Data Analytics for Anti-Corruption in Public Procurement

Viktoriia Poltoratskaia
Central European University
10 PUBLICATIONS **0** CITATIONS

Mihaly Fazekas
Central European University
166 PUBLICATIONS **2,331** CITATIONS

Data Analytics for Anti-Corruption in Public Procurement

Viktoriia Poltoratskaia* and Mihály Fazekas*
*Central European University and Government Transparency Institute

Published in Sope Williams and Jessica Tillipman (eds.) Routledge Handbook of Public Procurement Corruption. Chapter 4. Routledge.

1. Abstract

This chapter describes and analyses the ways in which data analytics can be used for countering corruption in public procurement. It also discusses the main challenges for effective data-driven anti-corruption in public procurement. We propose a simple conceptual framework, which distinguishes petty and grand corruption. Depending on the type of corruption, different accountability mechanisms should operate for successfully implementing data-driven anti-corruption. We describe the main challenges and limitations of available datasets, including data quality, scope, depth, and accessibility, as well as most recent measurement approaches. We provide some empirical examples of successful implementation of data analytics for countering corruption in public procurement. We conclude by summarizing the lessons learnt and focus on practical steps to make data analytics more effective for anti-corruption.

2. Introduction

The rapid spread of open data and advanced quantitative techniques for anti-corruption analysis is widely perceived as a watershed, providing researchers and civil society with a vast number of analytical tools amenable to monitoring governments comprehensively (Harrison et al., 2012, Lima&Delen 2020, Park&Kim 2020). Yet the use of data, including big data, has many limitations and pitfalls that must be taken into account and matched by appropriate methodology and realistic expectations of what information the data can provide in regards to corruption strategies and red flags.

One of the most prominent examples of open data providing researchers, civil society, and the general public with the opportunity to investigate and assess corruption is public procurement data. On the one hand, this prominent role of public procurement in anti-corruption analytics has been fuelled by the unparalleled transparency, detail, and accessibility of public procurement data, in spite of it being far from perfect as we will see below. On the other hand, public procurement is among the most corrupt government activities (OECD 2016) with one of the greatest societal impact. Corruption in the sector is fuelled by large amounts of public money, technical and legal complexity and a great degree of official discretion. Given the widespread use of public procurement data for studying corruption, this chapter is aimed at answering two questions:

a. What are the most promising and feasible uses of data analytics for anti-corruption in public procurement?
b. How can data analytics best support effective anti-corruption in public procurement?

To answer these questions, we propose a simple conceptual framework distinguishing grand and petty corruption in public procurement. This distinction leads to different accountability mechanisms, reflecting on the actors and scope of corrupt activities. It is more straightforward to fight petty corruption with the help of data analytics which contributes to civil society holding governments to account (vertical accountability) and governmental oversight bodies monitoring public spending. Nevertheless, grand corruption is harder to counter, even when it can be precisely identified through advanced data analytics and data tools. These challenges are particularly applicable in public procurement where grand corruption is more prevalent than petty corruption.

The question of data analytics effectiveness in combating corruption is a topic actively discussed in the recent literature. Many of these studies aim at reviewing and further developing strategies for improving transparency and better embedding data analytics in various ministries, agencies and civil society actors (Homburg & Snellen 2007, Bănărescu 2015, Máchová & Lnénicka, 2017). Yet other authors focus on mechanisms embedded in the structure of political institutions and power relationships, which establish limitations on the use of data analytics for preventing corruption (Lederman, Loayza&Soares, 2005, Soudijn&Been, 2020). This chapter contributes to the literature by offering a balanced assessment of opportunities and challenges of employing data analytics in the fight against corruption in public procurement. We also offer selected practical examples which show it is possible to overcome challenges to effectively fight corruption.

This chapter is organised as the following, it first develops a conceptual framework by looking at various types of corruption in public procurement, in particular distinguishing petty and grand corruption. Next, we provide an overview of the main challenges the data itself can impose on anti-corruption activities. Data scope, quality, depth, and accessibility can set serious limits to data analytics[1]. We also touch upon some of the most recent innovative measurement approaches to corruption in public procurement, which open up new avenues for

[1] Hereinafter we focus on the features of the dataset to make it suitable for the future analysis, which is different from general discussion of data quality or compliance with the data standards (including OCDS standards). The description of database parameters as well as their application to the analysis was introduced by Cingolani et al, 2015.

corruption detection and the assessment of anti-corruption interventions. Then, we review a few notable examples of implementing data-driven anticorruption. These concrete examples show how efforts to work with data analytics to reduce corruption in public procurement have different impact depending on contextual accountability mechanisms and administrative capacities. Finally, we offer some concluding thoughts, highlight evidence gaps, and offer pointers at the way forward.

3. Conceptual framework: data and accountability

The phenomenon of corruption has various definitions in the literature, focusing on either different aspects of institutional and public harm caused by corruption (Mo 2001, Rothstein 2014, Gründler&Potrafke, 2019, Akkoyunlu, Sule&Ramella 2020), or on the procedural aspects of corruption mechanisms (World Bank 2000, Knack 2007). Defining corruption before moving to measurement and analysis is a necessary first step. In order to tackle corruption one should define its origins and tailor countering measures to target its root causes. Therefore, this section focuses on conceptualising corruption through two major frameworks.

First, corruption can be defined as a principal-agent problem, assuming elected politicians as principals are able to control and hold bureaucrats accountable for their actions (Rose-Ackerman 1975, Moe 2005). In this framework, bureaucrats are seen as agents, performing on behalf of their principals and delivering public goods. The principal-agent problem implies that both agents and principals have different interests, and due to information asymmetry principals cannot be sure that agents indeed behave in their interest. Applying this problem to the concept of corruption, many types of petty corruption can be seen as the examples of principal-agent dilemma (Persson et al. 2013). For example, the supplier’s field engineer bribing a supervising public official in order to get acceptance of low quality construction works.

