 of 21 June 1993) considers as a crime the act of frauding the competitive nature of the bidding process, in order to obtain any advantage resulting from winning that process. Santos and de Souza (2016) observe that collusion risk patterns can be classified in (a) economic advantages or (b) indications of a previously combined action between the competitors.

We provide the characterization of collusion risk patterns, together with the description of the related risk for the public institution that executes the bidding process or spending. These are more complex risk patterns, since each of them involves two or more companies. Some algorithms used for identifying these risk patterns are presented in Section 4.

- *Identical bids*: two companies A and B participate in different bidding processes, always with the same bid values C_A and C_B. Related risk: the two companies are possibly in collusion. The two values C_A and C_B have been possibly prearranged and therefore there is no real competition.
- *Concerted bids*: two companies A and B participate in different bidding processes, with bid values C_A and C_B such that their ratio C_A/C_B or their difference $C_B - C_A$ is always the same. Related risk: the two companies are possibly in collusion. The values C_A and C_B have been possibly prearranged and therefore there is no real competition.

- *Top losers*: company A participates in a large number of bidding processes but never (or almost never) wins. Related risk: company A only participates with the aim of giving an appearance of competitiveness. Possibly, A is a shell company in collusion with some winning company that controls it.
- *Common partners*: two companies A and B with a common partner P participate in the same bidding processes. Related risk: the two companies are possibly in collusion. Competition is impaired because P can influence the bids of both A and B.
- *Common addresses*: two companies A and B with the same publicly declared address participate in the same bidding processes. Related risk: competition is impaired because there is possibly some relation between the two companies.
- *Common economic group*: two companies A and B in the same economic group (e.g., with the same parent company) participate in the same bidding processes. Related risk: the two companies are possibly in collusion. Competition is impaired because no matter which company wins, the same economic group will be awarded.
- *Common registration data*: two companies A and B with the same IP address participate in the same online bidding processes. Related risk: competition is impaired because there is possibly some relation between the two companies.

3.2. Company-level risk patterns

In addition, we compute a number of company-level statistics and risk patterns that characterize the participation of each company in public spending. For each company, we check:

- *Blacklisting*: presence in federal or state-level blacklists.
- *Links with political campaigns*: if the company was a supplier or made donations during political campaigns.
- *Incompatible company size*: if a self-employed person (*Microempreendedor Individual*, in Portuguese) won public contracts above income limits defined by the Brazilian legislation.
- *Contract earliness*: if the date of the first contract or bid of the company is very close to its registration in the dataset of *Receita Federal*.
- *Large number of activity classifications*: if the number of different economic activity classifications in the registration data of the company is excessively large or the activity types are incompatible with each other.
- *Cost overruns*: high value of the ratio between the final and original values of contracts of this company.
- *Direct contracts*: high value of contracts without a call for bids.
- *High winning rate*: high value of the ratio between the number of winning bids and the total number of bids of the company.
- *Conflicts of interest*: when a person strongly related to the administration of some government body owns a company contracted by this same body (see Fig. 4).

Whenever any of these risk patterns is detected for some company, it indirectly affects the risks associated with the other companies of its partners, as well as the risks of other companies belonging

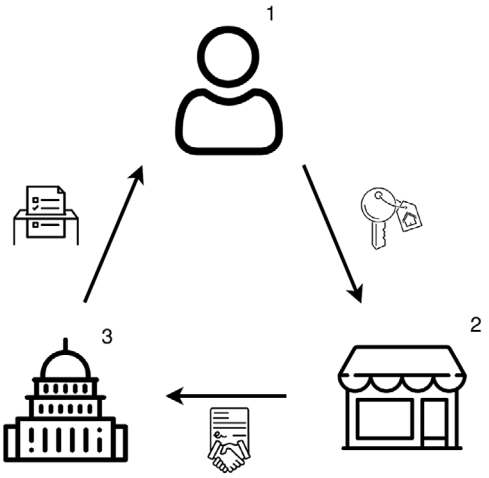


Fig. 4. A typical case of a conflict of interest occurs when a politician (1) is a partner of a company (2) contracted by the same government body (3) where the politician was elected. This pattern can be detected using the datasets of company partners (edge 1-2), of public spending (edge 2-3), and of personal data of politicians and candidates (edge 3-1).

to the same economic group of this company. Furthermore, the risks of all partners of the aforementioned companies are also affected. The risk analysis of all related companies and physical persons must use these risk patterns for showing to the decision maker the richest available information.

