# PLUS: A Semi-automated Pipeline for Fraud Detection in Public Bids

MICHELE A. BRANDÃO, Instituto Federal de Minas Gerais, Brazil and Universidade Federal de Minas Gerais, Brazil

ARTHUR P. G. REIS, BÁRBARA M. A. MENDES, CLARA A. BACHA DE ALMEIDA, GABRIEL P. OLIVEIRA, HENRIQUE HOTT, LARISSA D. GOMIDE, LUCAS L. COSTA, MARIANA O. SILVA, ANISIO LACERDA, and GISELE L. PAPPA, Universidade Federal de Minas Gerais, Brazil

The diversity of sources and formats of public bidding documents makes collecting, processing, and organizing such documents challenging from the point of view of data analysis. Thus, the development of approaches to deal with such data is relevant since the analysis of them allows to expand of the inclusion of people as they have more access to public decisions and expenditures, increase transparency in the public sector, and give citizens a greater sense of responsibility for having different points of view on the government's performance in meeting its public policy goals. In this context, we propose PLUS, a semi-automated pipeline for fraud detection in public bids. PLUS comprises a heuristic meta-classifier for bidding documents and a data quality module. Both modules present promising results after a proof of concept, reinforcing the relevance of PLUS for automating the bidding process investigation. Then, we present two applications of PLUS on real-world data: the construction of audit trails for fraud detection and a price database for overpricing detection. Such applications evidence a significant reduction of specialists' work searching for irregularities in public bids.

CCS Concepts: • **Applied computing → E-government;**

Additional Key Words and Phrases: Government data, public bids, fraud detection, audit trails

**ACM Reference format:**

Michele A. Brandão, Arthur P. G. Reis, Bárbara M. A. Mendes, Clara A. Bacha de Almeida, Gabriel P. Oliveira, Henrique Hott, Larissa D. Gomide, Lucas L. Costa, Mariana O. Silva, Anisio Lacerda, and Gisele L. Pappa. 2024. PLUS: A Semi-automated Pipeline for Fraud Detection in Public Bids. *Digit. Gov. Res. Pract.* 5, 1, Article 5 (March 2024), 16 pages.
https://doi.org/10.1145/3616396

## 1 INTRODUCTION

Traditionally, governments must document the acts and facts related to administrative processes and all the decision-making processes involved in exercising governmental activities. Examples of these documents are those related to public bids, which record public goods' purchase and sale transactions. In short, a bidding process

We would like to thank the partial support from MPMG through the project Analytical Capabilities.

Authors' addresses: M. A. Brandão, Street Vera Lúcia de Oliveira Andrade, 800 - Vila Esplanada, Ribeirão das Neves - MG, 33805-488, Ribeirão das Neves, Minas Gerais, Brazil and Avenue Presidente Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901; e-mail: michele.brandao@dcc.ufmg.br; A. P. G. Reis, B. M. A. Mendes, C. A. B. de Almeida, G. P. Oliveira, H. Hott, L. D. Gomide, L. L. Costa, M. O. Silva, A. Lacerda, and G. L. Pappa, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; e-mails: arthurpetrocchi@ufmg.br, barbaramit@ufmg.br, clarabacha@ufmg.br, gabrielpoliveira@dcc.ufmg.br, henriquehott@dcc.ufmg.br, larissa.gomide@dcc.ufmg.br, lucaslage@ufmg.br, mariana.santos@dcc.ufmg.br, anisio@dcc.ufmg.br, glpappa@dcc.ufmg.br.

2639-0175/2024/03-ART5 $15.00
https://doi.org/10.1145/3616396

is defined as a set of procedures through which governmental entities purchase or hire products and services. The importance of these documents to the national legal system is undeniable and one of the basic rules of Public Administration. However, despite its great relevance, the bidding process is vulnerable to irregular activities, which are prone to trigger political and economic problems. In this sense, a challenging task is how to process documents from bidding processes to identify such irregularities. Besides the huge number of documents, other data-related problems make this task more difficult, including the diversity of sources, lack of standardization, and data fragmentation.

According to Mistry and Jalal [19], **Information and Communications Technology** (**ICT**) can be utilized to provide public services more efficiently and disseminate information. Its results suggest that: as the use of ICT-related e-government increases, corruption decreases. Meanwhile, Ghedini Ralha and Sarmento Silva [7] studied cartels' formation in Brazil's public procurement processes. Identifying cartels is problematic because it requires the analysis of several public bidding processes, which usually exceed the scope of a single government agency. Besides, cartels can operate in various government departments, cities, and even states of the Federation, which demands sophisticated analyses of massive datasets. Identifying fraud in public bidding is a complex task, as it involves a large volume of non-standardized documents and requires manual inspection—which demands time and significant mobilization of human resources. In this context, we propose a semi-automated pipeline for fraud detection in public bids (PLUS), capable of auditing trails for fraud detection and identifying fraud in public purchases.

**Open Government Data** (**OGD**) is a global movement/philosophy whose goal is to make government data available to all.[1] The idea of OGD is to promote transparency, accountability, and value creation to make public institutions more transparent and accountable to citizens. In 2011, in Brazil, the Access to Information Law[2] established that government data should be available through open access, including government bids and bidding processes. This is the type of data we analyze in this article, with a focus on processes happening in the state of Minas Gerais, Brazil. Once the data is collected and stored, we apply the PLUS methodology, which is divided into two modules: data classification and quality assessment. The classification module categorizes bidding documents according to their types to allow for automatic information of relevant data, such as the product being bought or the companies that have bid for it. The quality assessment module analyzes the dataset, correcting any identified data problems. We emphasize that dealing with governmental data is challenging since most of the data available on the web are big, non-standardized, and dynamic [10].

Hence, the scientific objectives of this research are the definition of a consistent and reproducible pipeline that other works can appropriate to analyze long text documents (mainly in the public bid context but not only), the description of a heuristic meta-classifier that allows categorizing the data, and the proposal of customized indicators that helps to examine data quality. Regarding applications, the main objectives are the definition of audit trails and the methodology of overprice detection. Both can be generalized to different contexts.

The main contributions of this work are the following: a discussion about the existing research and a debate of the related work relevant to the particular topic addressed in this manuscript (Section 2); a framework entitled PLUS, a semi-automated **P**ipe**L**ine for Fraud Detection in P**U**blic Bid**S**, which comprises the classification and data quality modules (Section 3); a proof of concept of PLUS to validate the proposed pipeline in real-world government data (Section 4); the construction of audit trails, a sequence of steps to identify evidence of specific irregularities, for detecting frauds in public bids as first practical application of PLUS (Section 5); the discussion of overprice detection in the price database as a second practical application of PLUS (Section 6); a summary of the results and a relation with digital government and artificial intelligence in the public sector (Section 7); a brief discussion of the main challenges and limitations of the presented methodology that may open research questions to other works (Section 8); and a conclusion of the results and the study's contributions (Section 9).

[1] OGD: https://www.oecd.org/gov/digital-government/open-government-data.htm

[2] Law n° 12.527 http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm

## 2 RELATED WORK

This section presents related work of the use of government open data (Section 2.1) and the detection of fraud in the public sector (Section 2.2). Following, we discuss the differences and contributions of this work in comparison to other works in the literature (Section 2.3).

