

VigIA: prioritizing public procurement oversight with machine learning models and risk indices

Andrés Salazar¹ , Juan F. Pérez²  and Jorge Gallego³ 

¹Department of Economics, Universidad del Rosario, Bogotá, Colombia.

²Department of Industrial Engineering, Universidad de los Andes, Bogotá, Colombia

³Office of Evaluation and Oversight, Inter-American Development Bank, Washington, DC, USA

Corresponding author: Juan F. Pérez; Email: jf.perez33@uniandes.edu.co

Received: 19 August 2023; **Revised:** 30 August 2024; **Accepted:** 16 October 2024



Keywords: GovTech solutions; machine Learning; Public procurement

Abstract

Public procurement is a fundamental aspect of public administration. Its vast size makes its oversight and control very challenging, especially in countries where resources for these activities are limited. To support decisions and operations at public procurement oversight agencies, we developed and delivered VigIA, a data-based tool with two main components: (i) machine learning models to detect inefficiencies measured as cost overruns and delivery delays, and (ii) risk indices to detect irregularities in the procurement process. These two components cover complementary aspects of the procurement process, considering both active and passive waste, and help the oversight agencies to prioritize investigations and allocate resources. We show how the models developed shed light on specific features of the contracts to be considered and how their values signal red flags. We also highlight how these values change when the analysis focuses on specific contract types or on information available for early detection. Moreover, the models and indices developed only make use of open data and target variables generated by the procurement processes themselves, making them ideal to support continuous decisions at overseeing agencies.

Policy Significance Statement

Overseeing agencies are tasked with the key but very challenging task of preventing inefficient and irregular practices in public procurement. Here we describe the development of tools to support these entities in the early detection of these practices. The tools are designed and developed with the user in mind, prioritizing explainability and easy access to the data required, especially making use of open data sources. The results reveal which operational variables associated with the procurement process are key for the early detection of inefficiencies. The tools can further support decisions regarding the prioritization of investigations and resource allocation in overseeing agencies.

  This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.



1. Introduction

Public procurement plays a key role in the successful operation of public entities and their ability to deliver services to citizens. Given the size and number of public entities in any country, their procurement operations can be very large and wide-ranging, posing many challenges to their monitoring and oversight. This is particularly relevant as the activities developed through public procurement are funded by taxes, their goals are of public interest and their impact is significant to the citizens' well-being. Traditionally, public procurement oversight has been performed by specialized entities, where teams of investigators monitor public contracts grouped by sectors, geographical areas, and buyer entities. However, these oversight entities have limited budgets and personnel while the vast size of the public procurement activity makes the careful monitoring of each process infeasible. This limitation naturally leads to the selection and prioritization of procurement processes to oversee. Here, the experience of the investigators is central to deciding whether to monitor a procurement process and in which level of detail. Whereas this mode of operation is a necessity given the current state of affairs, the availability of procurement data opens up opportunities to incorporate new technologies that can better support the monitoring and oversight tasks in public procurement.

To support the overseeing process, machine learning methods have been proposed to facilitate the detection of inefficiencies in public procurement. Examples of this approach include the work of Colonnelli et al. (2022) which focuses on data from anti-corruption audits in Brazil, De Blasio et al. (2020) which consider white-collar crime in Italian municipalities, and Decarolis and Giorgiantonio (2022) that look into the tendering processes in roadwork projects in Italy, among others. However, many of these works focus on the so-called active waste, as defined by Bandiera et al. (2009), which refers to deliberate actions by decision-makers that lead to waste, as in corruption cases. This is in contrast to passive waste, where inefficiencies are caused by a lack of skills or incentives. While active waste may attract more attention, Bandiera et al. (2009) found that passive waste accounted for 83% of the total estimated waste in Italian state agencies. To address this gap, in this study we adopt a holistic view for the procurement process assessment, considering both passive and active waste. Passive waste is captured through machine learning models that focus on inefficiencies, while active waste is reflected via risk indices (IRIC) that focus on irregularities.

As described by Lakkaraju et al. (2017), when building data-based solutions for public procurement overseeing, it is easy to fall for the so-called "selective labeling problem." This problem arises when the indices or models are based on data from cases detected by entities like the Comptroller's Office or the judicial system, which therefore leaves out all other cases that were not caught by these entities. This leads to biases in the analysis that limit their reach to what has already been detected and prioritized by the judicial system. To bridge this gap, in this study, we develop models and indices that rely on objective measures of both inefficiencies and irregularities without passing through the filter of any other party. Further, these data also have the advantage of being updated regularly (daily), increasing the effectiveness of the methods proposed in this paper to support decisions regarding investigation prioritization and resource allocation.

In addition, many existing studies in this area have focused on black-box models with a large number of variables, for instance in Lopez-Iturriaga and Sanz (2018); Ash et al. (2021); Decarolis and Giorgiantonio (2022); De Blasio et al. (2022). While introducing an extensive set of variables enables these models to reach high accuracy levels, it also hinders their explainability, which is key to gaining trust and facilitating their use by investigators at the overseeing entities. We approach this gap by developing models with a limited set of variables and show that these can achieve good performance while maintaining their explainability. Thanks to these models, we are also able to shed light on which contract characteristics are most important when anticipating instances of inefficiency. That is, it makes it possible to identify the red flags to which the authorities should pay more attention. The analysis used allows not only to identify the variables with the greatest predictive power to detect inefficiencies in public procurement but also to understand how the correlation between these variables and the outcome of interest varies throughout the distribution of the contract features. In addition, it shows how the predictive capacity of these tools changes depending on the type of contract and the moment of the process in which the prediction is carried out.

Finally, the models and indices proposed in this study are put together in VigIA, a tool that supports the early detection of procurement processes with a high risk of resulting in inefficiencies and irregularities. As inefficiencies, we specifically focus on cost overruns and delays and consider as irregularities other

aspects that can impact the transparency and competition in the procurement processes. To detect inefficiencies we propose machine learning models, while for irregularities we develop risk indices, all at the contract level. As VigIA was an actual tool developed for an overseeing entity, many design decisions were made to favor the tool transfer and adoption.

In the next section, we provide an overview of recent experiences regarding the introduction of predictive models and risk indices in public administration, particularly in public procurement. Next, we provide context regarding the procurement system considered, before discussing the data, methods employed, and the key findings. The article concludes with a discussion of the key lessons learned and future work.

2. Literature review

Several studies have explored the implementation of machine learning algorithms to address public policy problems. These applications have occurred in fields as varied as health (Kleinberg et al., 2015), education (Rockoff et al., 2011), public finance (Zumaya et al., 2021), violence prevention (Chandler et al., 2011), justice (Kleinberg et al., 2018), poverty reduction (Blumenstock et al., 2015), food security (Hossain et al., 2019), official communications (Jungblut and Jungblut, 2022), among many others. Most of these applications fall into the category that Kleinberg et al. (2015) define as “prediction policy problems,” that is public policy situations where causal inference is not so relevant, but the prediction of the circumstances under which an intervention will be more effective is. This prediction can then support decision-makers in allocating scarce resources. In this study we focus on the use of machine learning algorithms and risk indices, to anticipate inefficiencies and irregularities in public procurement processes and thus better allocate resources to oversee these processes.

2.1. Machine learning for public procurement overseeing

In the nascent literature that seeks to implement machine learning and artificial intelligence tools to predict corruption and waste in public procurement, the analyses vary on the level of aggregation from the larger ones at the municipality level, down to the entity, and further to the contract level. At a high level of aggregation, we find that Colonnelli et al. (2022) employ data from anti-corruption audits in Brazil to train a set of machine learning models and assemble them through the SuperLearner approach (Polley et al., 2011) to predict municipal corruption. After grouping variables into categories to assess the predictive power of each category, the authors find that variables associated with the private and financial sectors are more important than political or public sector characteristics. Also, in the De Blasio et al. (2020) study, the focus lies on the prediction of white-collar crime at the municipal level in Italy, using classification trees. The authors find that their models reach high levels of accuracy, that variables related to labor and housing markets are the most important, and that the use of predictions based on machine learning would improve the authorities’ fight against corruption. Finally, Lima and Delen (2020) carry out an even more aggregate exercise, attempting to predict corruption at the country level. The authors identify the variables with the greatest predictive power, highlighting the integrity of the government, property rights, judicial effectiveness, and education.

At a more granular level, Decarolis and Giorgiantonio (2022) carry out an analysis at the contract level, focusing on roadwork projects in Italy. The authors seek to determine which characteristics of the tendering design, the so-called “red flags,” are best at predicting the risks of corruption during the contract awarding stage. Among other factors, the authors find that the employment of an awarding criterion based on multiple parameters is highly predictive of procurement irregularities. To the best of our knowledge, there are not many other works that focus on predictive models at the contract level. Works at the same level of granularity but not using predictive models include Fazekas and Kocsis (2020), which focus on corruption risk indices; Kenny and Musatova (2010), which evaluate red flags to investigate governance failure, collusion or corruption in projects using a sample of World Bank water and sanitation projects; Szucs (2023), which studies discontinuities in procurement outcomes and the density of the contract values, found to be related to manipulation by the buyers to avoid auctions; and Fazekas and Wachs (2020), which

develops contract-level indices for Check Republic and Hungary to detect corruption and the resulting market distortions.