We note that risk patterns associated with the partners of a company as physical persons, which are listed in Section 3.3, are also considered as risk patterns of the company itself.

3.3. Person-level risk patterns

We also computed a number of risk patterns associated with physical persons. We used them for identifying whether some partner of a company meets the following criteria:

- *Blacklisting*: presence in federal or state-level blacklists.
- *Politically exposed persons*: if the partner is a candidate or an elected official in at least one election, a member of the directorship of a political party, or a politically exposed person.
- *Links with political campaigns*: if the partner received payments for goods or services or made donations during political campaigns.
- *Cash transfer programs*: if the partner is a beneficiary of a conditional cash transfer program (*Bolsa Família*, *Garantia Safra*, *Seguro Defeso*, *Benefício de Prestação Continuada*, and *Programa de Erradicação do Trabalho Infantil*).

Once again, we observe that many patterns do not represent by themselves individually a significant risk of the existence of fraud or corruption. However, the risk of fraud increases when combinations of these risk patterns are identified for some physical person or company. In the course of any investigation making use of the DSS, investigators of the law enforcement agencies should decide which combinations of risk patterns are more relevant for each specific investigation.

Table 1
Identical bids

Municipality	C_B	C_A
1	R$ 100,000	R$ 70,000
2	R$ 100,000	R$ 70,000
3	R$ 100,000	R$ 70,000

Table 2
Concerted bids with same ratios

Bidding process	C_B	C_A	Ratio C_B/C_A
1	R$ 15,000	R$ 10,000	1.5
2	R$ 150,000	R$ 100,000	1.5
3	R$ 450,000	R$ 300,000	1.5

4. Methods and algorithms

In this section, we describe some algorithms developed for performing automatic or semiautomatic analysis of a large amount of public procurement data to extract specific risk patterns from the data. Since the collusion risk patterns are more critical and involve more sophisticated algorithms, we focus on algorithms for the risk patterns described in Section 3.1.

4.1. Identical and concerted bids

In the basic and most simple case of identical bids, two companies A and B participate in different bidding processes, possibly in different municipalities, always with the same bid values C_A and C_B; see Table 1.

Although nothing can be affirmed, this situation is an indication that the two companies have some kind of relationship. The two companies are possibly in collusion, in particular if this situation is observed over several municipalities. The two values C_A and C_B have been possibly prearranged by the two companies.

A variation of this risk pattern is that of concerted bids. In this case, the bid values C_A and C_B are not the same over different bidding processes but, instead, either the ratio C_A/C_B or the difference $C_A - C_B$ is the same, as illustrated in Table 2.

In order to search for concerted bids, for each bidding process we need the CNPJ (the Brazilian national register of legal entities) of each company making a bid and the value it proposes, as illustrated in Table 3. In this example, company 00.000.000/0000-00 bids twice a value corresponding to 1.33 times the bid of company 11.111.111/1111-11 (rows in red and blue).

Table 4 outputs the number of occurrences of concerted bids, indicating the number of times any two companies A and B made bids with the same ratio C_A/C_B between their values.

Algorithm 1 is used to compute concerted bids. Specific data science tools, features, and optimizations were used in the implementation. We used *Pandas* (McKinney, 2017) for data manipulation

Table 3
Data for identification of concerted bids

	Company	
Bidding process	CNPJ	Bid value
1	00.000.000/0000-00	R$ 10,000
1	11.111.111/1111-11	R$ 7,500
1	22.222.222/2222-22	R$ 9,000
2	00.000.000/0000-00	R$ 15,000
2	11.111.111/1111-11	R$ 11,250
2	22.222.222/2222-22	R$ 14,000
3	33.333.333/3333.33	R$ 10,000
3	44.444.444/4444.44	R$ 5,000

Table 4
The number of occurrences of concerted bids: the company with CNPJ 00.000.000/0000-00 bids twice a value corresponding to 1.33 times the bid of company with CNPJ 11.111.111/1111-11 (in bidding processes 1 and 2)

CNPJ A	CNPJ B	Ratio C_A/C_B	Occurences	Bidding processes
00.000.000/0000-00	11.111.111/1111-11	1.33	2	1,2

and analysis, together with other libraries such as *NumPy* and *Matplotlib*, in Algorithm 1 and other algorithms.