### 2.1 The Use of Government Open Data

The significant advances in ICT have profoundly impacted almost all areas of society. In particular, digital initiatives have redesigned the dynamics of governmental power, allowing for greater popular participation and transparency in public actions [1]. In this context, different research efforts have analyzed government institutions' advantages, challenges, and potential uses of ICTs [21, 23, 24, 26].

One of the main topics in the literature is the use of OGD [27]. OGD has datasets about government actions, expenditures, and investments distributed in an accessible and transparent manner to the public [29]. Hsieh et al. [9] highlight the importance of OGD by using historical data from public mediations to build a machine-learning model to guide users to solve their conflicts with a lower-cost solution. Furthermore, Joshi and Saha [11] propose a framework for automatic knowledge extraction from the federal regulatory code of the U.S. government and its representation through an ontology. They transform the data from the federal regulation code in open text format into a structure for companies and non-specialized government agencies to conduct searches on various topics and streamline their work to comply with the legislation.

OGD is usually made available through data portals by responsible government institutions. The number of OGD portals at the federal, state, and city levels has grown globally, along with the number of surveys that analyze the access to these portals and data by different users [16]. Many government agencies only focus on having an OGD portal that works and provides the data, disregarding the access of audiences with different needs and fundamental principles, such as its usability, usefulness, quality, and reuse of data [22].

Different evaluation methods are available in the literature to identify the strengths and weaknesses of government portals in multiple contexts. An example is the Berners-Lee model, Five-Star Linked Open Data, which assesses the maturity of OGD based on five categories: (1) Data is available on the web with an open license, (2) Available as machine-readable structured data, (3) Available in a non-proprietary format, (4) Published using open standards from the W3C, and (5) All of the above and links to other Linked Open Data to provide context [17]. Another example is the Window theory, proposed by Matheus and Janssen [16], which evaluates the transparency of portals based on 42 factors grouped into the following categories: data quality, system quality, organizational characteristics, and individual characteristics.

In Brazil, not many works on the evaluation and use of OGD portals are available. Matheus et al. [17] is one of the few. They use the Five Stars model to evaluate Brazilian portals at federal, state, and city levels. The authors emphasize that accessing data from such portals in other projects is difficult. Also, dos Santos Brito et al. [4] and de Oliveira and Silveira [3], after evaluating the use of the data available on the portals of different contexts, identify the challenges most researchers face with this data, namely: dataset quality, data access and format, multiple and decentralized data sources, and lack of standards for data publishing. Finally, Kawashita et al. [13] developed a questionnaire to conduct a usability analysis of OGD portals and validated it with data from 26 portals of Brazilian states. The authors list the main factors against the advancement of the use of portals, including the absence of an organizational culture favorable to open data and the fact that many public managers do not know what open data is or are not interested in increasing the level of social control over them [13].

### 2.2 Fraud Detection in the Public Sector

Fraud detection is a general problem in the public sector, and different approaches have tackled it [8, 20]. A common aspect of these approaches is that they have to deal with big data, which is a frequent characteristic of OGD. In particular, Handoko and Rosita [8] present a quantitative study that shows the importance of big data in fraud detection. Fraud detection approaches are commonly based on artificial intelligence techniques. For example, Mongwe and Malan [20] apply an unsupervised learning algorithm in government data for detecting fraud in financial statements of the public sector.

Specifically, in Brazil, there is a shortage of analyses using OGD due to existing barriers reported in works like [3]. Few works in this context include [6, 14, 15], and [28] that show how to use OGD for fraud detection and cartel formation in public tenders. Also, Lima et al. built a database to evaluate fraud detection models in the public service from the processing of the open text of the Brazilian Official Diary. Also, Gabardo and Lopes use social network analysis techniques to verify the formation of cartels among companies participating in bids in the civil construction area, and Lima et al. model the relationships between companies as a network (network/co-bidding network) and extract bidding patterns to calculate fraud indicators. Finally, Velasco et al. [28] describes a **decision support system** (**DSS**) to address the existing limitations in fraud analysis in the Brazilian public sector. The proposed DSS applies data mining techniques to extract rules and risk patterns from public procurement from several states. Their idea is to provide indicators for auditors of suspicious activities by companies and individuals involved. As a result of the DSS application, public authorities in Paraíba opened relevant investigations. As evidenced by the works cited, analyzing fraud in bids is a relevant topic, partly due to the known damage that the misuse of public funding can cause to society as a whole [14].

### 2.3 Our Work in the Face of the State-of-the-Art

The analysis of the related work reveals that research on OGD falls into three main categories: OGD data processing [11, 13], OGD quality evaluation [3, 4, 11, 13, 16, 17, 22], and applications that deal with OGD data [4, 6, 9, 11, 14, 20, 28]. In this context, the proposed pipeline—PLUS—fits into these three categories, as it processes the data in the classification module, analyses the quality in the data quality module, and uses the processed public data into two real applications, namely, audit trails and to build a reference price database.

Furthermore, the literature review shows that there are no works that propose a framework for extracting and making knowledge available from public bidding. In this sense, PLUS is an innovative pipeline as it makes available to the specialist, in a practical and semi-automated way, essential data for the analysis of bids that otherwise would not be possible, given the amount of information available. In addition, our pipeline allows the discovery of information and knowledge for fraud detection in public bids.

## 3 PLUS: A SEMI-AUTOMATED PIPELINE FOR FRAUD DETECTION IN PUBLIC BIDS

This section presents PLUS, a semi-automated **P**ipe**L**ine for Fraud Detection in P**U**blic Bid**S**. Figure 1 shows an overview of the main steps of the pipeline. Given a set of documents—here represented by public bids documents made available for open access by the government, a classification module receives the documents, preprocesses, and categorizes them according to their type (minute, public notice, and so on). Next, the information extracted from these documents is stored in an **Enterprise Data Warehouse** (**EDW**), whose priority is to enable quick data queries. Then, a data quality assessment module analyzes the quality of the built dataset, correcting any identified data problems. Finally, the dataset built with public bids can be used by different applications. In this work, we describe two of them: audit trails for fraud detection and a reference price database that helps identify fraud in public purchases. Both applications are discussed in Sections 5 and 6, and each step of the proposed pipeline is described next.

### 3.1 Classification Module

The first module of PLUS refers to a document classifier that categorizes bidding documents according to their types. Specifically, given a set of documents, the classifier will assign one of the different bidding document categories (e.g., public notice, minutes, homologation, among others) to make them easier to manage, search, filter, or analyze. In this section, we present the main steps of this module, from preprocessing the text extracted from the collected and filtered documents to developing the final classification model, as depicted in Figure 2.

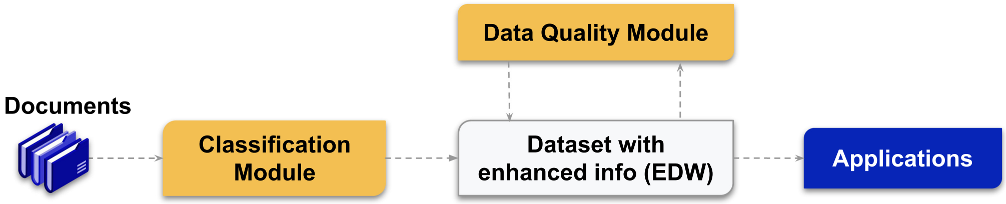


Fig. 1. PLUS: **P**ipe**L**ine p**U**blic bid**S**—such pipeline has three main modules, which are the data quality module, preprocessing module, and classification module.