Closely related to this work, in the study of Gallego et al. (2021), the authors carry out an analysis at the contract level partially using open data for public procurement in Colombia. The authors train black box machine learning models employing a large (100+) number of features available along the process lifecycle and considering processes of all types. As the target variable, the authors use corruption investigations, breaches of contract, and implementation inefficiencies. They find that the models reach high levels of predictive power, with the budget and execution period as the characteristics with the most predictive power. As we develop a tool to support resource allocation in a procurement oversight entity, we also work at the more granular level of the contract.

2.2. Risk indices for public procurement overseeing

We also note that in recent years part of the literature has been interested in designing objective corruption risk indices using open data in public procurement. These indices have emerged in response to traditional measures of corruption that were mostly based on perception surveys. Studies such as Fazekas et al. (2016), Charron et al. (2017), Fazekas and Kocsis (2020), and Gnaldi et al. (2021), propose and discuss indices that capture typical red flags of public procurement and that can serve as early warnings.

Rodríguez-García (2022) shows a risk index of corruption by political parties in Latin America based on an analysis of laws and regulations. Fazekas and Kocsis (2020) built a tendering composite score at the contract level for European countries, with a focus on the lack of competition dimension. In Colombia, Zuleta et al. (2019) provided a robust corruption risk index that takes into account three main dimensions based on a methodology developed by The Mexican Institute for Competitiveness (IMCO) (IMCO, 2018). However, they built the indices at different levels of aggregation, such as political parties and buyer units (public entities). In this paper, we contribute to this body of literature by proposing indices of this nature for the Colombian case based on Open Data, which can be used as early indicators of contractual irregularities in public procurement and serve as a complement to the machine learning models that focus on inefficiencies.

2.3. Considering active versus passive waste

Bandiera et al. (2009) distinguish between active and passive waste in public procurement. Active waste occurs when public decision-makers deliberately benefit from wasting resources, often through corrupt practices, while passive waste arises from inefficiencies due to a lack of skills, incentives, or excessive regulation, without any direct benefit to the decision-maker. The current literature has predominantly focused on predicting active waste. For example, Ash et al. (2021) use cases of corruption detected by random audits of the General Comptroller's Office of Brazil, Lopez-Iturriaga and Sanz (2018) utilize corruption cases reported by courts or the media in Spain, and Decarolis and Giorgiantonio (2022) focus on judicial measures for road contracts in Italy. Similarly, De Blasio et al. (2022) address white-collar crimes in Italy. However, examining passive waste is equally, if not more, important. Bandiera et al. (2009) found that passive waste accounted for 83% of the total estimated waste in their study of state agencies in Italy. Our study significantly addresses this gap by considering both active and passive waste. Our outcomes of interest encompass active waste, likely correlated with our irregularity measure (IRIC), and passive waste, strongly associated with our inefficiency measures (cost overruns and delivery delays). Therefore, our approach predicts waste in public procurement holistically, without prioritizing one form over the other.

2.4. Selective labeling

Even when focusing on active waste, the literature has often relied on measures that likely suffer from significant measurement error, or worse, from what Lakkaraju et al. (2017) term the "selective labeling problem." Using measures of corruption detected by entities like the Comptroller's Office and judicial systems implies that models focus on visible corruption, but miss the invisible corruption. This approach

may concentrate on those who “got caught,” potentially excluding the more skillful offenders. In the worst-case scenario, the corruption investigators themselves might be biased (possibly influenced by corruption), leading to instances of genuine corruption not being classified as such, and vice versa. This could result in biased predictions with limited utility. Our study addresses this gap by using objective measures for both inefficiencies and irregularities that do not necessarily depend on labeling by involved parties. A contract that experiences delays or cost overruns, takes a long time to be awarded, or is given to a multipurpose contractor is objectively recorded in the system as such.

2.5. Black-box vs explainable models

Existing studies have generally employed black-box models that are complex to explain and implement, often relying on a large number of predictive variables. For instance, Lopez-Iturriaga and Sanz (2018) use neural networks, despite using few macroeconomic variables at the provincial level. Ash et al. (2021) utilize Gradient Boosting Machines (GBM) with nearly 800 variables. Decarolis and Giorgiantonio (2022) employ lasso, ridge, and random forests with a similarly extensive set of variables. Similarly, De Blasio et al. (2022) use classification trees in their models with almost 100 variables. While these approaches can enhance model accuracy, our study demonstrates that a relatively simpler approach with a restricted set of variables can achieve good performance levels. Addressing this gap is crucial if the goal of these models is to be implemented practically, assisting agencies in optimizing their scarce resources in the fight against the waste of public resources.

2.6. The case study of Bogota and VigIA

The case study of Bogota contracts and the models employed in this paper effectively address the identified gaps. The open data from Bogota allow us to construct measures related to both active and passive waste, as these data include indicators of cost overruns and delivery delays, as well as specifics of the procurement processes. This enables us to build objective indicators of inefficiencies and irregularities that are less susceptible to measurement errors or the selective labeling problem. Furthermore, the data set is sufficiently rich in both variables and the number of contracts, allowing for automated pre-selection of predictors while using simpler models that are easier to interpret and implement. By focusing on a single local government, such as the city of Bogota, we can control for various political, economic, and institutional factors that might correlate with public resource management, without losing variability or scale in terms of the type and volume of public procurement. This comprehensive approach ensures that our models are both practical and robust, providing valuable insights for optimizing resource allocation in public procurement oversight.

3. Background

This study focuses on the case of the city of Bogota, Colombia, where public contracting is governed, mainly, by Law 80 of 1993¹ and 1150 of 2007,² which define different mechanisms through which public procurement can be carried out. The most common are public bidding, abbreviated selection, merit competition, and direct contracting. These mechanisms make it possible to differentiate between competitive and non-competitive processes, mainly in terms of the potential number of bidders that a procurement process may have. Direct contracting implies the discretionary selection of the supplier of the good or service in question, for which it is considered a non-competitive mechanism. In fact, the law states that direct contracting should only be used in exceptional cases. However, it is a fairly common contracting mechanism, as illustrated by Figure 1, which shows that

¹ See <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=304> (In spanish). Last accessed 02/17/2023.

² See <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=184686> (In spanish). Last accessed 02/17/2023.

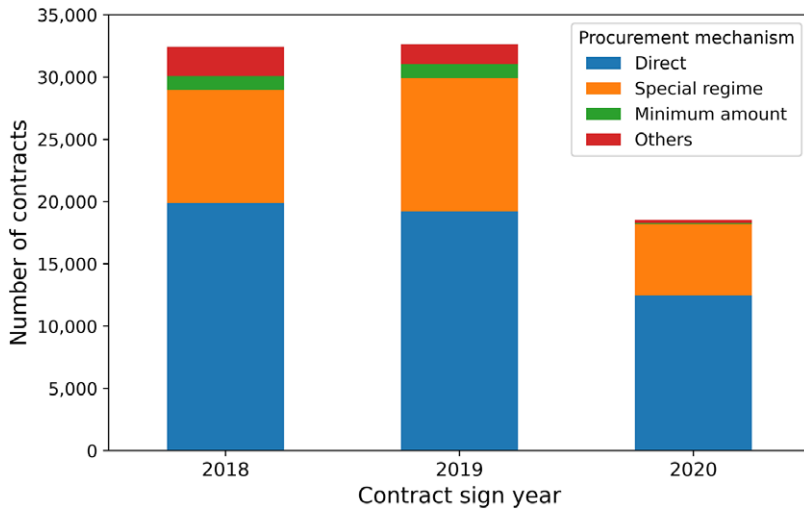


Figure 1. Number of contracts reported virtually by procurement mechanism across the years in Bogotá.

direct contracting is in fact the most common procurement mechanism among the contracts reported in Bogotá between 2018 and 2020.

In recent years, the use of electronic platforms in public procurement has been encouraged at different levels. The Electronic System for Public Procurement (SECOP by its acronym in Spanish) in its current version (SECOP II), is a transactional tool where public agencies carry out their purchases electronically. To this end, entities must use the platform to create the procurement processes, which may result in one or many contracts. In this study, we focus on the contract as the unit of analysis but incorporate characteristics of the process under which the contract was signed, such as the advertisement period duration or the number of bidders.

For later reference we also point out five key dates in a contract life cycle that are used in the models and indices developed: (i) signing date, when the contract is signed by both parties; (ii) start date, when the contractual responsibilities of the supplier start; (iii) and (iv) execution start and end dates, when the actual execution starts and ends; and (v) end date, when the contractual responsibilities of the supplier end. While ideally, these dates should occur sequentially in the listed order, we actually find that in many instances the signing date occurs after the start date. This will become evident in the results presented in Section 6 and will in fact represent a red flag for inefficient contracts.

Although adoption rates vary significantly across the country, in 2017, the Mayor's Office of Bogotá committed to fully adopting the SECOP II platform. As a result, all public procurement in the city is now conducted through this platform, offering a comprehensive view of the platform's data. Furthermore, tables containing information registered on the platform are published daily as open data, making them perfectly suited for the tools developed in this study.

3.1. The Veeduría Distrital

The Veeduría Distrital (or District Oversight Office), created by Law 1241 of 1993,³ is the agency in charge of ensuring morality and administrative efficiency in the city of Bogotá, with a unique nature in the country for being of the sub-national order, but independent of its executive body (Rodríguez Arévalo et al., 2021). One of the key objectives of this entity is to exercise *preventive* control in the use of the city's public resources. However, the office is relatively small, in terms of budget and staff. To put it in perspective, in 2020 the assigned annual budget represented only 0.1% of the total city budget (Rodríguez

³ See <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=106394> (In spanish). Last accessed 03/04/2023.