Algorithm 1. Concerted bids

1: $R \leftarrow$ list of all bids, each one containing the identification of the bidding process, the CNPJ of the bidding company, and the bidding value.
2: Remove from R all bids related to bidding processes with only one bidding company.
3: For each bidding process, compute $C = R \times R$ restricted to the companies participating in this process.
4: Remove the rows of C corresponding to pairs of companies with the same CNPJ.
5: For each row of C, compute the ratio (or the difference) between C_A (first company) and C_B (second company).
6: Remove duplicate rows of C with symmetrical information (same bidding process, but with companies swapped).
7: Count repetitions in C of each different triple (first company, second company, ratio value).

Table 5 illustrates the relevant information stored in C after steps 1–6 of the above algorithm have been performed. There are two repetitions of the triple (00.000.000/0000-00,11.111.111/1111-11,1.33), one for bidding process 1 and another for bidding process 2. The final results already appeared in Table 4.

4.2. Common addresses

This is the situation when two companies A and B with the same publicly declared address participate in at least one common bidding process. There is a risk of collusion between the two companies, with the result of the bidding process being prearranged.

Table 5
Results after steps 1–6: rows marked in red are redundant and will be removed before the final counting step

Bidding process	CNPJ *A*	C_A	CNPJ *B*	C_B	Ratio
1	00.000.000/0000-00	10,000	11.111.111/1111-11	7,500	1.33
1	00.000.000/0000-00	10,000	22.222.222/2222-22	9,000	1.11
1	11.111.111/1111-11	7,500	00.000.000/0000-00	10,000	0.75
1	11.111.111/1111-11	7,500	22.222.222/2222-22	9,000	0.83
1	22.222.222/2222-22	9,000	00.000.000/0000-00	10,000	0.90
1	22.222.222/2222-22	9,000	11.111.111/1111-11	7,500	1.20
2	00.000.000/0000-00	15,000	11.111.111/1111-11	11,250	1.33
2	00.000.000/0000-00	15,000	22.222.222/2222-22	14,000	1.07
2	11.111.111/1111-11	11,250	00.000.000/0000-00	15,000	0.75
2	11.111.111/1111-11	11,250	22.222.222/2222-22	14,000	0.80
2	22.222.222/2222-22	14,000	00.000.000/0000-00	15,000	0.93
2	22.222.222/2222-22	14,000	11.111.111/1111-11	11,250	1.24
3	33.333.333/3333.33	10,000	44.444.444/4444.44	5,000	2.00
3	44.444.444/4444.44	5,000	33.333.333/3333.33	10,000	0.50

Each address is composed of five main elements: state, municipality, district (*bairro*), street (or similar), and number. The exploratory analysis of the data showed that state and municipality names are highly standardized. However, district and street names can be written in several ways. For example, the street named "Rua São João" also appears as "R. São João," "R. S. João," or simply "São João," among others, in addition to misspellings. In order to identify this risk pattern, the main computational challenge consists in determining, given the addresses of two companies *A* and *B* participating in the same bidding process, whether both addresses actually represent the same geographical location.

We designed and implemented a three-phase approach for the identification of similar addresses. Its three phases are

1. clusterization of district names;
2. clusterization of street names; and
3. identification of addresses in the same district and street clusters that have the same street number.

We used algorithm DBSCAN: density-based spatial clustering (Ester et al., 1996) for the clusterization of district and street names, implemented in the Scikit-learn library (Pedregosa et al., 2011). DBSCAN uses a customized distance metric between the objects to be clusterized. We used the well-known Levenshtein (1965) distance as the metric of dissimilarity between two strings, which is equal to the minimum number of character edits (insertions, deletions, or replacements) required to transform one string into the other. The algorithm assigns district and street cluster identifiers to each address.

In practice, given the high standardization of state and municipality names, the same address detection algorithm is applied inside each municipality, reaching the number of 5570 algorithm instances (corresponding to the total number of municipalities in Brazil) that can be executed concurrently.

However, the complexity of the instances varied considerably. In a naive implementation, the largest instance, corresponding to the municipality of São Paulo, showed an execution time of several months. Specific algorithm optimization and tuning procedures were required for this case. The parallel implementation uses several threads and a low-level implementation of the Levenshtein distance, which combined reduced the execution time by a factor of 20.

We observe that the company addresses written in a very distorted or wrong way cannot be identified in some cases because severe errors and misspellings would prevent the clusterization of these addresses by the clusterization algorithm.