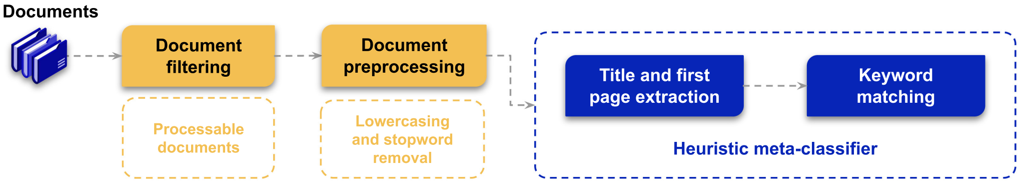

Documents
Document filtering
Processable documents
Document preprocessing
Lowercasing and stopword removal
Title and first page extraction
Keyword matching
Heuristic meta-classifier

Fig. 2. Classification module overview.

*3.1.1 Document Filtering.* Once collected, bidding documents go through a filtering process, separating them into non-processable (scanned/corrupted documents) and processable (documents from which it is possible to extract text directly). To do this separation, we use Python's PDFPlumber[3] library, which cannot extract text from scanned, corrupted documents or images. When no text is extracted from the document, it is considered unprocessed. Documents considered processable proceed to the next steps of the classification module.

*3.1.2 Document Preprocessing.* After filtering the documents, their text is preprocessed using a set of functions, which includes lower-casing and removal of proper names, e-mails, URLs, pronouns, adverbs, special characters, accents, stop-words, hours, number symbols, numbers, contracted and shortened words, single letters in the text and extra spaces. We also tokenize the text to provide appropriate input to the classifier.

*3.1.3 Heuristic Meta-classifier.* One of the main problems of classifying documents of public bids into types is that the number of possible classes is unknown. There is a fixed number of documents required by law to be made available for each bidding process, but there are processes with fewer and processes with a larger number of documents. For example, some institutions add more than 40 types of documents for each bid, including spreadsheets of price quotes, modified versions of the call, and so on. Hence, we started with a simple process to understand how rare were some classes and how difficult the task was. We developed a heuristic meta-classifier based on keywords. From the constant interactions with the bidding documents, we perceived structural patterns of the documents and keywords that proved suitable attributes for separating the most frequent classes of documents. Although the meta-classifier is simple, it is the first to categorize bidding documents into four meta-classes by only analyzing the title and first page of long text documents and searching for keywords. Algorithm 1 presents the main steps of the heuristic meta-classifier based on keywords, which analyzes the occurrence of these words in the title and content of each bidding document. Initially, for each document, the meta-classifier extracts the title and the first-page content and checks whether the keywords associated with each meta-class are present in the title or first-page content. If so, the document is assigned to its associated meta-class. Otherwise, the document is categorized as the "Others" meta-class. The definition of meta-classes considers types of documents assumed essential to a bidding process. From the empirical knowledge acquired

[3] PDFPlumber: https://github.com/jsvine/pdfplumber

by the analysis of the documents, we establish four meta-classes: Minutes, Public Notice (covers public notice documents and invitations sent in bids of the Invitation modality), Adjudication/Approval (covers documents of adjudication, approval, or that present both information in the same document) and Others (includes files belonging to other types of documents, e.g., errata, annexes, contracts, and descriptive memorials). Table 1 presents the keywords defined to identify each meta-class in the documents. The evaluation of the heuristic meta-classifier is presented in Section 4.2.

**ALGORITHM 1:** Heuristic Meta-classifier

```
Input: Documents of bidding processes in PDF
Output: Meta-class assigned to each document
begin
    for each PDF document of each city c do
        Extract the title and first-page content of the document;
        Declare countWordsTitle variable; // Occurrence of keyword in title, by meta-class
        Declare countWordsContent variable; // Occurrence of keyword in first-page content, by meta-class
        for each meta-class do
            Update countWordsTitle with the number of keywords that occurred in the title;
            Update countWordsContent with the number of keywords that occurred in the first-page content;
        if "Others" meta-class keywords exist then
            meta_class ← "Others"
        if "Adjudication/Approval" meta-class keywords exist then
            meta_class ← "Adjudication/Approval"
        Sort countWordsTitle in descending order;
        Sort countWordsContent in ascending order;
        if there is a keyword occurrence in the first-page content then
            meta_class ← meta-class associated
        else
            meta_class ← "Others"
return List of labelled documents by meta-class
```

## 3.2 Data Quality Module

The second module of PLUS is responsible for data quality analysis. The data quality area has multiple definitions, but the most accepted is that data quality is associated with a context, i.e., the data may be suitable for one scenario but not another [12]. Therefore, many works analyze quality in a specific domain [2]. Another definition concerns multiple dimensions identified by attributes representing particular data characteristics [25]. Thus, this module determines the data quality in a data warehouse built from big data of public bids. We use ***Great Expectations*** (**GE**) as our data quality tool.

GE is an *open-source* data quality tool that uses a mechanism similar to unit tests for data validation, where each validation is done by a module called *expectation*. Here, *expectations* are called indicators. GE provides several native indicators that perform generic data validations, such as checks for field types, ranges of values, and null records. In addition, GE offers the possibility of creating custom indicators, through which it is possible to implement specific business rules for each data warehouse table [5]. Considering the context of fraud detection in public bids, we choose GE as a data quality tool for three main reasons: (i) the customized indicators allow the implementation of data quality indicators for specific business problems; (ii) *Data Docs* generates a graphical interface containing the results of the executed indicators, which facilitate the analysis of the results by the user, and (iii) the *Profiler* makes a pre-analysis of the structure of the stored data, then shows an overview of the data along with some interesting indications of data checks to be implemented.

Table 1. Meta-classes and Associated Keywords

| Meta-class | Keywords | Meta-class | Keywords |
|---|---|---|---|
| Minutes | `minutes`, `public session` | Public notice | `invitation`, `notice` |
| Adjudication/ Approval | `adjudication`, `approval` | Others | `schedule`, `addition`,<br>`order of service`, `answer`,<br>`extract`, `official diary`, `warning of`,<br>`rectification`, `administrative contract` |

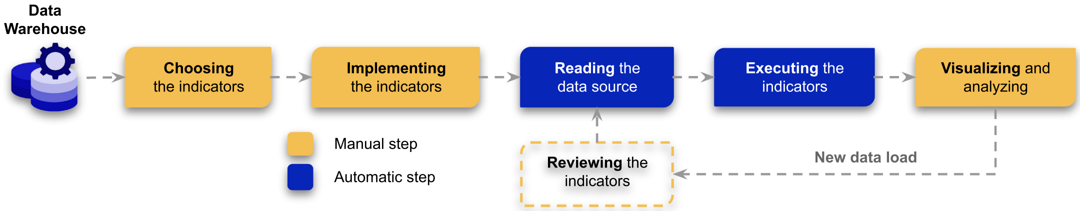


Fig. 3. Methodology of the data quality module.

Figure 3 presents the data quality analysis flow to correct potential problems and that consists of five main steps, ranging from choosing specific indicators for each data source to viewing and analyzing the results by analysts. It also includes an optional sixth step, which is the review of the indicators. A good part of this flow is iterative and automatic, as the last four steps are executed whenever a new load is performed on the data source, thus covering the new data added. Each step of this pipeline is detailed next.