Alternatively, corruption can be defined as a collective action problem. This approach distinguishes between political elites and the general public, where a large and diffuse citizenry faces obstacles to effectively controlling a well organised group of political office holders. The collective action problem arises because citizens need to get organised in order to pursue their collective interests of controlling government; yet they often fail because individual citizens’ self-interests trump collective goals (Ostrom 2004). For example, all individuals would be better off from making politicians and powerful bureaucrats enforce rigorous competition and value for money principles in public procurement. However, their individual costs of monitoring and holding them to account (e.g. checking complicated tender documents) outweighs individually benefits. Within this framework, corruption arises when those in power have the means and opportunities to exploit their position for private gain to the detriment of the wider society (Mungiu-Pippidi 2013, Marquette&Peiffer 2015).

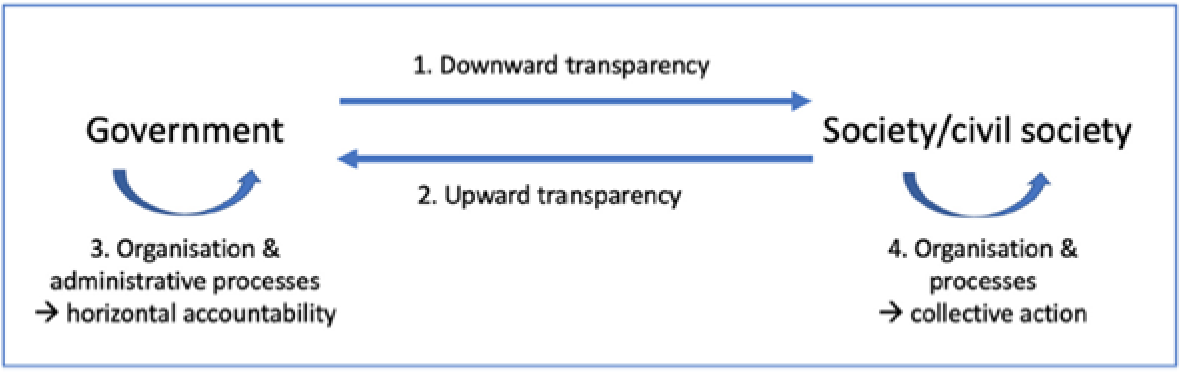
These two theoretical lenses focus our attention on two different types of corruption. Principal-agent problem typically describes petty corruption existing among street level bureaucrats, low-level politicians, and low level employees of suppliers. The inability of politicians to hold procuring officials to account can lead to conspiracy between buyer and supplier to gain illegal profit. Yet, this approach does not provide suitable explanation for grand corruption among a closed elite or state capture. This is where understanding corruption as a collective action dilemma describes the phenomenon of corruption better, providing a framework for effective solutions. Here, the wider public lacks the ability to hold politcians and top bureaucrats to account, opening the door for collusion between public and private elites. Nevertheless, these 2 models are not exclusive, rather they should be seen as 2 complementary lenses useful understanding complex corrupt phenomena (Marquette&Peiffer 2018).

3.1. Accountability mechanisms: actors, information flows, powers

The accountability mechanisms that can potentially help resolving the problems of corruption are different depending on the type of corruption described above. These different mechanisms can be differentiated according to three elements: the participating actors, their powers and information flows between them (i.e. data and data analytics). While actors can be grouped in many ways, one widely used distinction is between citizens and government (i.e. elected officials and bureaucrats). The actors can be seen in a more diverse optic within each group. For example, within the government there are politicians and bureaucrats at different levels (federal vs municipal), as well as those who are directly accountable to their voters (elected politicians) and those who are appointed or nominated by other actors (e.g. prime minister appointed by the parliament or a state secretary appointed by the minister). Similarly, citizens can be more or less involved in holding governments to account depending on their roles, whether it is civil society, journalists, business groups or interest groups.

Considering information flows between actors, data and data analytics can help reduce corruption in two ways: through downward and upward transparency (Heald 2006a, Hood, 2006, Adam&Fazekas, 2021) (Figure 1). First, downward transparency implies that government activities become more open to citizens. In the case of public procurement, the introduction of electronic procurement platforms - which collect data about processes, actors and outcomes - provides NGOs and researchers with a tool to hold public buyers to account (Bauhr et al, 2020). Such measures can include detection of red flags, identification of potentially risky buyers, unreliable suppliers, as well as territories prone to higher corruption risks. Second, upward transparency can solve the problem of information asymmetry in a similar manner, but the opposite direction: government agencies can more easily get feedback and information from citizens about the performance of officials discharging their public duties. For example, law enforcement agencies tasked with investigating and sanctioning corruption can act on media investigations or information reported by civil society (Lagunes, 2021). Therefore, understanding these accountability mechanisms enables us to design effective anticorruption interventions in public procurement resolving information asymmetries between governments and citizens (Heald 2012). In this framework, data analytics and increasing public procurement transparency are key tools for fighting corruption (Köbis et al, 2022).