Arévalo et al., 2021). This implies that the Veeduría, each year, must monitor hundreds of thousands of contracts with relatively scarce resources. Therefore, developing analytical tools to optimize these processes was a latent need for the entity.

In fact, the results presented in this study have their genesis in a project submitted by the Veeduría Distrital to the CAF-Development Bank of Latin America and the Caribbean in the context of a call to support the strategic use of data and artificial intelligence in public entities. Among more than 80 projects presented by public entities across the region, the proposal of the Veeduría Distrital was the winner.⁴ As a result, a key focus of this study is to develop tools that can be easily communicated to public procurement investigators at overseeing agencies, such that resource allocation decisions are properly supported and understood at the process and contract levels. Furthermore, this also leads us to build differentiated models for different types of contracts and focus on models that only employ information available at early stages in the process, thus supporting preventive control.

4. Data

The main data source for this project was the public contracting platform SECOP II, maintained by *Colombia Compra Eficiente*, complemented by the record of sanctions imposed by the overseeing agency, *Superintendencia de Industria y Comercio* (SIC). Both of these sources are available on the Colombian open data platform, *Datos Abiertos Colombia*.⁵ In this section, we explain how these data were processed and transformed to be used by the models and indices presented in the next section.

4.1. Electronic contracts data

SECOP II is a transactional platform that connects public entities and suppliers to carry out official procurement processes online. To support traceability and transparency, much of the information generated in a process is made publicly available as data tables with distinct levels of aggregation. We used the electronic contracts data table *SECOP II—Contratos Electrónicos*⁶ as the main table for our analysis, where the unit of observation is the contract, which corresponds to the most granular level of aggregation. From this table, we selected the contracts generated by territorial entities in Bogotá, and the autonomous corporations associated with the city. In addition, we only considered the contracts that report an execution end date before Jan-01-2021, to focus on those contracts that could have reported cost overruns or delivery delays, thus obtaining an initial set of 87,387 contracts and 66 variables.

4.2. Procurement process data

We considered the procurement processes registered in the *SECOP II—procesos de contratación*⁷ data table, which groups the needs of the projects for which a buyer opens a tender and may be composed of many contracts. In this data table, a reconciliation was required as there were multiple records for the same process identifier (3.9% of the records). Hence, we selected the variables of interest, eliminated duplicate rows (in all their fields), and finally grouped the rows by *procurement process ID*, reconciling conflicting information; for instance, between several publication date records for the same process identifier, we used the oldest date. Finally, we obtained a table with 156,520 procurement processes. Next, we merged the procurement processes table with the electronic contracts table by *procurement process ID*,

⁴ See <https://www.caf.com/es/actualidad/noticias/2020/07/veeduria-distrital-de-bogota-gana-el-llamado-a-instituciones-publicas-de-caf/> (In Spanish). Last accessed: 05/15/2023.

⁵ <https://datos.gov.co>. Last accessed: 05/15/2023.

⁶ Data collected on Feb-04-2021 from Datos Abiertos Portal (<https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-II-Contratos-Electr-nicos/jbjy-vk9h>).

⁷ Data collected on Feb-04-2021 from Datos Abiertos Portal (<https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-II-Procesos-de-Contrataci-n/p6dx-8zbt>).

to obtain information at the process level for each contract, such as the number of offers, advertisement period variables, etc.

4.3. *Provider data*

We also obtained information about the supplier company type and its date of first registration on the platform, from the *SECOP II—Proveedores Registrados*⁸ data table, where we dropped duplicated values and rows with missing or inconsistent supplier identifiers. Additionally, we explored the use of information from sources other than SECOP II to have a wider set of explanatory variables. We used the record of 14,150 sanctions imposed on persons and companies by the *Superintendencia de Industria y Comercio* (SIC) between 01/01/2015 and 12/31/2020, available on Colombia's Open Data portal.⁹ These sanctions are imposed in cases of cartelization, unfair competition, and bad consumer service. With these data, we computed the number and value of penalties for a supplier before the beginning of execution of each contract. Also, we used the United Nations Standard Products and Services Codes (UNSPSC) to aggregate the products or services procured in larger sets and to use this as a predictor variable.

4.4. *Outcome variables*

In order to create the outcome variables for the inefficiency models, we used the record of modifications in contracts in the data table *SECOP II—Adiciones*.¹⁰ These data have the contract identifier, the date, and the type of modification (cost overruns, delivery delays, general modification, conclusions, and others). With this information, we grouped by contract to get the number of cost overruns or delivery delays reported for each reported contract. Next, we linked them to the contracts table by *Contract ID* to obtain the outcome variables. In this manner, contracts without cost overruns or delivery delays are marked with 0, and those with either issue with 1.

4.5. *Missing data*

After completing the previous steps, we obtained a table with 87,387 rows and 95 variables. We perform a first purge of variables to drop those that, due to their definition or being almost constant (over 99.99% of the observations have the same value) are not relevant for the models and indices, leaving 52 variables. Regarding missing values, we found that 8.9% of the contracts have at least one missing column: 5.2% missing values in variables at the procurement process level, 2.9% missing values in the signing date, and 1.2% missing values in the supplier registration date in the platform. To impute the signing date, we used the contract start date minus the median number of days between the contract signing date and the start date by type of contract. In the case of the supplier registration date, missing values are filled with the signing date. For the remaining 5.2% missing values, we created the variable *Have procurement process* that takes the value of 0 in the case of missing values for the procurement, and 1 otherwise. This variable is used to filter the observations used for the inefficiency models and also serves as an input variable for the irregularities indices.

4.6. *Feature engineering*

We performed feature engineering to summarize some variables, create new ones, and clean outliers. First, for all categorical variables, we joined categories according to their meaning, and put together those categories with a relative frequency of less than 0.1% into the category "other." In addition, we used the date variables to create characteristics that capture the duration in days between different events. For

⁸Data collected on Feb-04-2021 from Datos Abiertos Portal (<https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-II-Proveedores-Registrados/qmzu-gj57>).

⁹Data collected on Feb-04-2021 from Datos Abiertos Portal (<https://www.datos.gov.co/Comercio-Industria-y-Turismo/Sanciones-impuestas-en-firme-por-la-SIC-a-personas/i3z6-57ui>).

¹⁰Data collected on Feb-04-2021 from Datos Abiertos Portal (<https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-II-Adiciones/cb9c-h8snhttps://www.datos.gov.co/Gastos-Gubernamentales/SECOP-II-Adiciones/cb9c-h8sn>).

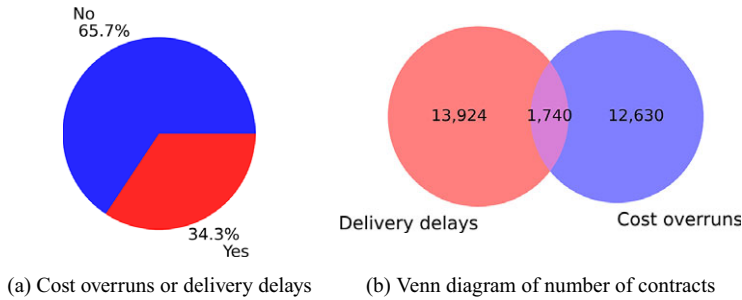


Figure 2. Distribution of contracts according to the presence of cost overruns or delivery delays.

example, Sign-start contract days, the time between the signing and the start of the contract, and *start-end execution days*, the time between the start and end of the execution. It is important to highlight that although counterintuitive, we found that it is a common practice to sign a contract after its start date, which causes negative values in some of these duration features. Regarding the outliers, we dropped the observations where the contract value is greater than the 99.5 percentile across contracts different from public works. As a result, we obtain a table where 34.3% of the contracts have cost overruns or delivery delays, as displayed in Figure 2a. Also, in Figure 2b we see that delivery delays are somewhat more common, and only 1740 contracts have both types of inefficiencies.

4.7. Data selection according to the process stage

Our processed dataset is composed of 87,027 contracts and 47 variables (see [supplementary material A and B](#)) with variables associated with three distinct stages in the procurement process: pre-contractual, adjudication, and execution stages. First, the pre-contractual variables are those available from the publication of the process on the platform until its adjudication. Second, the adjudication variables are those related to the assignment of the process to a supplier. Third, the execution variables are only those available once the actual service or product is delivered after adjudication, including, for instance, the execution start and end dates, the payments made in advance, and the inefficiency outcome variables.

5. Methods

In this section, we describe the methods we employ to develop the models and indices to address two complementary aspects of corruption in public procurement, namely, inefficiencies and irregularities.

5.1. Inefficiencies models

The main goal of the models we propose is to predict whether a contract is likely to present inefficiencies, measured as cost overruns or delivery delays. Thus, we are faced with a binary classification problem where the categories are inefficiencies versus no inefficiencies. To tackle this problem there is a wide range of classification methods in the statistical learning literature (Hastie et al., 2009; Bishop, 2006; James et al., 2013). As mentioned in Section 3, the models here developed are to be employed by the inspectors at oversight agencies, who decide which contracts to oversee. Thus, the models are required to be explainable, i.e., to allow a clear understanding of the reasons behind the alerts raised by the models, and thus improve the acceptance and use of the tool (Wenzelburger et al., 2022). This central concern guided our selection of classification methods towards transparent and explainable alternatives instead of black-box approaches, such as neural networks (Goodfellow et al., 2016).