*3.2.1 Choosing the Indicators.* After selecting a database table to be analyzed, this first step consists of a manual inspection of the table structure and content to define the quality indicators that will be implemented. The person conducting this stage must have technical and business knowledge so that the indicators chosen are adequate. This is the best time to use GE's *Profiler* component, which performs the pre-analysis of the data and verifies which native indicators[4] are more suitable to be executed on the analyzed data. These native indicators are standard and generic rules implemented internally within the tool, including validating the data domain and verifying that the data follows a regular expression or a range of values. On the other hand, custom indicators validate business rules specific to the data domain. For the context considered in our work, we recommend using native and customized indicators. After extensive analyses, which are better discussed in Section 4.3, we list all the native and custom indicators recommended for our context in Table 2.

*3.2.2 Implementing the Indicators.* This step refers to implementing the indicators chosen through the *GE* tool. This can be done using, for example, Pandas,[5] Spark,[6] or SQLAlchemy[7] data frames. The definition of the data frame type depends on the system architecture and the volume of data that will be analyzed. For instance, Spark is better for dealing with big data, whereas Pandas can be used to handle tables with few records. We use the data frames to store the data that will be given as input to the next step.

*3.2.3 Reading the Data Source.* In this step, GE reads each table of the data warehouse. It is worth noting that it is necessary to read the entire content of the table for the indicators to be executed. As mentioned before, such data is stored in data frames on which the indicators will be executed.

[4] List of native indicators existing at GE https://greatexpectations.io/expectations
[5] Pandas: https://pandas.pydata.org/
[6] Apache Spark: https://spark.apache.org/
[7] SQLAlchemy: https://www.sqlalchemy.org/

Table 2. GE Indicators Recommended for Fraud Detection in Public Bids

| | **Indicator (*expectation*)** | **Description** |
|---|---|---|
| **Native** | expect_column_values_to_not_be_null | Column values must be non-null |
| | expect_column_values_to_be_unique | There must be no duplicate values in the column |
| | expect_column_min_to_be_between | The smallest column value must be within the range [min, max] |
| | expect_column_values_to_be_in_type_list | Column values must be of the specified type |
| | expect_column_values_to_be_in_set | Column values must belong to a set of values |
| | expect_column_values_to_be_between | Column values must be in the range [min, max] |
| | expect_column_values_to_match_regex | Column values must follow a certain regular expression |
| **Customized** | expect_value_less_revenue | Bids must have a value less than or equal to the entity's income (city or state) |
| | expect_table_fato_licitacao_to_have_guests_if_invite | Invitation-only bids must have invited bidders |
| | expect_dates_to_match_across_tables | Reported dates must be in valid chronological order |
| | expect_only_one_year_of_activity | Bids must have only a single year of activity |
| | expect_sum_of_item_values_to_match_fato_licitacao | The sum of the values of the items must match the value of the bid |

*3.2.4 Executing the Indicators.* After reading the table, the implemented indicators are executed and the results are presented in an interactive graphical interface generated by the *Data Docs* component.

*3.2.5 Visualizing and Analyzing.* The last step corresponds to the visualization and analysis of the results of the indicators in the interactive graphical interface. From this analysis, it is possible to verify cases that indicate errors in the load and/or in the data format, as well as to take the necessary actions for the correction.

*3.2.6 Reviewing the Indicators (Optional).* If necessary, this step can be performed right after the new data loads in the evaluated tables and comprises the reassessment and implementation of new indicators according to needs and demands that may arise.

## 4 EVALUATION OF PLUS

In this section, we present an evaluation of PLUS to validate the proposed pipeline in real-world government data. Section 4.1 details the dataset of public bids built from OAD. Next, Section 4.2 presents the results for the classification model, and Section 4.3 shows the results for data quality.

### 4.1 Data

As mentioned before, one of the major difficulties when dealing with public bids is the massive volume of data. Brazil is a country of continental dimensions comprising 27 federated units and 5,570 cities. Each of these entities has bidding processes to satisfy its demands. In this work, to carry out a proof-of-concept of the proposed pipeline and its subsequent use in two applications (audit trails and reference price database), we consider data from Minas Gerais, the second most populous Brazilian state and the one with the largest number of cities (853). In particular, we use data from the Prosecution Service of the Brazilian state of Minas Gerais (in Portuguese, *Ministério Público do Estado de Minas Gerais* or simply MPMG).

For bids, we consider public data from the Computerized System of Municipal Accounts (SICOM,[8] city level) and the Transparency Portal of the Government of Minas Gerais[9] (state level). We also collected public bids from the transparency portals of 18 cities in Minas Gerais, obtaining documents from different classes to evaluate the proposed heuristic meta-classifier. Regarding the bidders (i.e., companies), we consider data from the Federal Data Processing Service (SERPRO) to obtain stakeholder information and the Registry of Disreputable and Suspended Companies (CEIS) to verify possible sanctions. All this information is aggregated in the MPMG data infrastructure and then processed by PLUS.

[8] SICOM: https://portalsicom1.tce.mg.gov.br/

[9] Transparency Portal of Minas Gerais: https://www.transparencia.mg.gov.br/compras-e-patrimonio/compras-e-contratos

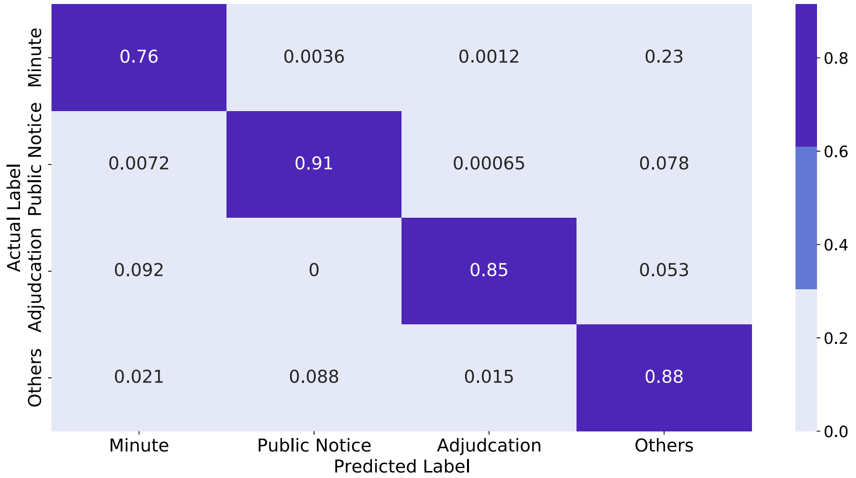


Fig. 4. Confusion matrix of the heuristic meta-classifier.

Overall, we consider 378,137 bids (14,565 at the state level and 363,572 from the city level) covering the period from 2014 to 2021. The bids are divided into 17 modalities, including competition, invitation, contest, and trading. In addition, our dataset contains information from 103,858 different bidding companies that participated in such processes. This data includes relevant information for detecting spurious relationships (e.g., different companies with exactly the same address and partners) and restrictions (inactive registration, single bidder, etc.).