Figure 1. Four impact mechanisms through which data and e-procurement systems can have an effect on corruption


1. Downward transparency
Government
Society/civil society
2. Upward transparency
3. Organisation & administrative processes → horizontal accountability
4. Organisation & processes → collective action

Source: adapted from Adam&Fazekas (2021)

Yet, increasing transparency *between* government and citizens may not remedy corruption. It is necessary to target and reshape power structures within government (horizontal accountability) as well as within society (e.g. supporting societal collective action) (Figure 1). For these purposes, data and data analytics can be useful too. First, e-government tools and the data they generate can help standardising and automating administrative processes and activities of the government (Fazekas&Blum, 2021). In turn, these decrease the scope of officials' discretion and reduce the direct contacts between citizens and bureaucrats. Limiting official discretion makes corruption harder to conduct; while fewer contacts between officials and individuals make coordination and collusion among the corrupt harder to achieve (Jiménez, Hanoteau&Barkemeyer 2022). Second, societal collective action can be supported by data analytics through provision of information via different platforms such as watchdog portals, therefore creating opportunities for mass mobilization against corruption which can further lead to organised anti-corruption movement (Rodrigues 2014, Bauhr 2017).

While some progress can be achieved in fighting grand corruption with data analytics, its capacities are limited and less straightforward in comparison to its effectiveness against petty corruption (Adam&Fazekas, 2021). As public procurement is often driven by grand corruption, applying data analytics to public procurement corruption problems is often challenging. While data and increased transparency can help improving accountability, some conditions need to be met. This is what the next section discusses in detail.

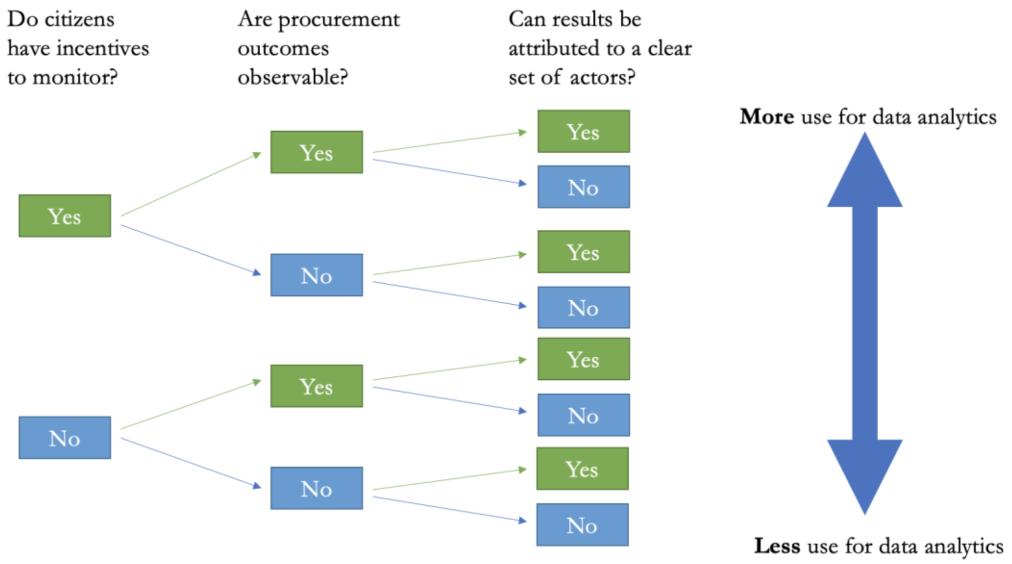
3.2. Preconditions for data-based accountability

Preconditions for data-driven accountability play out differently for the 4 mechanisms above. We start by looking at accountability relationships between citizens and governments (downward/upward transparency). In order for data to become an effective instrument for accountability between governments and citizens, three preconditions should be met:

- Motivation of actors,
- Observability of outcomes, and
- Attributability of results.

First, people using open data should be motivated to monitor public actors and procurement outcomes. Depending on what effect the tender has on their lives, they will be more or less motivated to invest time and effort into monitoring government. For example, parents of students attending a public school are most likely motivated to check the quality of school meals or school equipment procured. However, the quality of complex defence procurement will have little impact on citizens' lives (in peace times at least) lowering their motivations for monitoring. Data and data analytics can have a profound impact on such motivations. E-procurement platforms generate a wide array of detailed data on public procurement processes and outcomes. This information can be used to highlight impact and relevance of procurement spending to various stakeholder groups such as public service users. It can be impactful in highlighting the costs of corruption (Abdou et al, 2022) and the wide-ranging impacts of bad investments (financial, societal), potentially motivating actors to control corruption better.

Figure 2. Schematic representation of the relationship between procurement attributes and the strength of accountability between governments and citizens


Do citizens have incentives to monitor?
Are procurement outcomes observable?
Can results be attributed to a clear set of actors?
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
More use for data analytics
Less use for data analytics

Source: Authors, adapted from World Bank (2016)

Second, observability of outcomes is needed for any meaningful control of government. Observing public procurement outcomes can be particularly challenging as it is often characterised by high degrees of technical and legislative complexity. This is especially true for high value, high impact projects such as highway construction or government IT infrastructure. However, a wide class of public investment projects only show their adverse outcomes years later. For example, a road whose foundation is insufficiently stable may serve drivers well for some years while eventually it deteriorates beyond repair. Data analytics in public procurement can be powerful in uncovering hidden patterns both in terms of corruption relationships, procedural biases and showing insufficient outcomes early on[2]. For example, procurement legislation often requires long advertisement periods but allow for exceptions for shortening these time periods. Data analytics make it possible to spot patterns consistent with exploiting exceptional rules to the benefit of connected suppliers. In turn, these details can be used to develop tailored risk indicators pointing citizens, even without detailed legal knowledge to risky tenders. More broadly, automatically generated risk factors or red flags which can be applied to large-scale procurement dataset can improve anti-corruption targeting.