We have therefore opted to employ logistic regression and random forests as the classification methods of preference, due to their explainability. On the one hand, as we look to predict how likely a contract is to present cost overruns or delivery delays, or either, logistic regression offers a simple mechanism to associate this

likelihood with the features of the contract, such as the type of the procurement process, the type of contractor, or the type of goods/services procured. On the other hand, a random forest employs a large number of simple decision trees to determine whether a given contract is likely to present cost overruns or delivery delays, or neither. Whereas each decision tree is relatively simple and has partial information, combining many of them allows the random forest to provide predictions that are both more accurate and have a smaller variance. Due to its construction, with decision trees and random forests, it is possible to estimate the feature importance, which can guide the user, in this case, the investigator, into the reasons why a specific contract is classified as irregular or not. Furthermore, we estimate partial dependency plots to derive insights into the specific relation between the contract characteristics and the inefficiency prediction. The combination of these tools facilitates the communication and explanation of results to the oversight investigators.

5.1.1. Machine learning performance measures

A model's performance is measured through a number of metrics to capture the model's ability to predict the actual outcome, i.e., whether a contract incurs in inefficiencies or not. The model predicts a positive (resp. negative) result when it labels a contract with (resp. without) inefficiencies. If the prediction is correct it is called a true positive (TP) or true negative (TN), whereas if the prediction fails it is called a false positive (FP) or false negative (FN). The model's *accuracy* can then be defined as the ratio of correct predictions to the total number of predictions, i.e.,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Next, the model's *precision* captures the fraction of positive predictions that were correct, i.e.,

$$\text{Precision} = \frac{TP}{TP + FP}.$$

Also, the *recall* measures the fraction of actual positives that were correctly predicted by the model, i.e.,

$$\text{Recall} = \frac{TP}{TP + FN}.$$

While it is desirable that both precision and recall are large and close to one, these two metrics tend to be in conflict and an increase in one may be achieved by decreasing the other. Finally, we consider the receiver operating characteristic (ROC) curve, which graphs the Recall, also known as the True Positive Rate (TPR), and the False Positive Rate (FPR), defined as

$$\text{FPR} = \frac{FP}{FP + TN}.$$

Each point in the ROC displays the FPR (x axis) and the TPR (y axis) for a different classification threshold, i.e., the minimum value between 0 and 1 chosen to classify a contract as a positive. This allows the ROC to provide a result that is independent of such threshold. To aggregate the information in the ROC in a single number, the area under the curve (AUC) is computed, which is a number between zero and one. The closer the AUC is to one, the better the model's ability to provide true positive predictions over false positives.

To evaluate the models we shall focus on the measures defined above (accuracy, recall, precision, AUC) as well as on the confusion matrix, which displays the TP, FN, TP, and FP measures directly.

5.2. Irregularities indices

To complement the inefficiencies models, we built a Contract Irregularity Risk Index (IRIC, by its acronym in Spanish), and a weighted version (IRICP), that provide a measure of irregularities risk in a contract once it is signed. The goal of these indices is to cover other areas of the public procurement

process that can be indicative of irregularities, beyond the cost and time inefficiencies considered with the machine learning models.

We defined the indices following the findings of Zuleta et al. (2019), as an aggregation of multiple variables that identify irregularities in three dimensions, as in IMCO (2018): lack of competition, lack of transparency, and anomalies in the procurement process. However, while Zuleta et al. (2019) and IMCO (2018) define the indices at the entity and even more aggregate levels, we build the indices at the contract level. We thus revise the variables in each dimension to better fit the open data available, so as to avoid reliance on private or hard-to-obtain data sources.

To build these indices we constructed a record for each supplier based on contracts previously assigned to it at the time of the contract signature. In this record, we include the number of previous contracts with cost overruns or delivery delays, the number of different economic activities performed by the supplier, and the number of public contracts previously assigned to the supplier. With these records and the actual contract information, we built 11 binary variables (0/1) that define the IRIC, as follows.

1. Lack of competition

- (a) **Single or non-proponent:** a value of one represents a procurement process in which a single bidder, or none, submits an offer on the platform. This phenomenon is associated with high discretion in the selection process. A less competitive procurement process tends to favor companies that are politically connected, increases prices, and results in lower productivity (Baltrunaite et al., 2020; Szucs, 2023).
- (b) **Multipurpose suppliers:** this equals one if the number of different economic activities developed by the supplier is higher than one. The selection of a multipurpose supplier, which provides multiple and diverse goods and services, may indicate poor care about the suitability of the procurement needs, and increase the risk of rent extraction and corruption in Colombia's public procurement (Open Contracting Partnership, 2020).
- (c) **Exceptionally high supplier history:** used to identify companies that frequently win public offers, which could indicate favoritism. To define what constitutes a high value we compute the 95th percentile of the empirical distribution by type of contract. From this calculation, we obtained that a supplier is marked with a 1 in this variable if it has previously won more than three contracts in the case of professional services, or more than 6 contracts otherwise.
- (d) **Direct contracting:** this procurement mechanism implies that the entities directly engage with suppliers to pick who can submit bids, making the procurement process less competitive and increasing the risk the corruption (Fazekas and Kocsis, 2020). Thus, if a contract is awarded through the direct contracting mechanism, it is marked with a 1.
- (e) **Special regime:** in Colombia, the special regime mechanism was created for public entities that operate in competitive markets and execute public resources, such that contracts under this mechanism are exempt from the Colombian General Contracting Statute. Therefore, the use of this procurement mechanism requires special conditions that grant particular benefits to contractors, hence it must be carefully justified and cannot be used frequently. However, there are entities that employ this mechanism despite not operating in competitive markets, evidencing its incorrect use and the subsequent restrictions to competition in procurement processes (Zuleta et al., 2019). Thus, if a contract is awarded through the special regime mechanism, it is marked with a 1 in this variable.
- (f) **Extreme advertisement period:** during this period the procurement process is open to receive offers, calculated as the number of days between a tender publishing date and the submission deadline. A very short advertisement period may indicate an advantage for informally informed suppliers against others, which may signal corruption risks (Decarolis and Giorgiantonio, 2022; Fazekas and Kocsis, 2020). Also, a very long advertisement period may be caused by legal challenges, which may also signal risks (Fazekas and Kocsis, 2020). Thus, to define what duration can be labeled as extreme for the advertisement period, we reviewed its empirical distribution differentiating between contracts for professional services and others (see [supplementary material C](#)). We selected the 99th percentile of the distribution such that a

process shows an extreme advertisement period if its duration is greater than 14 days for professional services contracts, or greater than 31 days for other types of contracts. For the lower limits, we identified the first percentile as 1 day for both types of contracts, such that an advertisement period of zero days is marked as extremely short.

2. Lack of transparency

- (g) **Errors or missing data:** although it is not mandatory to fill all the information fields on the platform, some key data should never be missing, since its absence limits external evaluation and may indicate an irregularity. Therefore, this variable is marked as one if any of the following fields are absent or not duly registered: the supplier's ID is missing or inconsistent, the procurement mechanism justification is missing, or the contract value is missing or abnormally high for the goods or services to be acquired. From the contract value distributions, we consider abnormally high those above 221 million COP for professional services contracts, above 207 thousand million COP for public works, and 5 thousand million COP for all others.
- (h) **Extreme decision period:** this period is the time required to choose the supplier since the time the process closes for submissions. Abnormal decision periods can signal corruption risk at both extremes: if the period is excessively short, it may reflect a premeditated assessment, while if it is very long it may be due to legal challenges raised against the decision-making process or the initial award decision (Fazekas and Kocsis, 2020). As the platform does not provide the date on which the chosen supplier is announced, we used as a proxy the number of days between the submission deadline and the contract signature date, and fill missing values with 0. As before, we estimated the distribution of the decision period by differentiating by contract type and used the 5th and 95th percentiles to mark the threshold of what is considered an excessively short or long decision period, respectively. We obtained that a regular decision period takes between 1 and 43 days for professional services contracts, and between 1 and 55 days for other contract types (see [supplementary material C](#)). Thus, if the decision period is outside of these bounds, it is considered extreme and this variable is marked with a 1.

3. Anomalies in the procurement process

- (i) **Supplier with cost overruns:** it is marked with a 1 if the supplier assigned to the contract had previous contracts with cost overruns, which indicates that the supplier has been inefficient in the use of resources.
- (j) **Supplier with delivery delays:** similar to the previous one, this variable is marked with a 1 if the supplier assigned to the contract had previous contracts with delivery delays, another measure of inefficiency.
- (k) **Absence of a procurement process:** one of the characteristics that identifies an irregular contract is that its offer was not made public before contracting. If the contracting process to which a contract corresponds is not found on the platform, this variable is marked with a 1.

With the aforementioned variables, we calculate the IRIC index as their arithmetic mean to obtain a value between 0 and 1 for each contract. A higher IRIC represents a contract more likely to be irregular.