## 4.2 Classification Module Results

First, we present the evaluation of the proposed heuristic meta-classifier, which in general has a good performance, as shown in Figure 4. Each document considered in the classification module was manually labeled according to the actual meta-class. As the actual meta-class of most documents is present in the file title, the manual labeling process is performed by checking the title of each document and the meta-class labeled by the meta-classifier. In this process, both pieces of information are sufficient for reliable labeling; therefore, calculating the degree of divergence between the participating labelers is waived.

In total, we evaluated 6,337 documents from 18 cities in Minas Gerais. We generate a confusion matrix from the documents labeled with the actual meta-class to visualize the meta-classifier's accuracy, as depicted in Figure 4. Overall, the results indicate that the heuristic meta-classifier performed well, especially for the *Public Notice* meta-class, with a hit rate above 90%. The second meta-class that presented a high hit rate is *Others*. According to the confusion matrix, the classifier confused this meta-class with the meta-class "Minutes". Such errors demonstrate that establishing more specific classes of documents can be a promising alternative to mitigate the confusion of the proposed classifier.

## 4.3 Data Quality Module Results

In this section, the main results of the quality indicators of GE in the real data of the public bids are presented. In this analysis, the five main tables that gather bidding data were considered: (i) general bidding information; (ii) bidders entitled to participate in bidding processes; (iii) bidders approved as winners in bids; (iv) bid items; and (v) bidding commissions (bodies set up to act in bids).

For each table, specific native indicators were chosen according to their context. In addition, customized indicators were implemented according to pre-defined business rules. Table 3 presents the number of successes and failures in the indicators for each analyzed table, as well as the total number of implemented indicators.

Table 3. General Statistics of Performance of Indicators at GE

| Table | Success | Fails | Total |
|---|---|---|---|
| Bid | 88 (68.22%) | 41 (31.78%) | 129 |
| Licensed Bids | 25 (71.43%) | 10 (28.57%) | 35 |
| Homologated Bids | 32 (43.24%) | 42 (56.76%) | 74 |
| Bids Items | 54 (65.06%) | 29 (34.94%) | 83 |
| Commission | 47 (83.93%) | 9 (16.07%) | 56 |

Table 4. Number of Failures Captured by Quality Indicators

| Error/Table | Bid | Licensed Bid | Homologated Bid | Bid Item | Commission |
|---|---|---|---|---|---|
| Null values | 18 | 1 | 15 | 6 | 0 |
| Values out of expectations | 12 | 3 | 10 | 5 | 1 |
| Incoherent data type | 8 | 2 | 3 | 8 | 7 |
| Duplicate values | 0 | 1 | 3 | 1 | 0 |
| Others | 3 | 3 | 11 | 9 | 1 |
| **Total** | **41** | **10** | **42** | **29** | **9** |

Table 5. Custom Indicator to Check how Many Records do not Respect the Chronological Order of Bid Dates (Date 1 ≤ Date 2)

| Date 1 | Date 2 | Number of records | % |
|---|---|---|---|
| Public notice date | Publication date of the notice | 7,672 | 2.03 |
| Public notice date | Publication date on the venue | 8,728 | 2.31 |
| Publication date of the notice | Expected date of receipt of documentation | 2,186 | 0.58 |

Table 4 presents the most common errors detected in the analyzed tables. One of the most frequent errors is the presence of null values in columns that should be filled according to business rules. For example, it is not expected that fields containing the year in which bids were exercised have null values. Other common errors include the presence of out-of-standard values and/or expected range and incoherent data type. In addition, some tables have duplicate records, and this type of error occurs for two reasons: (i) data load errors and (ii) the data warehouse used has the limitation of not supporting integrity constraints to avoid this duplicity. Since GE detects these duplicate records, it serves to mitigate a data warehouse limitation.

In addition, GE's custom indicators allow us to verify more complex business rules that cannot be verified by native indicators. An implemented business rule checks the chronological order of date fields present in the bidding records, as the dates must respect the order of the bidding process. For example, the date of the bidding notice must be prior to its publication, as the preparation of the notice is the first step of the process, and the receipt of the documentation only happens after the notice is published. Table 5 shows the number of cases that do not respect this chronological order. Since there are few records in this situation, there may have been a problem with the imputation or data loading. This result reinforces the need for a thorough analysis of the extraction, treatment, and loading processes by the user.

## 5 AUDIT TRAILS FOR FRAUD DETECTION

In this section, we present the proposal and construction of audit trails for detecting fraud in public bids as the first practical application of PLUS. Section 5.1 presents the definition of the audit trails. Then, Section 5.2

presents two types of trails to identify distinct irregularities in bids: bidders with common links and bidders with restrictions. Finally, Section 5.3 shows the trails to detect specific inconsistencies in public bids, such as bidders with shared stakeholders, bidders with inactive registration, and so on.

## 5.1 Definition

Fraud detection in bids is complex to detect since these frauds are not punctual and isolated. Instead, they involve direct, indirect, and even temporal interactions between the entities involved. In general, checking irregularities in public bids is a real-time manual process. That is, the entire bidding process is monitored—from the publication of the notice until the contract is signed and executed—by people. As the volume of bids increases, automatic forms of monitoring, analyzing, and cross-referencing data in search of law violations have become very useful to assist this process. The human role in the process is still fundamental, as people know the laws and have the ability to interpret them. However, combining human monitoring with computer systems helps to optimize and improve fraud detection.

In this context, an audit trail is a sequence of steps necessary to identify evidence of specific irregularities in public bids. The audit trail helps specialists detect fraud in a large volume of data, as it allows for selecting the data of interest from the bidding databases. An example of an audit trail is to identify bids with different bidders that have at least one partner in common.

## 5.2 PLUS for Fraud Detection

To assist in the fraud detection process, we build the audit trails according to business rules provided by MPMG specialists. We use the specialists' knowledge to divide such rules into two groups: (i) rules based on common links between bidders and (ii) rules based on bidder restrictions. Therefore, we model two types of audit trails, which depend on one or more user-defined parameters to be executed. Thus, we build generic execution flows that can be easily adapted to implement the business rules belonging to both groups. Finally, the output of both types of trails is a list of public bids whose bidders violate the business rules.

*5.2.1 Bids Containing Bidders with Common Links.* This audit trail focuses on searching registration data in common among competing bidders (i.e., companies) in the same bidding process. Examples include companies with the same e-mail address, phone number, or shared stakeholders, indicating a possible connection between the companies. Such common links may represent fraud alerts, which require further manual investigation by specialists. Therefore, this trail combines information from all bidders in the same bidding, looking for information that can generate signs of links between companies.

*5.2.2 Bids Containing Bidders with Restrictions.* In contrast, this audit trail considers bidders individually and aims at finding companies with irregular registration or restrictions with government agencies that make them ineligible to bid. Examples include companies with inactive CNPJ (company registration number), companies that are the only bidders in a public bid, and companies with legal restrictions. To do so, we check the company's registration status in the data sources already mentioned, such as CEIS and SERPRO.