4.3. Top losers

A set of bidding processes carried out in a given municipality during a certain time window implicitly defines a directed weighted graph $G(V, A, w)$, whose vertex set V is formed by the participating companies, each of them uniquely identified by its CNPJ. For each finished bidding process with N participants, we assume that there is one winning company and $N - 1$ losing companies (the current algorithm can be straightforwardly adapted in case this assumption is not satisfied). The set of arcs $A \subseteq V \times V$ of the graph is such that arc $(x, y) \in A$ if and only if company x won company y in at least one bidding process. The weight $w(x, y)$ of arc (x, y) denotes the number of times company x won over company y during the considered period.

Graph $G(V, A, w)$ is very large, since there are millions of registered companies and tens of thousands of these companies sometimes may participated in bidding processes during some time interval. Therefore, we consider a refinement of $G(V, A, w)$ corresponding to a subgraph formed only by the most relevant vertices (companies) and arcs (relations).

We refer to companies that participate in a large number of bidding processes, losing all (or almost all) of them as *top losers*. Top loser companies are relevant because they are often shell companies deliberately used to fraud bidding processes. Two parameterized criteria are used in the search for top loser companies:

- *Participations in bidding processes*: the companies that are considered as possible top losers are those with the largest number of participations in bidding processes, that is, those that have participated in at least X bidding processes, where X is a parameter.
- *Winning rates*: the winning rate r_x of a company x is defined as the number of wins in bidding processes divided by the total number of participations in bidding processes. The candidate companies to be top losers are those with the lowest winning rates, that is, those whose winning rate is smaller than or equal to Y, where Y is a parameter.

For some values of the parameters X and Y, the two above primary filtering criteria may produce a large number of top loser candidates. In such cases, these two criteria could be adjusted to reduce the number of companies that will be considered as top losers, with the goal that each of them could be investigated in detail by agents of the law enforcement agencies.

More formally, the set of top loser candidates T_L is defined as

$$T_L = \left\{ x \in V \ : \ \sum_{y \in V} w(y, x) \geq X, r_x \leq Y \right\}.$$