Additionally, to take into account the contract size we introduce the IRICP. To obtain the weights for each contract, we calculated the log transformation of the contract value, $\log(\text{contract_value} + 1)$, and get the 75th percentile (labeled α) by contract type, to obtain $\alpha = 17.51$ for professional services, and $\alpha = 18.11$ for non-professional services. We set the weight as the ratio $\log(\text{contract_value} + 1)/\alpha$, and built the IRICP as

$$\text{IRICP}_i = \frac{\log(\text{contract_value}_i + 1)}{\alpha} \text{IRIC}_i,$$

$$\alpha = \begin{cases} 17.51, & \text{if } \text{Contract type} = \text{"Professional services"}, \\ 18.11, & \text{otherwise.} \end{cases}$$

Therefore, for contracts with the same value of IRIC, the IRICP index prioritizes those with a higher value. This approach addresses the concern that high-value contracts entail greater risk, as failure to identify fraudulent contracts of this nature could result in significantly higher financial losses.

6. Results

In this section, we present the key results obtained with the models and indices developed to predict inefficiencies and irregularities in public procurement for the city of Bogotá. While we developed a large number of models, with different objectives and sets of variables, here we focus on the main findings and differences among the models developed.

6.1. Models for cost overruns

For our base model, we start with all the variables available for analysis, as described in Section 4. As the target variable, we select *Cost overruns*, which tags those contracts that incurred cost overruns. The selection of this target variable implies a large class imbalance as the majority of contracts do not incur cost overruns. We thus perform a resampling of the data to obtain a new dataset with an equal number of observations for both classes (with and without cost overruns).

Once the resampling has been performed, we standardize the explanatory variables and train a random forest model to determine the set of most significant features. The random forest can compute a feature importance score as it estimates, for each branch in a tree, the amount of discriminatory power of a given feature. The objective here is to select a relatively small subset of variables while maintaining the predictive power of the model as high as possible. The selection of this small subset of variables is key to our goal of providing an explainable model to the investigators. This process results in the selection of just seven variables: Value, Sign-to-start contract days, Start-to-end contract days, Sign-to-start execution days, Start-to-end execution days, Days supplier registered, and Sector Culture (which represents 14.1% of the contracts). Descriptive statistics for these variables (over the full dataset) are presented in Table 1.

As these seven variables turn out to be the most important given their strong predictive power on whether a contract ends up having cost overruns or not, it is relevant to discuss them. To this end, we employ the partial dependency plots derived from the random forest model trained to predict cost overruns. First, the contract value turns out to be a key predictor of cost overruns, where a higher value indicates a higher probability of a contract having them. While the model results indicate that high-value contracts are in fact prone to cost overruns, the partial dependency plot in Figure 3, indicates that this is true for intermediate-value contracts too. This plot captures the marginal dependency of the target variable, i.e., the likelihood of observing a cost overrun, as a function of the explanatory variable, i.e., the contract value (Hastie et al., 2009). The top left plot in Figure 3 shows that contracts with values in the middle range are also prone to cost overruns, much more than contracts with a relatively small value (the contract value is displayed on a normalized scale around zero). We also explicitly mark the inflection point in the original scale to determine a threshold of what can be considered a contract value sufficiently high to require special attention. In this

Table 1. Descriptive statistics of numeric variables in the cost overruns models

Index	Mean	std	25-th perc	50-th perc	75-th perc
Value (COP)	72,454,998	797,445,592	10,025,890	20,444,050	41,753,400
Sign-start contract days	-9.84	37.53	-4	0	1
Start-end contract days	217.44	126.46	110	187	333
Sign-start execution days	-1.76	53.87	-3	0	2
Start-end execution days	211.79	126.6	104	181	331
Days supplier registered	333.33	302.64	67	281	474

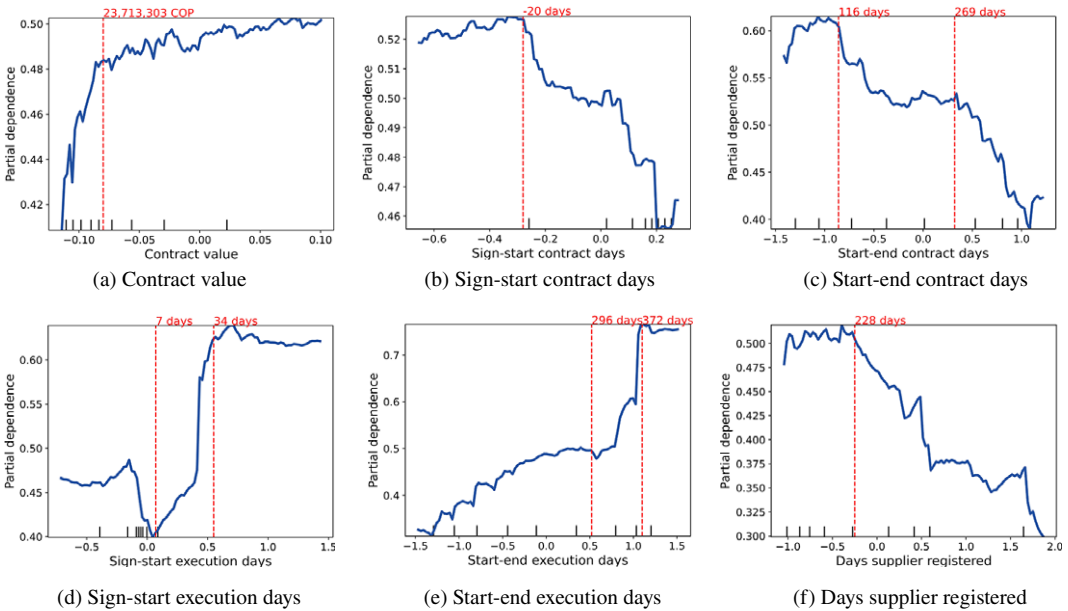


Figure 3. Partial dependency plots of explanatory variables in the Random forest model to predict cost overruns. All variables in the x axis are standardized, and red dashed lines indicate tendency breakpoints with the value in the original scale.

case, it corresponds to 23,713,303 COP, which is a rather low value, somewhat above the median displayed in Table 1. This highlights the need to widen the investigators’ reach beyond high-value contracts to tackle intermediate-value contracts that may also display cost overruns.

Let us now consider the *Sign-to-start contract days* variable, which shows that a shorter time between the contract signature and its start increases the likelihood of cost overruns. This can be observed in the top center plot in Figure 3 where the decreasing shape implies that a higher likelihood of cost overruns is observed for low values of the *Sign-to-start contract days* variable. The mark in the original scale shows that when this difference reaches -20 days or less, i.e., the contract is signed *at least 20 days after* its start date, the likelihood of cost overruns becomes very high. While it is not unusual for contracts to be signed slightly later than their start date (as shown by the 25th percentile of this variable in Table 1), when this delay gets large it becomes a predictor of cost overruns.

A similar relation is observed with the *Start-to-end contract days* in the top right plot of Figure 3, where we observe three main regions for the contract duration: when it is larger than 269 days (about 9 months), the contract has a small likelihood of cost overruns; when this duration is between 116 and 269 days (about 4 to 9 months) the likelihood of cost overruns increases; and when the duration is under 116 days, cost overruns become very likely. Conversely, a large delay in starting the execution after signing the contract, i.e., variable *Sign-to-start-execution days*, is associated with a larger likelihood of cost overruns, as shown in the sharp increase in the bottom left plot in Figure 3. Here we see that a delay of 34 days or more results in a significantly larger likelihood of cost overruns.

Also, a larger duration of the execution, as captured in the variable *Start-to-end execution days* indicates a higher chance of cost overruns, reflecting a possible greater risk involved in larger projects that involve longer contracts. This can be observed in the increasing shape of the bottom center plot in Figure 3, where we observed a particularly large increase in likelihood for execution lengths of over 1 year. Interestingly, the number of days the supplier has been registered in the platform, reflected in the variable *Days supplier registered* — bottom right plot in Figure 3, is also a key predictor, as a shorter time since registration increases the likelihood of cost overruns. This can signal a greater risk when contracting young firms with possibly little experience in public procurement. This risk decreases after the company has been registered in the platform for 228 days (about 7.5 months) or more.

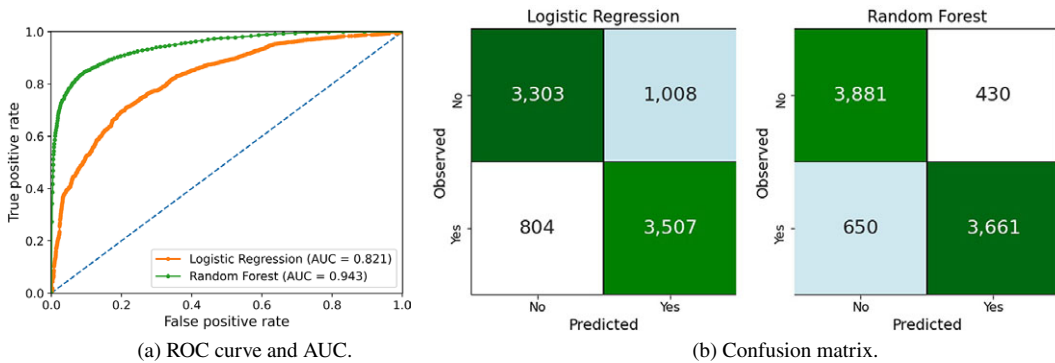


Figure 4. Results for logistic regression and random forests to predict cost overruns.

Finally, the categorical variable Sector appears as a key predictor of cost overruns, specifically when it marks contracts in the Culture Sector. We observe that contracts in this sector have a lower chance of requiring cost overruns. This may be due to the nature of the contracts in this sector or as a consequence of better practices in contracting, which is a question requiring further investigation.