## 5.3 Results

After defining the trails, it is possible to apply them in the construction of real-world audit trails, whose objective is to raise alerts of potential fraud in public bids according to the previously established business rules. We build 11 distinct trails: four derived from the trail of bids containing bidders with a common link and seven derived from the trail of bids containing bidders with restrictions. Table 6 presents the list of all audit trails and the trails from which they were derived. As mentioned in the previous section, the output of each trail is a list of bids with a fraud alert. For each bid, we include information such as the number of bidders involved in the common link or restriction, the registration number of these bidders, and what is the common link (e.g., e-mail address, phone

Table 6. Audit Trails Built from the Two Proposed Types of Trails

| Code | Type | Audit trail |
|---|---|---|
| T01 | Common links | Bids containing bidders with common stakeholders |
| T02 | Common links | Bids containing bidders with common e-mails |
| T03 | Common links | Bids containing bidders with common phone numbers |
| T04 | Common links | Bids containing bidders with common address |
| T05 | Restriction | Bids with a single bidder |
| T06 | Restriction | Bids containing bidders with inactive registration |
| T07 | Restriction | Bids containing bidders who are frequent winners |
| T08 | Restriction | Bids containing bidders who are frequent losers |
| T09 | Restriction | Bids containing bidders before the start of activities |
| T10 | Restriction | Bids containing bidders with legal restrictions |
| T11 | Restriction | Bids containing bidders whose stakeholders are civil servants |

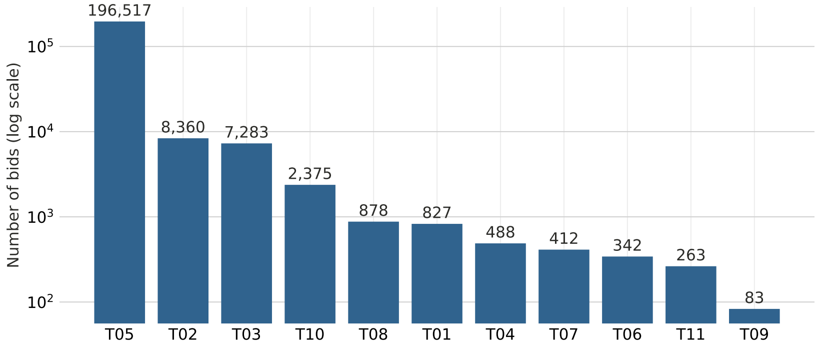


Fig. 5. Number of bids with irregularity alerts by audit trail (log scale).

number) for the trails from the first type. All results are used by MPMG specialists in the manual investigation stage of the fraud detection process.

Figure 5 shows the number of bids with fraud alerts for each audit trail. Overall, 211,459 bids have some alert (55.9% of the total), most of them coming from the trail T05 (bids with a single bidder), with 196,517 bids. Despite the high number, such a result is reasonable since many bids from small cities do not present great competition between companies. When excluding the trail T05, only 15,430 bids present some fraud alerts (note that bids can have alerts for more than one trail). This number represents a small fraction of the total (4.1%), which is in line with the expected since the number of bids with evidence of fraud is much lower than the number of bids that do not. Trails T02 (bidders with common e-mails) and T03 (bidders with common phone numbers) comprise the majority of such bids, but a brief manual inspection reveals that many companies have the e-mails and phone numbers of their accounting offices[10] in their registration. Therefore, audit trails are very useful in the fraud detection process but they must not be used alone. Instead, they are an important mechanism for specialists to filter and prioritize bids to be analyzed in the manual investigation stage of the fraud detection process.

## 6 REFERENCE PRICE DATABASE FOR OVERPRICE DETECTION

This section details the definition of the Reference Price Database (Section 6.1), the application of the steps of PLUS (Section 6.2), and the results (Section 6.3).

[10] In Brazil, it is not illegal for an independent accounting office to represent several companies.

### 6.1 Definition

A Reference Price Database is an essential tool for bidding processes in national territory, helping to consult or calculate reference values for contracting. This tool is essential for the public authority to have speed in preparing public notices and helps bring more security, flexibility, and transparency to public accounts in general. Specifically, it can help to find reference prices easily throughout the national territory, generating cost savings, as one can verify whether or not the proposed prices in the bidding process match the reality of the market. The Reference Price Database receives as input the information on the bidding items extracted from the bidding minutes present in the curated datasets resulting from PLUS.

The data source for the Reference Price Database developed is the Computerized System of Municipal Accounts (SICOM), which contains general information on the bidding objects. This information is usually present in the price registration minutes of the bidding processes. The Price Registration Minute is a binding and mandatory document, i.e., it is a document that generates the expectation of contracting, where prices, suppliers, supply conditions and all participating bodies are recorded.

### 6.2 PLUS and the Reference Price Database

Recall that PLUS Classification Module classifies documents according to their respective category. For this application, only documents referring to *price registration minutes* are considered in the following steps of PLUS. Hence, in the Data Quality module, the tables referring to the minutes go through the data quality flow to deal with any data problems. Finally, we create the final table that feeds the Reference Price Database with the information properly filtered and processed. Specifically, this table gathers information about the items and the corresponding bids (e.g., city, year, month). From constructing the final table until its use by the Reference Price Database, we also apply a clustering algorithm to group similar bidding items to define price reference groups.

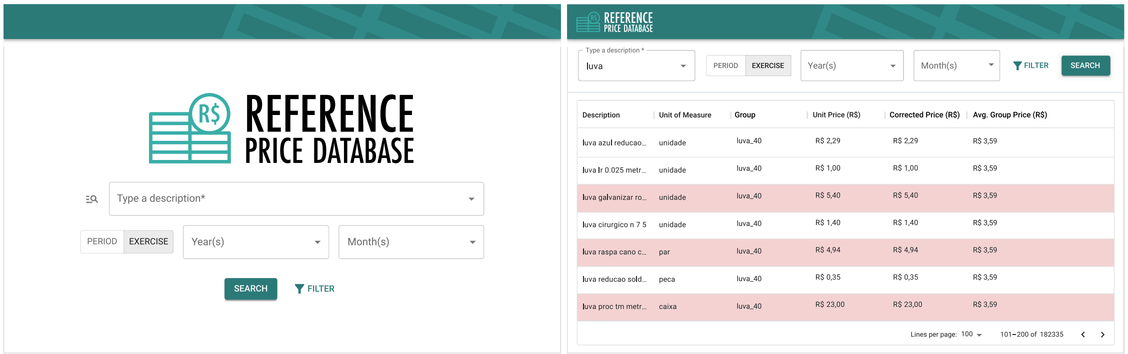
### 6.3 Results

After applying the pipeline proposed, we use the final table generated as input to group the bidding items according to their textual descriptions. Thus, one can implement overprice detection strategies from the grouping of items. Specifically, with aggregated items, one can compare the unit price of a bidding item with the average price of the group it belongs to. So, an overprice alert can be raised if the item in question's unit price is above its group's average price. In summary, using a statistics-based strategy, one measures how many standard deviations the record is from the mean of the data for each observation. In our context, the method checks how many standard deviations the unit price of a given item is from the average price of the group to which it belongs. For example, following a common rule, if the standard deviation is greater than C, where C is usually set to 3, the observation is marked as an outlier, i.e., as an overpriced item.

Implementing the first version of the Reference Price Database relies mainly on the presentation of bidding items, price statistics for groups of items obtained after items are clustered, and the detailing of prices and visualizations for groups of items. Figure 6 (left) shows the main screen of the first version of the tool, where the list and information of the searched item are presented to the user, in a pageable table format, according to the description and the selected filters. Figure 6 (right) shows, in a visual way, how the overprice alert is presented to the user. In the example, three items from the glove_40 group would be overpriced, where the unit price is much higher than the group's average price. In such cases, rows representing the items are highlighted in red.