Third, corruption can only be effectively countered if corrupt acts and their detrimental impacts can be attributed to particular politicians or bureaucrats responsible and/or illicitly benefitting. Once the responsibility for corruption is established, stakeholders can act on this information. For example, voters can vote corrupt politicians out of office or companies can relocate to less corrupt and better run regions. Determinants of corruption and identifying responsible actors are hard and often intractable due to complexity of procurement processes and institutional arrangements. Data and data analytics can help point at reasons for corruption and responsible actors. In some cases, even in the absence of democratic political regime and fair elections, "effective technocrats" are needed to pursue authoritarian modernisation through "pockets of efficiency" (Geddes 1994, Evans 1998). As a result, in societies with low accountability by definition of their political regime some sectors can still end up technologically developed with data-driven processes due to interest of ruling elites in effective policymaking[3] (Dobrolyubova et al. 2017).

Next, we look at horizontal accountability (i.e. corruption controls among government agencies). Here, given some degree of anti-corruption motivation of actors, data analytics can offer tools and the framework beneficial for agencies. As will be later shown in the case of Ukrainian e-procurement platform ProZorro, the first ministries to support the reform were those interested in spending less. Due to a highly corrupt and closed system prior to the reform, government agencies had to spend significantly more in the absence of the open competition,

[2] Importantly, most of the existing data analytics in the field of corruption cannot detect the corruption directly due to its hidden nature. Yet, it can point at weak spots prone to corrupt manipulation (Ponti et al, 2021). For example, detecting an unusually short period of time between publication of call for tender and submission of bids can signal potential wrongdoing – the buyer wants a specific company to win the tender, leaving other potential competitors unaware or unprepared for such a short notice. Whether this indeed results in a corrupt exchange (e.g. public officials getting kickbacks for abusing their power and favouring the supplier) has to be proven by investigative bodies.

[3] In contrast to popular belief, autocrats can be interested in effective policies increasing the wealth of the population or improving the quality of public services as these can provide more opportunities for ruling elites to extract rents (see Wright 2008, Pepinsky 2020).

as well as waste a lot of time on redundant bureaucracy. Moreover, data analytics can dramatically lower the cost of government agencies monitoring each other, for example checking compliance with the legislation can be partially automatised (e.g. compliance with time limits for advertising tenders) or certain abuses of legislation can be pinpointed (e.g. slicing contracts to avoid the regulations applicable over a certain threshold). However the reasons for particular non-compliance and abuse of powers require in-depth analysis such as interviews with stakeholders and mapping of regulations, where data analytics is less useful. Nevertheless, attributing wrongdoing to particular actors can sometimes aided by data analytics in public procurement. For example, corruption prevention commissions existing in many countries are quite often aimed at monitoring politicians and other bureaucrats through asset declarations, various official registers and whistleblowing. Data analytics and general openness of data can help such agencies to tackle corrupt officials more easily.

Finally, we review the preconditions of data analytics improving accountability relationships within society. As outlined above, data and data analytics can help citizens overcome collective action barriers, for example by helping them realize their collective loss (motivation) or targeting scarce collective energies (targeting). However, sufficient data literacy of a considerable portion of the society is a necessary precondition for any such beneficial effects. If 1 or 2 civil society organisation does a superb job workign with public procurement data, but they fail to mobilize a large group of citizens due to lack of skills, the corrupt will msot likely go unpunished.

4. Challenges and limitations of data analytics for anti-corruption

The main challenges of effective data analytics for anti-corruption can be divided into two groups: data-related and measurement-related. Each group of factors can significantly impact the quality and precision of analytical results and their practical uses.

When it comes to data, analytics can be and often is limited by data scope, depth and quality as well as by data accessibility. This means that the data available for analysis does not comprehensively and reliably capture the transactions and actors of interest, for example some high corruption risk contracts are not recorded in the database or key information is incorrect in the database. The amenability of public procurement datasets for corruption assessment is determined by technical issues associated with storing and systemising such data (Fazekas&Sanchez, 2021). This is particularly relevant when it comes to cross-country comparisons due to differences, for example in legally defined reporting requirements. This can also be a problem within one country, as legal requirements might change over time, limiting possibilities for time-series analysis. Therefore before starting the analysis, public procurement data should be checked and whenever possible corrected.

When it comes to measuring corruption, there are two approaches (which are not mutually exclusive): data-driven and theory-driven. The main difference is the starting point of the research. This can be roughly defined by whether it is theoretical expectations which are tested using the appropriate data and methodology, or it is the available data that leads to discovery and guides research questions, the focus of interest and the eventual analysis. Theory-driven research can help developing a broader understanding of the studied questions by filling existing knowledge gaps or testing hypotheses. Whereas, data-driven approaches can unearth less straightforward mechanisms and provide researchers with valuable insights straight from the data, including those which might have been overlooked or not paid attention to previously.

4.1. Data

The following section describes the four dimensions of datasets (scope, depth, quality, and accessibility), through which public procurement datasets can be assessed. These dimensions directly relate to the features of any dataset used for analysis. 1) Data scope refers to the rows of the data table: how many of them are present and how well they cover the population of observations targeted by measurement. 2) Data depth is synonymous to granularity or detail, and refers to the number of variables or columns present in a database. In an ideal scenario, all releant features of an observation are captured in structured columsn which are relevant for analytical goals. 3) Data quality corresponds to the content of the cells in the table and assesses how accurately and reliably the information is presented in the dataset vis a vis the actual phenomena. 4) Data accessibility expresses whether the database can be easily collected and prepared for analysis due to the way it stored on its original source (e.g. html page or structured json file).

4.1.1. Data scope

Does an existing database capture all relevant transactions? Sufficiently good data scope implies that the data covers relevant procurement markets, as well as all the relevant buyers, suppliers and their transactions such as contracts and tenders. There can be a significant variation in the information that each country is making publicly available. For example, availability of tenders per market can differ significantly across countries. In some cases such markets as military supplies can be restricted from publishing, therefore it is very hard to monitor and check for corruption risks. The simplest metrics of data scope is the share of published public contracts value (based on official publication portals) in total public procurement spending (based on budget data) (Table 1)[4]. Additionally, countries differ in the threshold for tender price that requires publication of the call for tender. Unreasonably high thresholds significantly reduce the access to information and do not allow to monitor corruption risks in such tenders.