With this subset of variables, we train both logistic regression and random forest models, using 70% of the samples for training and leaving 30% for testing. Figure 4 presents the ROC, AUC, and confusion matrices for these models. We observe that the random forest model performs better, with a significantly higher ROC curve and an AUC of 0.943 versus 0.821 by the logistic regression. The confusion matrices reveal a similar result, where logistic regression is able to identify contracts with cost overruns correctly about 3 out of 4 times, while the random forest model is able to do so 6 out of 7 times.

6.2. Models for delivery delays

Another variable indicative of procurement inefficiencies is *Delivery delays*, where the duration of the contract execution is prolonged beyond the terms defined during the procurement process. This may indicate deficiencies in the procurement planning process, delays in execution, and/or lack of oversight from the buyer. Thus, it is relevant for the overseeing entity to identify contracts at risk of delivery delays. To this end, we developed similar models as in the previous section, but replacing the target variable with *Delivery delays*.

Here, again, we performed a selection of the most relevant variables to come up with a compact and explainable model. In this case, we find that the most relevant variables coincide with those in the cost overruns model, except for Sector (Culture). Thus, we end up with six predictors for this model. As before, the random forest model performs best with an AUC of 0.936, very close to the previous model.

6.3. Considering different contract types

As mentioned before, a common practice in Colombian public entities is the use of professional services contracts to procure human resources. These contracts tend to behave rather homogeneously, with relatively low values and fixed duration (typically 1 year) that prevent contract cost overruns or delivery delays. A large fraction of contracts are of this type, as confirmed by Figure 5, which illustrates how professional services contracts are the most common type for a duration of up to a year, while beyond 1 year, other types of contracts become more prevalent. Following these considerations, we developed separate models for the set of professional services contracts and all other contracts, separately. This allows us to obtain models that fit better the behavior of each contract type.

As Table 2 shows, all the performance measures improve when the analysis is restricted to professional services contracts. For instance, the accuracy slightly increases from 87.47% to 88.29%. The performance for non-professional services is instead worse than the overall case, with an accuracy of 79.16%. This can also be observed in Figure 6, where the ROC curve for non-professional services remains below the one

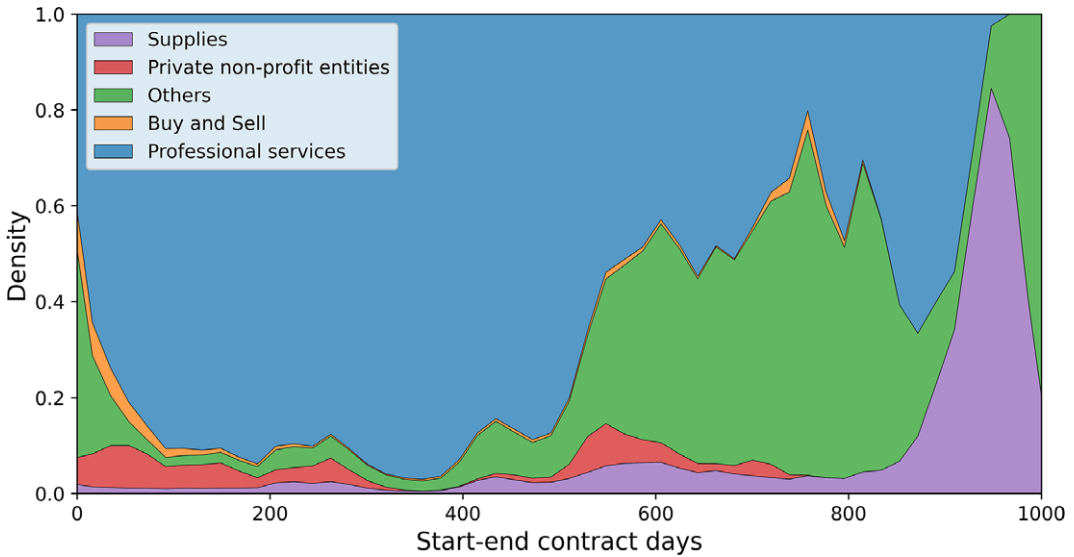
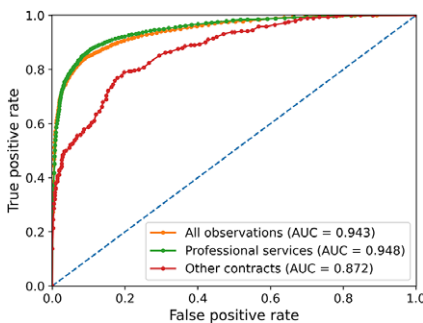


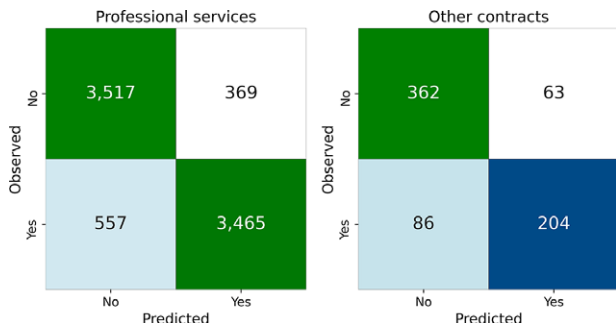
Figure 5. Distribution of the contract duration by type.

Table 2. Random forest performance measures with additions in value as target variable and different contract types

Contract type	Accuracy	Recall	Precision	AUC
All	87.47%	84.92%	89.49%	0.943
Professional services	88.29%	86.15%	90.38%	0.948
Not professional services	79.16%	70.34%	76.40%	0.872



(a) ROC curve and AUC.



(b) Confusion matrix.

Figure 6. Results for the random forests model to predict additions in value for different contract types.

for professional services and the overall one. The confusion matrices also capture this effect, which is likely related to the fact that non-professional services involve a wide set of procurement objectives and mechanisms, making it more difficult for the model to predict correctly whether a contract will present inefficiencies or not. This decrease in performance occurs even though the models that exclude professional services include three additional features after selection: the number of invited suppliers, duration of the advertisement period, and an indicator of products in the administrative services group.

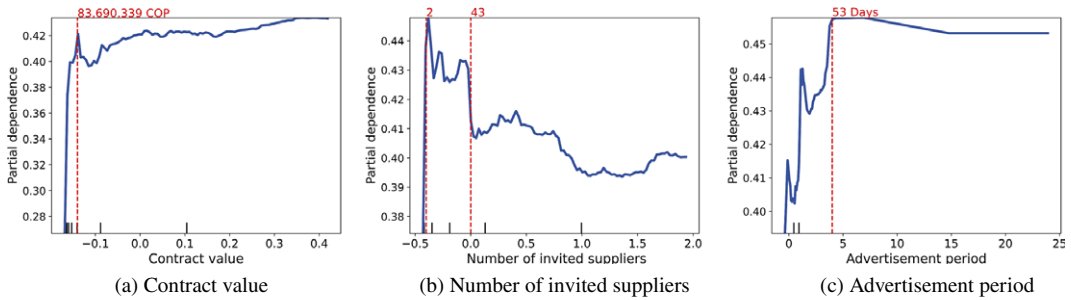


Figure 7. Partial dependency plots of explanatory variables in the Random forest model to predict cost overruns for non-professional services contracts. Variables in the x axis are standardized, and red dashed lines indicate tendency breakpoints with the value in the original scale.

While adding more variables would increase the model's performance, we decided to limit the number of variables to ensure the model remains easy to interpret for the investigators.

As this model includes additional variables, we study their partial dependency plots. Figure 7b shows how a very small number of invited suppliers increases the likelihood of cost overruns for non-professional services contracts. We also see that a significant decrease in this likelihood is observed for a number of suppliers above 40. For the advertisement period, we observe in Figure 7c that the likelihood of cost overruns increases with the duration of this period, and contracts with an advertisement period over 50 days display the largest likelihood. Finally, we would like to highlight the behavior against the contract value variable, displayed in Figure 7a, compared to the model for all the contracts, displayed in Figure 3a. While the trend is similar, here we observed a much larger cutoff value (83 vs 23 million COP) to mark contracts with a high likelihood of having cost overruns. This means that the models are able to capture these different behaviors caused by the different contract types, highlighting the value of training models for these different contract types.

6.4. Models for early intervention

One final consideration is the time at which the overseeing process is performed. Of special interest is the ability to raise early alarms on procurement processes that may lead to contracts with inefficiencies. To achieve this goal, we developed models where the set of variables is limited to those available during the procurement process up to the contract signing. We refer to this period as pre-execution. With this smaller set of variables it is harder to make accurate predictions, as they hold much less information than the full set of variables. However, the results of these models are particularly valuable for an overseeing entity as it can decide to direct early efforts toward contracts likely to result in cost overruns or delivery delays.

Here we start with 30 pre-execution variables on which we apply the selection method based on random forests to obtain a small set of features. In this case, however, the standard method lets only three variables pass through: Value, Days supplier registered, and Sector (Culture). While this small number of features makes the model easy to interpret, the performance suffers as all measures (accuracy, recall, precision, AUC) drop below 70%. We thus choose to modify the selection threshold to allow for more variables to be included in the model. After some tuning, we select a threshold of 1.3 times the median of the importance, i.e., only those variables with an importance score 30% above the median are selected. This choice allows us to improve the model's performance by including 5 additional variables after selection: Number of invited suppliers, Sector (Health), Procurement procedure justification, Product/Service category (Administrative services), and Product/Service category (Non-administrative services).