## 7 DISCUSSION ABOUT ARTIFICIAL INTELLIGENCE IN THE PUBLIC SECTOR

In this article, we show that given as input a set of public documents in PDF format, PLUS can generate as output a data warehouse that facilitates the extraction of information from the text documents. The pipeline output can benefit different contexts. For example, the input document can be a public expenditure document to analyze expenditure overpricing. It is relevant to compare this work with that of Janssen et al. [10], which



Type a description * luva

| Description | Unit of Measure | Group | Unit Price (R$) | Corrected Price (R$) | Avg. Group Price (R$) |
|---|---|---|---|---|---|
| luva azul reducao... | unidade | luva_40 | R$ 2,29 | R$ 2,29 | R$ 3,59 |
| luva lr 0.025 metr... | unidade | luva_40 | R$ 1,00 | R$ 1,00 | R$ 3,59 |
| luva galvanizar ro... | unidade | luva_40 | R$ 5,40 | R$ 5,40 | R$ 3,59 |
| luva cirurgico n 7 5 | unidade | luva_40 | R$ 1,40 | R$ 1,40 | R$ 3,59 |
| luva raspa cano c... | par | luva_40 | R$ 4,94 | R$ 4,94 | R$ 3,59 |
| luva reducao sold... | peca | luva_40 | R$ 0,35 | R$ 0,35 | R$ 3,59 |
| luva proc tm metr... | caixa | luva_40 | R$ 23,00 | R$ 23,00 | R$ 3,59 |

Lines per page: 100 101–200 of 182335

Fig. 6. Screenshot of the current version of the Reference Price Database. On the left is the home screen, and on the right is the resulting table after a query, showing overpricing alerts (in red) for the searched bidding items.

proposes a framework for data governance in trustworthy artificial intelligence systems. The aim of their system is to organize data to other systems that are similar to PLUS, whose goal is to generate data to be used for fraud detection in the public sector. As the data warehouse generated by PLUS has texts extracted from PDF documents and categorized in meta-classes, one can use these texts as input to a machine learning algorithm to identify further classes for the public bid documents.

Furthermore, Mehr et al. [18] claim that data-prepared and tread carefully with privacy should be one of the six strategies for applying artificial intelligence in government offices. By considering all the theories, methodologies, and practicals described in this article with the use of public bid documents, this work represents a step forward in this one strategy since all these contributions can be generalized to other OGD.

## 8 CHALLENGES AND OPEN RESEARCH OPPORTUNITIES

PLUS has some limitations that may be the focus of improvements in future works. These limitations are related to the challenge of dealing with many different documents, often made available without standardization on government portals. The main challenges and open opportunities are listed as follows.

**Assessment and discussions are based only on Brazilian data.** All the analyses and discussions presented in this work are related to Brazilian public bids. Indeed, some steps of PLUS are based on our knowledge of these documents. However, we believe that our proposed pipeline can be easily applied to other government data, as long as they follow similar data structures.

**PLUS is semi-automated.** The proposed pipeline has steps that require specialists' interference. Although this is important because we do not present a black-box pipeline, some situations may require a completely automatized pipeline. An example is when the context of data usage does not need specialists to make the necessary decisions.

**Meta-class *Others*.** As bidding documents are significantly distinct, in general, it was challenging to categorize them into very specific classes. Therefore, most of the documents were grouped in the Others meta-class. Thus, works considering all bidding documents or related bids need to explore files in this meta-class to understand them better.

## 9 CONCLUSION

In this article, we presented PLUS, a semi-automated **P**ipe**L**ine for fraud detection in p**U**blic bid**S**. Such a pipeline has two main modules: the classification module and the data quality module. The first referred to a document classifier whose main objective is categorizing bidding documents according to their types. The second one stated data quality analysis using the *GE* tool. PLUS receives as input a set of documents that are public bids and passes through the modules. A data warehouse stores these documents, and two real applications have used the output of PLUS for fraud detection tasks.

In particular, the classification module has three main steps: (i) document filtering, (ii) document preprocessing, and (iii) heuristic meta-classifier. We defined four meta-classes as bidding documents' categories: Minute, Public Notice, Adjudication/Approval, and Others. Next, based on keywords and document structure, the heuristic meta-classifier classifies each document into one of the available meta-classes. Overall, such a module presented a good performance, with more than a 90% hit rate.

On the other hand, the data quality module has six main steps: (i) choosing the indicators, (ii) implementing the indicators, (iii) reading the data source, (iv) executing the indicators, (v) visualizing and analyzing, and (vi) reviewing the indicators (optional). In this module, we have chosen the GE tool because it has graphical interface components that help visualize the results and allow a more significant amount of validations. GE also enables the creation of custom indicators verifying complex business rules not checked by native indicators. In other words, these custom indicators allow the implementation of specific validations for the data being analyzed. It is worth noting that GE helps to identify records that have problems caused by the impossibility of implementing data integrity restrictions, and thus, it works to mitigate this limitation of the data warehouse.

We then presented two applications for fraud detection using public bids and PLUS: the construction of audit trails for detecting fraud and a reference price database for overpricing detection. The results of the audit trails evidence some bids with irregularity alerts, which specialists can prioritize in the posterior investigation stage of the fraud detection process. Regarding the price database, the first version allowed the presentation of bidding items and price statistics for groups of items.

In sum, the main methodological contributions of this article include the classification and data quality modules of PLUS and the presentation of how PLUS can be applied to different scenarios. Finally, as the output of PLUS is a data warehouse, the practical contributions are the description of how to apply such output to define audit trails for detecting fraud and to create a reference price database for overpricing detection.

In future work, we plan to improve the classification module by considering more classes in the meta-classifier. Also, as the proposed pipeline is data-sensitive, we plan to analyze more aspects of the data quality module to improve the public bids data. Then, we project to insert functionalities at the reference price database to improve the fraud detection alarm task.

## REFERENCES

[1] Miguel Arana-Catania, Felix-Anselm van Lier, Rob Procter, Nataliya Tkachenko, Yulan He, Arkaitz Zubiaga, and Maria Liakata. 2021. Citizen participation and machine learning for a better democracy. *Digital Government: Research and Practice* 2, 3 (2021), 27:1–27:22. DOI: https://doi.org/10.1145/3452118

[2] Corinna Cichy and Stefan Rass. 2019. An overview of data quality frameworks. *IEEE Access* 7 (2019), 24634–24648.