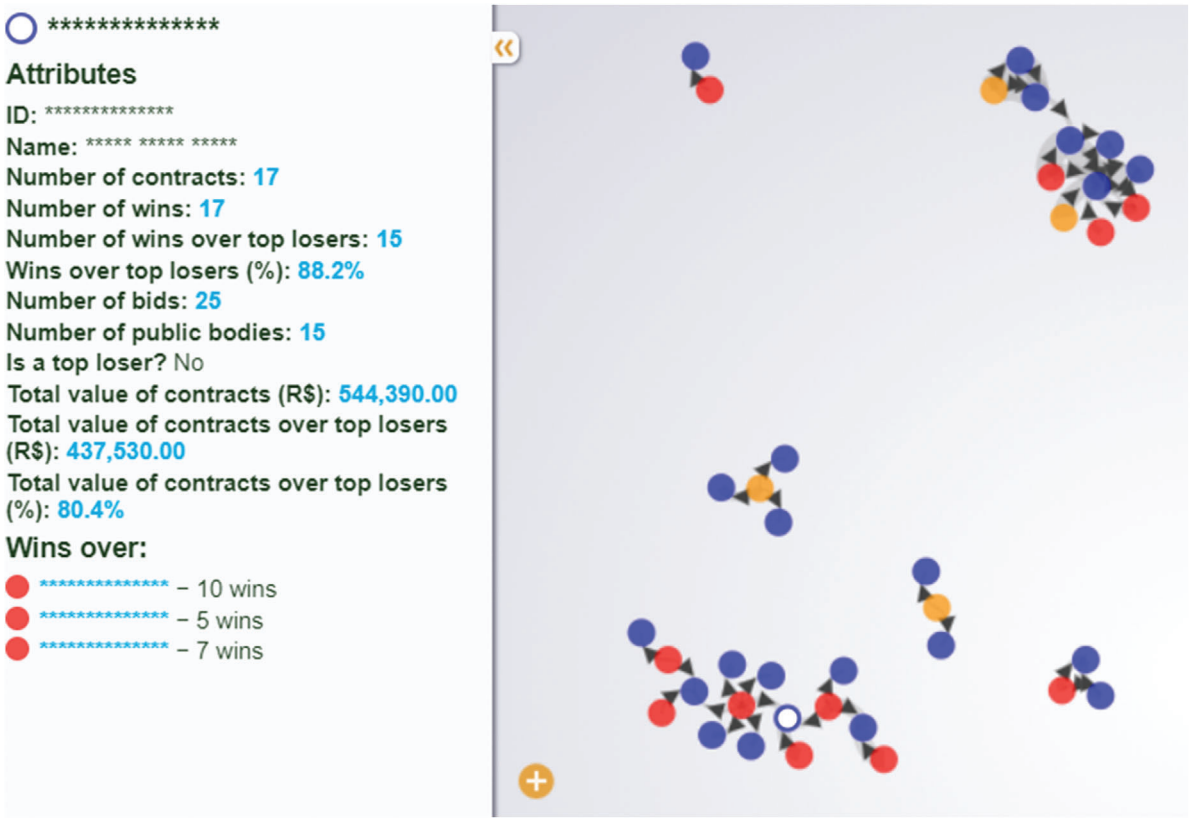


Fig. 5. Visualization produced by the decision support system panel showing some weakly connected components of the top losers graph built for the bidding processes of the state of Paraíba. Red and yellow vertices represent top loser companies with zero and more than zero victories, respectively. Blue vertices represent companies that won over top losers in at least k bidding processes.

The subgraph $G'(V', A', w)$ contains all arcs of $G(V, A, w)$ that are incident to the set T_L of top losers vertices, with

$$V' = T_L \cup \{x \in V \ : \ \exists y \in T_L \text{ such that } (x, y) \in A \text{ and } w(x, y) \geq k \text{ or } (y, x) \in A \text{ and } w(y, x) \geq k\}$$
$$A' = \{(x, y) \in A \ : \ x \in V', y \in V'\},$$

where T_L is the set of top losers defined as above and k is a filtering criterion that indicates a minimum number of victories over top loser candidates.

Figure 5 illustrates an anonymized instance of the top losers graph $G'(V', A', w)$ as it appears in the panel of the DSS. At its current stage, the interface is in Portuguese, but it can be easily customized for any language. A weakly connected component (Harary, 1969) of $G'(V', A', w)$ with 15 vertices (of which 9 are blue and 6 are red) can be found at the bottom of the figure. The blue vertex with a white interior corresponds to the selected company that is a frequent winner

of top losers. It is connected to three red vertices that represent top losers. Additional details about this company are shown in the left sidebar, displaying the number of bids of the company, the number of wins in these bids, the total contractual value these wins represented, and other useful information. On the sidebar, it is also possible to identify the top loser companies (red vertices, also anonymized) won by the chosen company, as well as the number of victories over each of them.

These companies closely related to top losers are very relevant because strong links between them and losing companies may be indicative of collusion. Companies that repetitively win over top loser companies often control them using straw persons. In order to determine the companies that will be further investigated in detail, some criteria should be considered, such as the number of victories over each top loser and the percentage of the values of contracts that a company has won against top losers. These secondary criteria are used by the user of the DSS to further refine the graph, choosing the companies that are more relevant for the investigation.

5. Experience in practice: quantitative and operational results

The DSS was implemented in the Public Prosecutor's Offices of the states of São Paulo and Paraíba and has been in use since 2018. Some quantitative and qualitative results obtained with the practical experience and the use of the system in real-life investigations are reported in this section.

5.1. Quantitative results

In order to measure the impact of the findings of the DSS, we summarize some quantitative results obtained using public spending data of the state of Paraíba:

- Direct contracts, without a call for bids: 2308 companies, with more than R$ 806.47 million in contracts.
- Conflicts of interest, when a politician owns a company contracted by the same government body where the politician was elected: 122 companies, with more than R$ 226.42 million in contracts.
- Companies whose partners are beneficiaries of the conditional cash transfer program *Bolsa Família*: 312 companies, with more than R$ 236.04 million in contracts.
- Top losers: 140 companies that won more than once over top losers, with more than R$ 237.32 million in contracts. This risk pattern considers as a top loser a company that participated in at least 15 bidding processes and lost more than 95% of them.
- Common partners: 857 companies that won bidding processes had common partners with other bidding companies, with more than R$ 3639.95 million in contracts.

We recall that the Brazilian Real was quoted at US$ 0.3023 on January 1, 2018. The above numbers give a clear picture of the number of irregularities and the amount of public procurement involved in different risk patterns that can be detected using public data. We recall that Paraíba is one of the smallest states in Brazil, which gives a projection of the possible figures for other, larger states.