Figure 8 shows how the ROC curve decreases when using the pre-execution variables only, compared to the full set of variables. The accuracy and recall, in this case, are 76.06% and 76.2%, respectively, which is confirmed by the confusion matrix, where we observe that the model has similar false positive and false negative rates, and is able to predict correctly whether a contract will have cost overruns or delivery delays

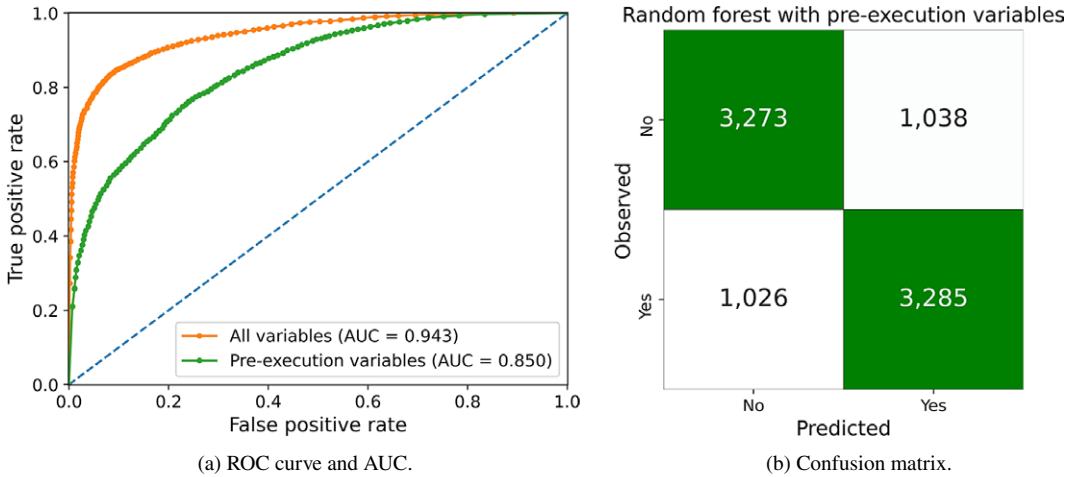


Figure 8. Results for the random forest models for additions in value using all variables vs pre-execution variables.

3 out of 4 times. From the partial dependency plots (see [supplementary material](#)) we also observe that contracts in the Health Sector have a larger likelihood of inefficiencies, as well as those in the Administrative service category. Instead, contracts that require a Procurement procedure justification and those in the Non-Administrative services category have a smaller likelihood of resulting in inefficiencies.

6.5. Inefficiency indices - IRIC

The previous models give us a set of tools to raise red flags to detect contracts with a high risk of inefficiency. In turn, the IRIC and IRICP indices provide metrics to measure the risk of irregularities in the procurement process and offer a tool to compare across entities and territories in terms of public procurement. Hence, we calculated the IRIC and IRICP for 18,796 contracts, where 16,681 are professional services and 1115 are of other types, signed in the year 2020 by public entities in Bogotá. The mean IRIC by buyer shows an aggregate measure that can be updated with each contract and provides a metric to track the probity of the entity. Likewise, the weighted version, IRICP, allows us to monitor the procurement process but accounts for the effect of the contract value magnitude to take special care of expensive contracts.

Table 3 shows that in Bogotá on average professional services contracts are more likely to have irregularities than other types of contracts. Also, most contracts display an IRIC between 0.27 and 0.36, which means that contracts usually exhibit three to four of the 11 red flags defined in Section 5.2. Additionally, Tables 4 and 5 show how IRIC aggregations allow us to compare between entities. We found

Table 3. Descriptive statistics of IRIC and IRICP, by contract type, for contracts signed in 2020

Index	count	mean	std	min	25%	50%	75%	max
Professional services								
IRIC	16,681	0.354	0.074	0.000	0.273	0.364	0.364	0.818
IRICP	16,681	0.339	0.073	0.000	0.270	0.342	0.367	0.930
Non professional services								
IRIC	2115	0.320	0.092	0.000	0.273	0.273	0.364	0.727
IRICP	2115	0.296	0.102	0.000	0.230	0.253	0.364	0.768

Table 4. Top entities by average IRIC for professional services contracts in 2020

Buyer	IRIC	IRICP
IDIGER	0.452	0.440
Alcaldía Local de Puente Aranda	0.426	0.403
Alcaldía Local de Kennedy	0.422	0.409
Cuerpo Oficial de Bomberos de Bogotá	0.411	0.385
Unidad de Rehabilitación y Mantenimiento Vial	0.410	0.398

Table 5. Top entities by average IRIC for non-professional services contracts in 2020

Buyer	IRIC	IRICP
Defensoría del Espacio Público	0.636	0.768
Alcaldía Local de Ciudad Bolívar	0.636	0.743
Terminal de Transporte S.A	0.636	0.681
Alcaldía Local de los Mártires	0.545	0.609
Secretaría General de la Alcaldía de Bogotá	0.466	0.441

that the buyer with the worst index on average in 2020 in Bogotá across professional services contracts was IDIGER,¹¹ which exhibits the highest IRIC, and similar values of IRICP. On the other hand, across non-professional services contracts the riskiest buyer was the Defensoría del Espacio Público¹² as it displays the highest IRIC and an even higher IRICP. We also note that the highest IRIC and IRICP values are much higher for non-professional services contracts than for professional services. This emphasizes the differences among these types of contracts and the need to have models and indices that take these differences into account.

7. Discussion and conclusion

The models, indices, and results presented in this paper originate from a tool developed for an overseeing agency with the aim of supporting their resource allocation decisions and the prioritization of investigations. There are, however, different alternatives for developing such tools. We have made a number of choices that we summarize here.

First, while many potential data sources could be identified, we decided to rely *only* on Open Data sources. While this may be limiting, as other studies have made use of more detailed information, we prioritized having continuous access to data to update the models without requiring large efforts from the overseeing agency. Naturally, a better quality of the published data, in terms of quantity, completeness, cleanliness, and documentation, would favor the development of better (more complete and accurate) models, but the availability of these sources as Open Data is key to the tool development and its future maintenance.

Second, we decided to lean towards explainable models, giving particular importance to the selection and use of as few variables as possible. We found this to be particularly relevant to develop tools for overseeing agencies, as their decisions face ample scrutiny. The explainability requirement is challenging due to the trade-off with accuracy. This led us to sacrifice accuracy in order to guarantee that the number of variables remained limited. Having explainable models was in fact important for us to identify the key

¹¹ IDIGER is the District Institute of Risk Management and Climate Change of Bogotá.

¹² Defensoría del Espacio Público is the public entity in charge of the defense, inspection, surveillance, regulation and control of the public space of Bogotá.

variables that can be used as flags by the investigators, such as the supplier experience or the duration of certain steps in the process.

From a technical point of view, the project also leaves an important lesson in terms of precision, false positives, and resource optimization. When training machine learning models like the ones developed here, researchers have the ability to tune the parameters of interest based on the number of false positives and negatives they are willing to tolerate. A very aggressive model, capable of detecting many true positives at the cost of detecting many false positives, can cause authorities to waste resources inspecting contracts that are not actually problematic. As audit costs are significant and resources are scarce, it is important to find the sweet spot between true and false positive detection.

Finally, we would like to highlight the contributions of this article to the development of machine learning-based models and irregularities indices as holistic complementary tools for the oversight of procurement processes. The development of these tools has also led us to identify key indicators of inefficiency in public procurement that are easy to interpret for domain experts without a background in machine learning.

The aforementioned design choices serve also as a baseline for future work. The Open Data employed focuses on features of the procurement process, which are readily available at the national level and could be found in other datasets, such as those from the European Union Tenders Electronic Daily (TED) portal, the US SAM data bank, and CanadaBuys, among others. Future work may thus test the scalability of the proposed methods to these and other contexts, and tailor the solution according to the specific data available and the local characteristics of the procurement process.

Also, regarding the features employed in the models and indices, the relationship between the presence of missing data and the outcomes should be explored. Although the presence of missing data is widely used in the risk index literature (Fazekas et al., 2016; Fazekas and Kocsis, 2020; Zuleta et al., 2019), the quantification of the relationship between missing values and the outcomes of risk indices and inefficiency models remains an open question. Furthermore, our research identified relevant variables related to the provider, but their construction is limited by data availability. Thus, future work could explore the inclusion of supplier and entity network measures, for instance, through graph analysis (Van Erven et al., 2017; Liu et al., 2016).

Finally, another avenue for future work is to measure the impact of this or similar tools on investigators' productivity, especially on their ability to promptly identify inefficient and irregular public contracts.

8. Practitioner notes

Multiple authors have nourished the public policy literature on the problem of predicting corruption and waste of resources in public procurement, providing a variety of alternative tools. These are diverse in i) level of aggregation, from the larger ones at the municipality level down to the more detailed ones at the contract level; ii) techniques used, including objective corruption risk indices and a variety of machine learning models; and iii) data sources, covering different sectors and variables. There are, however, some gaps in the literature that we address in this study:

- First, existing literature focuses mostly on the prediction and monitoring of active waste, while overlooking passive waste. This prioritizes corruption cases as detected by the judicial systems, but leaves aside inefficiencies caused by lack of skills, incentives, or the presence of political and regulatory obstacles. We tackle this gap by proposing both an index based on Open Data to detect active waste in terms of contractual irregularities, as well as machine learning models to detect passive waste in terms of inefficiencies. This holistic approach is able to better consider both active and passive waste in public procurement.
- Second, the literature has relied on data labeled by human decision-makers, such as Comptroller's Offices and judicial systems, which fall for the so-called "selection labeling problem". As a result, the models derived from such data carry a bias toward visible corruption but miss undetected cases. To address this gap, this study relies on objective measures of inefficiencies and irregularities

obtained from the procurement process itself, such as delivery delays or cost overruns. This enables the identification of cases beyond those previously detected by other decision-makers.