The presence or absence of information by itself can be used as a red flag, and considered as a signal of corruption risks. Yet it can be potentially misleading too, as with the example of

[4] For full details of methodology see Basdevant&Fazekas, 2022

unpublished bidders. Are they not published in order to prevent researchers and general public from demanding accountability? Or is data missing because of the lack of information storage capacities, deficiency in staff maintaining the website, variety of local regulations of storing information about tenders or any other technical reasons? Going deeper into the field can help answering such questions through analysis of legal requirements while what matters ultimately is the implementation of rules hence the data published. Crucially, the lack of transparency requirements can be the result of corruption directly.

Table 1. Data scope for selected countries

ISO code	Year	Procurement Spending (GTI), Billion Int. USD	Procurement Spending (Budget), Billion Int. USD	Ratio
AT	2020	37.5	68.7	54.6%
BE	2020	18.1	88.6	20.4%
BG	2020	8.1	14.4	56.3%
CY	2020	0.4	2.6	15.4%
DE	2020	53.5	719.8	7.4%
DK	2020	21.1	50	42.2%
EE	2020	5.4	7.2	75.0%
FI	2020	16	50.7	31.6%
FR	2020	206	481.5	42.8%
GR	2020	4.9	38.7	12.7%
HR	2020	3.6	14.8	24.3%
HU	2020	17.1	47.1	36.3%
ID	2020	45.9	333.2	13.8%
IE	2020	14.4	32.3	44.6%
IT	2020	62.1	278.4	22.3%
LV	2020	3.2	7.1	45.1%
NO	2020	16.1	53.1	30.3%
PL	2020	36.6	145.2	25.2%
PT	2020	23.4	34.4	68.0%
SE	2020	21.1	91.6	23.0%
SK	2020	16.8	24.5	68.6%
UG	2020	0.1	10.4	1.0%
UK	2020	428	432.2	99.0%

Source: Basdevan&Fazekas, 2022

4.1.2. Data depth

Do we have enough detailed information on recorded cases? By answering this question one can establish the limitations of data analysis regarding the details of corrupt transactions and their different types and forms. For example, some public procurement systems publish information about bidders, such as number and names of bidders per tender, amount of each bid, winning bid and in some cases the reason for winning, while in some countries this information is either unavailable or published only for certain types of procedures.

Moreover, a public procurement database can contain information on buyer and supplier names but lack information on their legal address, or tax IDs. This can limit developing certain indicators, as well as significantly complicate linking public procurement data to other datasets such as company ownership, asset declaration or politically exposed persons. Without persistent tax IDs, it is hard to track organisational performance, while linking public procurement data to other datasets could greatly increase corruption risk measurement accuracy. Moreover, the absence of time series data stored in the same structure and format will prevent the analysis of trends and finding external shocks or event influencing corruption risks. For example, after COVID-19 breakdown and introduction of the state of emergency certain markets became more prone to corruption risks (in particular medical supplies and COVID-related products, e.g. masks, ventilators, etc). Yet without consistent time series data it is impossible to observe if the change in corruption risks was caused by a state of emergency or something else.

Using unique IDs to link public procurement data to external datasets can prove to be powerful in enriching analytical results as well as validating risk indicators. For instance, Decarolis & Giorgiantonio (2022) links detailed public procurement data from 2 sources to corruption investigations data on suppliers' owners and managers. Linking such rich and detailed indicators allowed them to develop new red flags of corruption as well as validating a wide set of indicators using reliable investigative evidence.

4.1.3. Data quality

Are the data reliable and complete (Liu et al. 2016)? The simplest way to check this is look at the missing rate of key variables such a contract values and verify whether related publications are indeed published. For example, absence of call for tenders in public procurement processes is a key quality problem. This can be checked through the date of the call for tender publication and its ID reported in related publications such as the contract award. If the call for tenders cannot be found on the government publication portal, it most likely indicates serious data quality deficiency. Yet in some cases absence of data can happen due to data collection or processing error - something that is present on the procurement portal can be missing from the dataset used for the analysis. In some cases, governments themselves offer different versions of the data with different content, for example a structured data dump with fewer variables and a full html publication mage with more complete information. The most difficult type of data quality problem is when the information is present in the right format but it does not correspond to actual actions and behaviours. For example, if the contract value written in the contract award announcement differs from the contract value in the eventually signed contract between buyer and supplier.

4.1.4. Data accessibility

Can the public procurement dataset be accessed easily and reliably for quantitative analysis? Answering this question requires checking the data can be accessed exactly, i.e. if the data is in a machine-readable format or not, and in case it is - what type of machine readability is applicable. If the data stored as scanned pdfs or jpegs - it is very hard to transform it into an analysable dataset. Pdfs and word docs with standardised structure and possibility for machine reading the content are more amenable to machine processing and eventual data analysis. HTML pages[5], as well as Aplication Programmign Interfaces (APIs[6]) usually enable more flexibility and easier access to the data, yet html content may vary greatly hence it may be very challenging to work with. A typical problem is when the html pages follow national publication standards of great variety with some countries using over 30 different standard forms for reporting largely similar information (Czibik et al, 2015).

Some countries have websites providing access to public procurement data in a readily downloadable format such as data dumps. Such sources need to be verified before analysis and compared to the official websites. For instance, checking the number of observations by procurement type can be informative and easy to perform even without downloading the dataset.