The risk patterns calculated for the Public Prosecutor's Office of the state of Paraíba have been effectively used and applied as inputs for several investigations, which led, for example, to the operations that are mentioned in Section 5.2.

5.2. Operational results

The techniques exposed in this work and the quantitative results summarized in Section 5.1 depict unambiguously the need to implement analytical approaches in the public administration and the power branches of the State.

There is an utmost need of transforming the decision-making process in law enforcement agencies. It should be based on knowledge discovery in databases processes and data mining, making repressive and preventive actions more assertive and giving rise to irrefutable practical results, because of the reduction of operating costs and the systematic increase in the confidence of the actions of the State.

It is notable that the estimation of risks associated with companies contracted by public administration bodies (municipalities or state governs) using analytical approaches proves to be useful for law enforcement agencies, helping them to structure their repressive or preventive actions in the short term. Also, in the medium term, they can be useful to public administrators in municipalities and states, by allowing them to implement public policies, more in line with the social pains. Moreover, in the long term, it is beneficial for the population since it allows to exercise a democratic citizenry through the rational use of voting.

Even if it is necessary to optimize public resources by tracing prioritization strategies, it is also essential to use these resources in a justified and temperate manner, minimizing any risks associated with public contractors.

Modeling the characteristics of the companies using specific risk patterns results in a better performance in their risk-driven classification, in comparison with more traditional methods.

The use of the DSS in this scenario led to much more accurate analyses in many cases, which in the state of Paraiba resulted in the *Xeque-Mate* and *Calvário* operations, with previously unattainable pathways.

The *Xeque-Mate* operation (Polícia Federal, 2018; G1, 2018a, 2018b), in addition to several risk patterns, also made use of a specific algorithm to identify public servants who were not entitled to their salaries, since, in fact, they were not actual employees in the public institutions. It soon became evident that the earnings of these *ghost-servants* were shared with the political agents who appointed them.

The *Calvário* operation (JOTA, 2018; Correio da Paraíba, 2019; G1, 2019; Polícia Federal, 2020), on the other hand, used business rules (risk patterns) that allowed identifying farms of shell companies related to social organizations. The identification of these companies made possible an increase in assertiveness, opening the path to the entire money laundering chain.

6. Concluding remarks and extensions

In this work, we reported a successful combination of operations research methods (such as mathematical modeling, graph theory, and network analysis) with data science tools (such as data mining,

clustering algorithms, anomaly detection, and risk ranking) in the development of a DSS for law enforcement agencies in Brazil.

The DSS and the proposed algorithms made possible the detection of risk patterns in the amount of hundreds of millions of Brazilian Reals in public expenditures. They also provided inputs to several investigations, resulting in at least two large operations (*Xeque-Mate* and *Calvário*) conducted by Prosecutor's Office of the state of Paraiba in collaboration with other law enforcement agencies: the Federal Police of Brazil (*Polícia Federal do Brasil*) and the Prosecutor's Office of the state of Rio de Janeiro.

These investigations made possible the recovery of significant amounts of funds diverted from the public treasury. Part of the resources recovered by the Calvário operation has already been returned to society. In March 2020, the Prosecutor's Office of the state of Paraiba announced the decision to use R$ 825,000 of the recovered funds to acquire 15 machine ventilators for hospitals of the municipalities of João Pessoa and Campina Grande and for *Hospital Universitário Lauro Wanderley* (Paraíbaonline, 2020). Ventilators are critical breathing machines for the treatment of COVID-19 (the disease caused by the new coronavirus) patients. Some days later, following a request of the Prosecutor's Office, a judge of the Court of the Justice of the state of Paraíba determined that R$ 399,000 recovered from a whistleblower of the same Calvário operation should be immediately used for acquiring 2660 COVID-19 immunofluorescence tests (José, 2020).

Other countries face similar institutional problems related to corruption and to the need to implement systematic detection of risk patterns in public procurement data. Public procurement presents several procedural similarities across countries and across subnational units. Public expenditures datasets have been available for many years in virtually all developed countries and in some middle-income countries. Assuming the availability of the main sources of public data used in this work (such as public suppliers, bidding processes, corporate data of all companies and their partners, and beneficiaries of social and cash transfer programs), the technology presented here could be adapted and transferred to other countries. Corruption measures specifically related to public procurement have the potential of being compared across subnational units of the same country or even among different countries. This approach could be a first step for future studies that might be able to overcome the challenge of creating universally accepted objective measures and transnational proxies for corruption.