- Third, the machine learning models developed in previous studies generally use black-box models and/or large numbers of variables to reach good performance. This creates barriers among personnel non-specialized in such techniques, hindering the tools' adoption. Our approach bridges this gap by leaning towards explainable models, with a careful selection of as few variables as possible. While this may be limiting, it simplifies reproducibility, which is key given the ample scrutiny that overseeing agencies face, thus easing adoption and even shedding light on key red flags to consider in the procurement process.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/dap.2024.83>.

Data availability statement. The data that support the findings of this study are openly available in VigIA: data and artifacts at <https://zenodo.org/records/10699508> reference number 10699508. These data contain a cleaned and processed version of data extracted from the Colombian Open Data Portal <https://www.datos.gov.co/> on 02/04/2021.

Acknowledgments. We would like to thank CAF-Development Bank of Latin America for supporting the development of this project and *Veeduría Distrital de Bogotá* for the support and feedback during the implementation. In particular, we thank María Isabel Mejía and Santiago Rodríguez for their coordination efforts. In addition, we would like to thank the team that made possible the development of the VigIA web application: Miguel Valencia, Juan Camilo Ruiz, María Jose Prada, and Santiago Lozano. The findings, interpretations, and conclusions expressed in this article do not necessarily reflect the views of the Inter-American Development Bank.

Author contribution. Conceptualization: J.P., J.G.; Data analysis: A.S., J.P., J.G.; Data curation: A.S., J.P., J.G.; Methodology: A.S., J.P., J.G.; Writing: A.S., J.P., J.G.

Funding statement. This work was supported by the CAF-Development Bank of Latin America. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

- Ash E, Galletta S and Giommoni T (2021) A machine learning approach to analyze and support anti-corruption policy. *Available at SSRN*.
- Baltrunaite A, Giorgiantonio C, Mocetti S and Orlando T (2020) Discretion and supplier selection in public procurement. *The Journal of Law, Economics, and Organization* 37(1), 134–166.
- Bandiera O, Prat A and Valletti T (2009) Active and passive waste in government spending: evidence from a policy experiment. *American Economic Review* 99 (4): 1278–1308.
- Bishop CM (2006) *Pattern Recognition and Machine Learning*, 1st Edn. New York, USA: Springer.
- Blumenstock J, Cadamuro G and On R (2015) Predicting poverty and wealth from mobile phone metadata. *Science* 350(6264), 1073–1076.
- Chandler D, Levitt SD and List JA (2011) Predicting and preventing shootings among at-risk youth. *American Economic Review* 101(3), 288–92.
- Charron N, Dahlström C, Fazekas M and Lapuente V (2017) Careers, connections, and corruption risks: Investigating the impact of bureaucratic meritocracy on public procurement processes. *The Journal of Politics* 79(1), 89–104.
- Colonnelli E, Gallego J and Prem M (2022) Chapter 16: What Predicts Corruption?. In *A Modern Guide to the Economics of Crime*. Cheltenham: Edward Elgar Publishing, pp. 345–373.
- De Blasio G, D'ignazio A and Letta M (2022) Gotham city. predicting 'corrupted' municipalities with machine learning. *Technological Forecasting and Social Change* 184, 122016.
- De Blasio G, D'ignazio A and Letta M (2020) Predicting corruption crimes with machine learning. a study for the italian municipalities. Working Papers 16/20, Sapienza University of Rome, DISS.
- Decarolis F and Giorgiantonio C (2022) Corruption red flags in public procurement: new evidence from italian calls for tenders. *EPJ Data Science* 11(1), 16.
- Fazekas M and Kocsis G (2020) Uncovering high-level corruption: cross-national objective corruption risk indicators using public procurement data. *British Journal of Political Science* 50(1), 155–164.

- Fazekas M and Wachs J** (2020) Corruption and the network structure of public contracting markets across government change. *Politics and Governance*, Cogitatio Press 8(2), 153–166.
- Fazekas M, Tóth IJ and King LP** (2016) An objective corruption risk index using public procurement data. *European Journal on Criminal Policy and Research* 22, 369–397.
- Gallego J, Gonzalo R and Martínez J** (2021) Preventing rather than punishing: An early warning model of malfeasance in public procurement. *International Journal of Forecasting* 37(3), 360–77.
- Gnaldi M, Del Sarto S, Falcone M and Troia M** (2021) *Measuring Corruption*. Cham: Springer International Publishing, pp. 43–71.
- Goodfellow I, Bengio Y and Courville A** (2016) *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>.
- Hastie T, Tibshirani R and Friedman J** (2009) *The Elements of Statistical Learning*. New York: Springer.
- Hossain M, Mullally C and Asadullah MN** (2019) Alternatives to calorie-based indicators of food security: An application of machine learning methods. *Food Policy*, 84, 77–91.
- IMCO** (2018) Índice de riesgos de corrupción: El sistema mexicano de contrataciones públicas. Technical Report, IMCO, Instituto Mexicano para la Competitividad, A.C, México.
- James G, Witten D, Hastie T and Tibshirani R** (2013) *An Introduction to Statistical Learning*, 2nd Edn. New York, USA: Springer.
- Jungblut M and Jungblut J** (2022) Do organizational differences matter for the use of social media by public organizations? A computational analysis of the way the german police use twitter for external communication. *Public Administration* 100(4), 821–840.
- Kenny C and Musatova M** (2010) Red flags of corruption. In *World Bank Projects: An Analysis of Infrastructure Contracts*. Policy Research Working Paper 5243, World Bank, Washington, DC.
- Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J and Mullainathan S** (2018) Human decisions and machine predictions. *The Quarterly Journal of Economics* 133(1), 237–293.
- Kleinberg J, Ludwig J, Mullainathan S and Obermeyer Z** (2015) Prediction policy problems. *American Economic Review* 105 (5), 491–495.
- Lakkaraju H, Kleinberg J, Leskovec J, Ludwig J and Mullainathan S** (2017) The selective labels problem: evaluating algorithmic predictions in the presence of unobservables. In *KDD: Proceedings. International Conference on Knowledge Discovery & Data Mining*, pp. 275–284.
- Lima MSM and Delen D** (2020) Predicting and explaining corruption across countries: a machine learning approach. *Government Information Quarterly* 37(1), 101407.
- Liu J, Bier E, Wilson A, Gómez G, Alexis J, Honda T, Sricharan K, Gilpin L and Davies D** (2016) Graph analysis for detecting fraud, waste, and abuse in health-care data. *AI Magazine*. 37, 33–46. [10.1609/aimag.v37i2.2630](https://doi.org/10.1609/aimag.v37i2.2630)
- Lopez-Iturriaga F and Sanz I** (2018) Predicting public corruption with neural networks: an analysis of spanish provinces. *Social Indicators Research* 140, 975–998.
- Polley EC, Rose S and Van der Laan MJ** (2011) Super learning. *Targeted Learning: Causal Inference for Observational and Experimental Data*, New York, NY: Springer. 43–66.
- Rockoff JE, Jacob BA, Kane TJ and Staiger DO** (2011) Can you recognize an effective teacher when you recruit one? *Education Finance and Policy* 6(1), 43–74.
- Rodríguez Arévalo S et al.** (2021) Predicción de ineficiencias en la contratación pública de bogotá. Master's thesis.
- Rodríguez-García G** (2022) Measuring the risk of corruption in latin american political parties. de jure analysis of institutions. *Data & Policy* 4, e42.
- Szucs F** (2023) Discretion and favoritism in public procurement. *Journal of the European Economic Association* 22(1), 117–160.
- Van Erven G, Holanda M and Carvalho R** (2017) Detecting evidence of fraud in the Brazilian government using graph databases. In *Recent Advances in Information Systems and Technologies*. Cham, Switzerland: Springer International Publishing 217, pp. 464–473
- Wenzelburger G, König PD, Felfeli J and Achtziger A** (2022) Algorithms in the public sector. Why context matters. *Public Administration*, n/a(n/a).
- Zuleta MM, Ospina S and Caro CA** (2019) Índice de riesgo de corrupción en el sistema de compra pública colombiano a partir de una metodología desarrollada por el instituto mexicano para la competitividad. Technical Report 67, Fedesarrollo, Bogotá, Colombia. Laboratorio Latinoamericano de Políticas de Probabilidad y Transparencia. Un Proyecto de Cooperación Sur-Sur ATN O/C 16465-RG.
- Zumaya M, Guerrero R, Islas E, Pineda O, Gershenson C, Iñiguez G and Pineda C** (2021) *Identifying Tax Evasion in Mexico with Tools from Network Science and Machine Learning*. Cham: Springer International Publishing, pp. 89–113.

Cite this article: Salazar A, Pérez JF and Gallego J (2024). VigIA: prioritizing public procurement oversight with machine learning models and risk indices. *Data & Policy*, 6: e75. doi:[10.1017/dap.2024.83](https://doi.org/10.1017/dap.2024.83)