[3] Emilio Feliciano de Oliveira and Milene Selbach Silveira. 2018. Open government data in Brazil a systematic review of its uses and issues. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, DG.O 2018*. Marijn Janssen, Soon Ae Chun, and Vishanth Weerakkody (Eds.), ACM, 60:1–60:9. DOI: https://doi.org/10.1145/3209281.3209339

[4] Kellyton dos Santos Brito, Marcos Antônio da Silva Costa, Vinicius Cardoso Garcia, and Silvio Romero de Lemos Meira. 2014. Experiences integrating heterogeneous government open data sources to deliver services and promote transparency in Brazil. In *Proceedings of the IEEE 38th Annual Computer Software and Applications Conference, (COMPSAC'14)*. IEEE Computer Society, 606–607. DOI: https://doi.org/10.1109/COMPSAC.2014.87

[5] Harald Foidl, Michael Felderer, and Rudolf Ramler. 2022. Data smells: Categories, causes and consequences, and detection of suspicious data in AI-Based systems. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI (CAIN'22)*, Association for Computing Machinery, Pittsburgh, Pennsylvania, 229–239. DOI: https://doi.org/10.1145/3522664.3528590

[6] Ademir Cristiano Gabardo and Heitor Silvério Lopes. 2014. Using social network analysis to unveil cartels in public bids. In *Proceedings of the 2014 European Network Intelligence Conference, ENIC 2014*. IEEE Computer Society, 17–21. DOI: https://doi.org/10.1109/ENIC.2014.11

[7] Célia Ghedini Ralha and Carlos Vinícius Sarmento Silva. 2012. A multi-agent data mining system for cartel detection in Brazilian government procurement. *Expert Systems with Applications* 39, 14 (2012), 11642–11656. DOI: https://doi.org/10.1016/j.eswa.2012.04.037

[8] Bambang Leo Handoko and Ameliya Rosita. 2022. The effect of skepticism, big data analytics to financial fraud detection moderated by forensic accounting. In *Proceedings of the 2022 6th International Conference on E-Commerce, E-Business and E-Government*. 59–66.

[9] Hsun-Ping Hsieh, JiaWei Jiang, Tzu-Hsin Yang, Renfen Hu, and Cheng-Lin Wu. 2021. Predicting the success of mediation requests using case properties and textual information for reducing the burden on the court. *Digital Government: Research and Practice* 2, 4 (2021), 30:1–30:18. DOI: https://doi.org/10.1145/3469233

[10] Marijn Janssen, Paul Brous, Elsa Estevez, Luis S. Barbosa, and Tomasz Janowski. 2020. Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly* 37, 3 (2020), 101493.

[11] Karuna Pande Joshi and Srishty Saha. 2020. A semantically rich framework for knowledge representation of code of federal regulations. *Digital Government: Research and Practice* 1, 3 (2020), 21:1–21:17. DOI: https://doi.org/10.1145/3425192

[12] Clovis S. Junior and Carina F. Dorneles. 2021. Avaliação de Dimensões de Qualidade de Dados para o Agronegócio. In *2021: Proceedings of the 36th Brazilian Symposium on Databases, (SBBD'21)*. SBC, 283–288. DOI: https://doi.org/10.5753/sbbd.2021.17886

[13] Ilka Kawashita, Ana Alice Baptista, and Delfina Soares. 2022. Open government data use in the Brazilian states and federal district public administrations. *Data* 7, 1 (2022), 1–18. DOI: https://doi.org/10.3390/data7010005

[14] Marcos Lima, Roberta Silva, Felipe Lopes de Souza Mendes, Leonardo Rebouças de Carvalho, Aletéia P. F. Araújo, and Flavio de Barros Vidal. 2020. Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020*. Trevor Cohn, Yulan He, and Yang Liu (Eds.), Findings of ACL, Vol. EMNLP 2020, Association for Computational Linguistics, 1580–1588. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.143

[15] Marcos S. Lyra, António Curado, Bruno Damásio, Fernando Bação, and Flávio L. Pinheiro. 2021. Characterization of the firm–firm public procurement co-bidding network from the State of Ceará (Brazil) municipalities. *Applied Network Science* 6, 1 (2021), 77. DOI: https://doi.org/10.1007/s41109-021-00418-y

[16] Ricardo Matheus and Marijn Janssen. 2020. A systematic literature study to unravel transparency enabled by open government data: The window theory. *Public Performance and Management Review* 43, 3 (2020), 503–534. DOI: https://doi.org/10.1080/15309576.2019.1691025

[17] Ricardo Matheus, Manuella Maia Ribeiro, and José Carlos Vaz. 2012. New perspectives for electronic government in Brazil: The adoption of open government data in national and subnational governments of Brazil. In *Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance, (ICEGOV '12)*. David Ferriero, Theresa A. Pardo, and Haiyan Qian (Eds.), ACM, 22–29. DOI: https://doi.org/10.1145/2463728.2463734

[18] Hila Mehr, H. Ash, and D. Fellow. 2017. *Artificial Intelligence for Citizen Services and Government*. Ash Center, Harvard Kennedy School.

[19] Jamshed J. Mistry and Abu Jalal. 2012. An empirical analysis of the relationship between e-government and corruption. *International Journal of Digital Accounting Research* 12 (2012), 145–176.

[20] Wilson Tsakane Mongwe and Katherine M. Malan. 2020. The efficacy of financial ratios for fraud detection using self organising maps. In *Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence, (SSCI'20)*. IEEE, 1100–1106. DOI: https://doi.org/10.1109/SSCI47803.2020.9308602

[21] Sarah Moore. 2019. Digital government, public participation and service transformation: The impact of virtual courts. *Policy and Politics* 47, 3 (2019), 495–509. DOI: https://doi.org/10.1332/030557319X15586039367509

[22] Anastasija Nikiforova and Keegan McBride. 2021. Open government data portal usability: A user-centred usability analysis of 41 open government data portals. *Telematics and Informatics* 58 (2021), 101539. DOI: https://doi.org/10.1016/j.tele.2020.101539

[23] Samuli Pekkola, Maija Ylinen, and Nicholas B. Mavengere. 2022. Consortium of municipalities co-tailoring a governmental e-Service platform: What could go wrong? *Digital Government: Research and Practice* 3, 1 (2022), 6:1–6:16. DOI: https://doi.org/10.1145/3511889

[24] Gabriel Puron-Cid, Dolores E. Luna, Sergio Picazo-Vela, J. Ramón Gil-Garcia, Rodrigo Sandoval-Almazan, and Luis F. Luna-Reyes. 2022. Improving the assessment of digital services in government websites: Evidence from the Mexican State government portals ranking. *Government Information Quarterly* 39, 1 (2022), 101589. DOI: https://doi.org/10.1016/j.giq.2021.101589

[25] Monica Scannapieco and Tiziana Catarci. 2002. Data quality under a computer science perspective. *Journal of The ACM - JACM* 2 (2002), 1–12.

[26] Hans Jochen Scholl. 2020. Digital government: Looking back and ahead on a fascinating domain of research and practice. *Digital Government: Research and Practice* 1, 1 (2020), 7:1–7:12. DOI: https://doi.org/10.1145/3352682

[27] Kuang-Ting Tai. 2021. Open government research over a decade: A systematic review. *Government Information Quarterly* 38, 2 (2021), 101566. DOI: https://doi.org/10.1016/j.giq.2021.101566

[28] Rafael B. Velasco, Igor Carpanese, Ruben Interian, Octavio C. G. Paulo Neto, and Celso C. Ribeiro. 2021. A decision support system for fraud detection in public procurement. *International Transactions in Operational Research* 28, 1 (2021), 27–47.

[29] Bernd W. Wirtz, Jan C. Weyerer, and Michael Rösch. 2018. Citizen and open government: An empirical analysis of antecedents of open government data. *International Journal of Public Administration* 41, 4 (2018), 308–320. DOI: https://doi.org/10.1080/01900692.2016.1263659

Received 1 August 2022; revised 24 March 2023; accepted 6 June 2023