Among the extensions of this work, we mention the use of the electronic invoices as another source of information for constructing specific risk patterns. Unlike typical public expenditure databases containing aggregated data about contracts and bidding processes, electronic invoices contain a detailed description of the products that are sold to a public body. This information allows, with a high degree of confidence, the identification of fraud risk patterns related to the presence of inconsistencies between the quantities of a specific product (e.g., gasoline or medicines) that are sold to some municipality or state government, and the expected quantities this entity should consume, based on estimations, for example, from its population. It is also possible to identify unreasonably high prices of some products as outliers of the price distributions.

Another extension in progress is related to a more accurate identification of straw persons, that is, physical individuals who legally appear as partners (or have the legal appearance of being partners) of some company, but do so on behalf of another person, generally for illegal purposes. Values received by these individuals as beneficiaries of social programs; the Human Development Index associated with addresses of their companies; and amounts donated to and received from political

candidates during electoral campaigns are proxies of the income of these individuals that can be used in the identification of straw persons.

We are also interested in finding collusion patterns based on abnormal distributions of winning rates of some set of companies over different municipalities or states. Significant differences between winning rates in several geographical regions may be an indication of geographical cartelization.

Glossary

- *Bolsa Família*: the largest and most important social program in Brazil. It provides a monthly financial aid to low income Brazilian families.
- *Garantia Safra*: a social benefit that guarantees to farmers and their families a guaranteed minimum income in case of droughts or floods.
- *Seguro Defeso*: a social benefit to fishermen during the period in which they are prevented from fishing due to the need to preserve species.
- *Programa de Erradicação do Trabalho Infantil*: a social program that protects children and teenagers under the age of 16, providing financial support to their families if they attend school regularly instead of prematurely working at inadequate jobs.
- *Benefício de Prestação Continuada*: a social benefit equivalent to retirement to elderly and disabled people who cannot support themselves and that cannot be maintained by their families.
- *Receita Federal*: the Brazilian federal revenue agency.
- *Tribunal Superior Eleitoral*: the highest court of the Brazilian electoral justice.

Acknowledgments

The authors are thankful to the Prosecutor's Office of the state of Paraíba for its invaluable collaboration during this work. They are also grateful to one anonymous referee for the constructive remarks who contributed to improve this article. Rafael B. Velasco, Igor Carpanese, and Ruben Interian thank Daniel Ortega Nieto for his encouragement. Work of Celso C. Ribeiro was partially supported by CNPq research grants 303958/2015-4 and 425778/2016-9, by FAPERJ research grant E-26/202.854/2017, and by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Finance Code 001. The authors also thank Rodrigo Luna for his contribution to Section 4.2 of this article.

References

Broms, R., Dahlström, C., Fazekas, M., 2019. Political competition and public procurement outcomes. *Comparative Political Studies* 52, 1259–1292.

Correio da Paraíba, 2019. Calvário prendeu políticos e expôs maior esquema de corrupção da PB. Available at https://portalcorreio.com.br/retrospectiva-2019-operacao-calvario/ (accessed on 27 March 2020).

Dávid-Barrett, E., Fazekas, M., 2020. Anti-corruption in aid-funded procurement: is corruption reduced or merely displaced? *World Development* 132. https://doi.org/10.1016/j.worlddev.2020.105000.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press, Portland, OR, pp. 226–231.

Fazekas, M., King, L.P., 2019. Perils of development funding? The tale of EU funds and grand corruption in Central and Eastern Europe. *Regulation & Governance* 3, 405–430.

Fazekas, M., Tóth, I.J., King, L.P., 2014. Anatomy of grand corruption: a composite corruption risk index based on objective data. Technical Report, CRCB Working Paper 2013:02, Corruption Research Center, Budapest. Available at https://www.againstcorruption.eu/publications/anatomy-of-grand-corruption-a-composite-corruption-risk-index-based-on-objective-data/ (accessed 29 April 2020).

Fazekas, M., Tóth, I.J., King, L.P., 2016. An objective corruption risk index using public procurement data. *European Journal on Criminal Policy and Research* 22, 369–397.

G1, 2018a. Entenda como a "Xeque-Mate" derrubou prefeito e vereadores de Cabedelo. Available at https://g1.globo.com/pb/paraiba/noticia/entenda-como-a-xeque-mate-derrubou-prefeito-e-vereadores-de-cabedelo.ghtml (accessed 27 March 2020).