In some cases a country does not have a single, centralised public procurement platform. Therefore if the data is scraped from one of the national sources, it can be incomplete and only cover tenders from certain markets, over a certain threshold, or regulated by one of many public procurement laws. Another possible issue is non-official source of public procurement data, which can be potentially not regularly updated or not verified. Such shortcomings can be observed through basic descriptive statistics. A very uneven annual distribution of observations, absence of certain markets, thresholds different from ones specified by national regulations can be a signal of incomplete data.

[5] HTML is a language through which the information is stored and displayed on the webpage. HTML can either be well-structured (meaning that the same type of information has the same HTML path and therefore can be easily scraped through the script), or less standardized (e.g. each title of the tender has different HTML path depending on other elements of the webpage), therefore more problematic to download.

[6] API is an interface which establishes standard that is usually written for users to approach certain information on the webpage. In other words, if the user wants to access certain information on the webpage, there is an established and standardized mechanism of communication (API) between the software used to parse the data (e.g. Python) and the webpage itself.

4.2. New approaches for detecting corruption

Measuring a variety of corrupt phenomena has always been a very challenging task due to the hidden nature of corruption. Yet, in recent years there is a growing number of indicators developed to gauge corruption using large-scale, publicly available datasets, most notably in public procurement. As more and more countries began to publish detailed procurement data, opportunities opened for researchers to develop micro-level corruption indicators instead of focusing on cross-country aggregated scores and comparisons. These developments also enabled the tailoring of indicators to more specific fraudulent activities or corruption mechanisms.

An increasing number of models for detecting fraud and corruption in public procurement use historical data to predict the risk of wrongdoing, rather than actual measuring instances of corruption. Such models include logistic regressions, decision trees, clustering, neural networks and various other semi-supervised or unsupervised machine learning techniques (Modrušan et al. 2021). Yet, most models still need meticulous adjustments to adopting the most precise and valid indicators for the selected time period or context (e.g. country).

In many cases, even with newly developed and more comprehensive methodologies, limitations imposed by data quality remain. In particular, there is little possibility to resolve a problem of missing data without losing its reliability (i.e. some options like replacing missing values with average values can influence the overall representativeness of the variable). Similarly, data errors can be easily detected by statistical methods (e.g. simple distribution graph can show abnormal outliers), yet not much can be done besides the suggested solutions for missing values. In some cases new approaches and methods can help resolving this issue through finding such abnormalities and attempting to make sense of them.

One of the main challenges in defining corruption using data is to distinguish between "normal" and "abnormal" behaviours (Yang&Wu 2020). Increased use of big data allowed researchers to apply more comprehensive methodological approaches, such as machine learning techniques. For instance, Yang&Wu (2020) used dynamic unsupervised learning method to identify "suspicious" behaviour and risky transactions by looking at historical behaviour of the customer and searching for abnormalities. Naturally, machine learning techniques are highly dependent on the availability of accurate labels in the training data (i.e. whether the sample is representative and whether there a large-enough and reliable sample of clean and corrupt cases). Some techniques, such as clustering can be used as unsupervised methods and therefore do not require a training dataset, as well as pre-developed knowledge on abnormal transactions. Therefore the very definition of anomaly such as fraud is coming not from the expectations of researcher, but directly from the data. Nevertheless, most methods rely on theoretically sound understandings of corrupt deals based on qualitative case studies to look for specific high risk patterns in large scale datasets (Fazekas & Kocsis, 2020).

By using network analysis on micro-level procurement data (in particular contractual relationships between organisations), it is possible to identify not only whether there is a corrupt relationship, but to reveal if corruption has become systemic, leading to state capture. In a network perspective, state capture is defined as the clustering of high corruption risk connections; where corrupt behaviour is the norm and corrupt actors are capable of collective action in pursuance of their group goals (Fazekas&Tóth 2016). Furthermore, linking open and standardised beneficial ownership data with public procurement datasets provides multiple opportunities for enhancing the analysis of corruption, conflicts of interest, cartels and state capture.

5. Practical examples of data-driven anticorruption

When done right, data analytics can be useful for supporting anticorruption in two broad areas: i) to support investigations on the contract, organisation or market levels, and ii) to analyse policy reform and support policy evaluations.

To support concrete investigations in public procurement, first, data can be used for flagging new cases to investigate where corruption is more likely to hide. For example, based on comparison to average lengths of advertisement periods (the time interval between publishing the call for tenders and the deadline), tenders with too short advertisement periods can be selected for further investigation. Second, data analytics can help ranking a longer list of known cases. For example, when an investigative body receives a large number of whistleblower reports but it is uncertain which ones are worthy of investigation (i.e. likely to lead to a successful conviction) risk indicators can help ranking cases. A range of public procurement corruption risk indicators such as single bidding can be matched to the reported cases so that most risky ones are selected. Third, data analytics can help conduct the investigation itself. Once an entity is selected for investigation, there might be a need for identifying its most risky or relevant transactions or partners for further data collection and analysis. Corruption risk indicators (red flags) can help investigators zoom in on the parts of the case which are most likely to be corrupt and data analytics can support the identification of related high-risk actors and transactions (i.e. defining the boundaries of the investigation).