G1, 2018b. "Xeque-Mate": ex-prefeitos, empresário e radialista viram réus por compra de mandato na PB. Available at https://g1.globo.com/pb/paraiba/noticia/2020/01/27/xeque-mate-ex-prefeitos-empresario-e-radialista-viram-reus-por-compra-de-mandato-na-pb.ghtml (accessed 27 March 2020).

G1, 2019. Governador e ex-governador da Paraíba são alvos da "Operação Calvário." Available at https://g1.globo.com/google/amp/pb/paraiba/noticia/2019/12/17/ex-governador-ricardo-coutinho-e-alvo-da-setima-fase-da-operacao-calvario-na-paraiba.ghtml (accessed 27 March 2020).

Harary, F., 1969. *Graph Theory*. Addison-Wesley, Boston, MA.

José, M., 2020. A pedido do Gaeco/MPPB desembargador Ricardo Vital destina R$ 399 mil devolvidos por Livânia Farias na Operação Calvário, ao HU para combate ao coronavírus. Available at https://marcelojose.com.br/2020/03/27/a-pedido-do-gaecomppb-desembargador-ricardo-vital-destina-r-399-mil-devolvidos-por-livania-farias-na-operacao-calvario-ao-hu-para-combate-ao-coronavirus/ (accessed 27 March 2020).

JOTA, 2018. Ministério Público da Paraíba usa tecnologia e ciência de dados em investigações. Available at https://www.jota.info/coberturas-especiais/inova-e-acao/mppb-tecnologia-ciencia-operacoes-20112018 (accessed 27 March 2020).

Kalivas, J.H., 2019. Data fusion of nonoptimized models: applications to outlier detection, classification, and image library searching. In Cocchi, M. (ed.) *Data Fusion Methodology and Applications, Data Handling in Science and Technology*, Vol. 31. Elsevier, Amsterdam, pp. 345–370.

Kar, D., Spanjers, J., 2014. Illicit financial flows from developing countries: 2003–2012. Technical Report, Global Financial Integrity, Washington, DC.

KPMG, 2009. Forensic Advisory—A Fraude no Brasil: Retatório de Pesquisa 2009. Available at http://www.kpmg.com.br/publicacoes/forensic/Fraudes_2009_port.pdf (accessed 9 April 2020).

Levenshtein, V.I., 1965. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10, 707–710.

McKinney, W., 2017. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Sebastopol, CA.

Paraíbaonline, 2020. Gaeco comenta sobre respiradores comprados com recursos resgatados da "Calvário." Available at https://paraibaonline.com.br/2020/03/gaeco-comenta-sobre-respiradores-comprados-com-recursos-regatados-da-calvario/ (accessed 27 March 2020).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.

Polícia Federal, 2018. Polícia Federal deflagra 6 fase da Operação Xeque-Mate. Available at http://www.pf.gov.br/imprensa/noticias/2019/12/policia-federal-deflagra-6a-fase-da-operacao-xeque-mate (accessed 27 March 2020).

Polícia Federal, 2020. PF deflagra 8 fase da Operação Calvário e investiga esquema de desvio de recursos públicos em PB. Available at http://www.pf.gov.br/imprensa/noticias/2020/03-noticias-de-marco-de-2020/pf-deflagra-8a-fase-da-operacao-calvario-e-investiga-esquema-de-desvio-de-recursos-publicos-em-pb (accessed 27 March 2020).

Santos, F.B., de Souza, K.R., 2016. *Como Combater a Corrupção em Licitações*. Editora Forum, Belo Horizonte, Brazil.

Transparency International, 2006. *Handbook for Curbing Corruption in Public Procurement*. Transparency International, Berlin. Available at https://www.transparency.org/whatwedo/publication/handbook_for_curbing_corruption_in_public_procurement (accessed 27 March 2020).

Vassiliadis, P., 2009. A survey of extract–transform–load technology. *International Journal of Data Warehousing and Mining* 5, 1–27.

Weeks-Brown, R., 2018. Cleaning up: countries are advancing efforts to stop criminals from laundering their trillions. *Finance & Development* 55, 44–45.

Wibowo, A., 2015. Problems and available solutions on the stage of extract, transform, and loading in near real-time data warehousing (a literature study). 2015 International Seminar on Intelligent Technology and Its Applications, Surabaya, Indonesia, pp. 345–350.