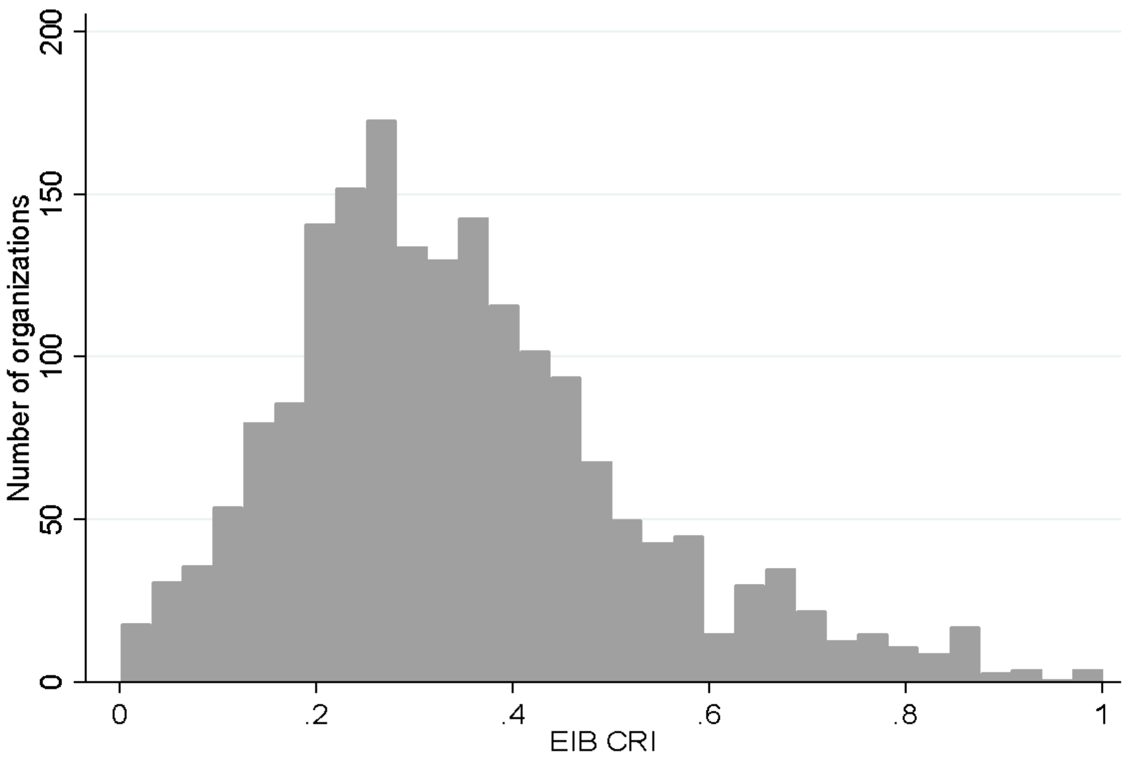
Data-driven approaches can be employed to analyse policy reform and contribute to policy evaluations. First, public procurement data systems containing data from different government data providers often vary in quality and suffer from a range of data quality problems, some of which are hard to identify. Data analytics can be useful for pointing out such data gaps, supporting policy interventions remedying them. As high quality public procurement datasets as essential or a range of governmental analytical tasks from budgeting to improving value for money, such reforms are likely to have wide ranging benefits. Second, public procurement regulations define the ways in which public purchases are done and by extension influence their outcomes. A range of regulatory features and their modifications, such as contract value thresholds for open procedures, are well suited for quantitative impact evaluations, establishing whether regulatory goals are met. Third, data analytics can also help understanding whether different organisational rules and controls are well suited to control corruption. For example, the different assignment of tender evaluation responsibilities or contact implementation rights have a profound impact on corruption in public procurement. These different impacts can be uncovered through data analytics, supporting better organisational design.

These diverse uses of data analytics in public procurement are demonstrated below with the help of a small number of concrete case studies.

5.1. European Investment Bank

European Investment Bank finances various EU projects of over 56-60$ billion annually (European Investment Bank, 2021). The allocated loans are then typically distributed through procurement process, resulting in a large number of contracts. In order to audit these projects and counterparts, the Fraud Investigation Division of EIB regularly conducts risk-based audits, so-called Prior Integrity Reviews, which are aimed at mitigating risks before financial losses occur. To identify counterparts to investigate, corruption risks in public procurement are assessed among many other factors. This methodology builds on the Corruption Risk Indicator (CRI) (following Fazekas&Kocsis, 2020) to rank organisations according to their risk (OECD 2019). The CRI methodology first identifies valid individual red flags, second it combines individual indicators into a composite score. One of the main red flags used, which is a direct measure of the tender's competitiveness, is single bidding (whether the tender received one bid or multiple bids). Other indicators' validity is tested through their correlations with single bidding while controlling for structural features such as main sector or year of contract award. For example, by analyzing the time period between publication of the call for tender and deadline for submitting bids, one can see correlation between short time period and low competitiveness (the shorter is the submission period, the lower is the number of bids received per tender). Thus, the new indicator can be added assigning high risk to short submission periods, and low to the longer ones. For the EIB analysis the following risk indicators were identified as valid: procedure type (open or not), call for tender (published or not), length of advertisement period, product description length, elgibility criteria lenght, length of decision period, single bidding, tax haven registration of supplier, and supplie/buyer capture. The final CRI score takes averages from each component and constructs a scale from 0 to 1 where 1 is a high-risk organization and 0 is a low risk.

Figure 3. Distribution of organisations contracted by EIB based on corruption risk in public procurement



Subsequently, the highest risk cases (right side of Figure 3) are further investigated using desk research, for example by looking at media reports. The shortlist of organisations, which shows high risks in both quantitative and qualitative assessment, are then selected for on-site audits by the EIB proactive integrity team (OECD 2019). This analysis is made possible by identifying over 500 000 government contracts of EIB counterparts encompassing those contracts which are directly financed by EIB but also which are funded from other sources, offering a comprehensive risk overview. The underlying public procurement data is gathered from publicly available official public procurement records across the EU. Within the above conceptual framework, data analytics applied by EIB supporting audits represents an example of horizontal accountability supporting accountability relationships among different governmental bodies. Greatly improving the targeting of monitoring by EIB, counterparts incentives to follow the rules and avoid corruption are strengthened.

5.2. Prozorro

Another example of preventing corruption using open data and increasing accountability is the establishment of the electronic procurement system ProZorro in Ukraine. The system rests on the cooperation among business, civil society and government which is a great example of both downward and upward transparency. Introduction of the platform was not only aimed at collecting the data and making it publicly accessible, but first and foremost at reforming the public procurement system in Ukraine in general, making it decentralised and transparent. It started from a few state agencies and ministries volunteering to test the platform (Nizhnikau 2022). As a result, more and more stakeholders became involved, spreading a word about high saving rates per contract. Similarly, more suppliers started using the system hoping to avoid previously high corruption risks. Gradually, the platform won political support too, becoming one of the most successful anti-corruption reforms in modern Ukraine (Nizhnikau 2022).

The establishment of ProZorro further boosted the development of monitoring mechanisms and amendments to the public procurement law in Ukraine, contributing to the overall reform of the public procurement system in the country which started in 2014. For example, in 2017 new amendments were introduced to the law regulating tendering procedures in Ukraine. The amendments specified automated monitoring tools and risk indicators, developed in collaboration between ProZorro officials and civil society (Nizhnikau 2022). ProZorro is a notable example of multiple efforts coming from both civil society and government to tackle corruption through increased transparency and automated data sharing. Interestingly, once the platformed proved its effectiveness against petty corruption, it became possible to adjust overall public procurement regulation and reduce grand corruption through more transparent criteria and monitoring mechanisms.

5.2. Opentender. eu

The watchdog portal opentender.eu allows governments, companies, NGOs, and the general public to easily access essential information as well as performance indicators on public procurement tender in 33 European countries plus the European Commission itself. It is unique in its scope, depth and wide set of users. The platform publishes data collected from official government data sources after processing and cleaning them, resulting in a single standardised and structured data structure. In most countries covered by the portal public procurement information is not readily downloadable, so additional work is required to scrape numerous html pages and transforming the data into a structured database. Besides the provision of open access to the data, as well as possibility to download it, opentender.eu provides and visualises transparency and integrity indicators. It not only offers an overview of countries, regions and markets but also allow users to drill down to individual suppliers and buyers observing their performance. Integrity indicators include procedure type (open or not), length of advertisement period (how long was the period for submitting the bids), decision period (in how many days the winning bidder selected), call for tender (whether it was published or not), single bidding (how competitive was the tender), new company (whether supplier is a newly established organization) and tax heaven (if the address of the supplier is registered in one of the countries considered as tax haven).

The users of the platform are diverse and come from a range of backgrounds. Some users utilize the market analysis functionalities to get a better insight into market structure, budding opportunities and the openness of competition. Typically, these insights are useful for companies considering entering a new market which they are interested in but don’t know in depth. This exemplifies the uses of data for improving upward accountability whereby more bidding firms chck up on potentially corrupt competitors. Another set of users look at integrity indicators in particular and use the portal for identifying and investigating potential corrupt transactions and corruption organisations. For these users, the portal allows for searching by risk level and ranking organisations or transactions by risk. Hence, for these users opentender.eu improves accountability mechanisms by making investigations better targeted, quite similarly to the above EIB example.

Similar to ProZorro, Opentender provides a possibility for users to assess the overall corruption risks in public procurement of a certain country or region. Due to the high level of aggregation, data analytics can also assist powerful actors in tackling grand corruption through analysing and comparing general trends and differences in corruption risks.

6. Lessons learnt and the way forward

As it was shown, data analytics has great promise yet considerable limitations when it comes to countering corruption. While it can be helpful and effective for combating petty as well as grand corruption , albeit the latter tends to be more challenging. To make a better use of data analytics for anti-corruption, one should first identify which type of corruption should be targeted. Is it oligarchic public-private relations or individual dishonest suppliers? A clear goal-setting is the foundation for determining the which impact mechanisms can be leveraged to counter corruption and hence how best to tailor the data analytics tool and its applications.

For using data to prevent petty corruption, the main task is to solve information asymmetry and increase accountability between agents and principals. While tackling grand corruption requires wider ranging changes - reshaping power dynamics, attracting people beyond civil society actors and journalists to follow up on data-driven insights and get engaged in anti-corruption activism. These tasks consist of many closely coupled steps which can be facilitated by data analytics but also require a range of conditions to be met.

One of the most important drivers for successful use of data analytics for countering corruption in public procurement is the participation of relevant stakeholders. This can be achieved through various means, including user-friendly and easily accessible analytics platforms running on open data, as well as clear perspective on benefits that each side can gain from combating corruption. As the ProZorro example demonstrates, the implementation of open data reforms and analytics can achieve wide ranging success even in a rigid patrimonial political structure once the gains are clear for all the sides and benefits outweigh the costs.

Data analytics should also be easily incorporated into the daily routine of public agencies. Manually investigating thousands of cases does not only limit the effectiveness of auditing, but requires large, often unavailable, resources from the monitoring agency. Applying well-tested and valid risk indicators to all observed cases and selecting entities with the highest risks can be incorporated into standard law enforcement procedures, as the case of EIB prior-integrity reviews show. Combining quantitative risk assessment with qualitative information from media reports, for example, can be combined into a powerful risk assessment pipeline.

Finally, data should not be considered as a universal remedy for corruption, both due to the complexity of phenomena and the wide-ranging challenges of profound reforms. In addition, data analytics for anti-corruption in public procurement is often limited by data scope, depth, quality, and accessibility. Many of these challenges can be overcome at least partially in order to produce widely useful watchdog portals as the opentender.eu portal demonstrates. The same dataset and comprehensive performance indicators can feed into and support decision making by a diversity of stakeholders ranging from bidding firms to investigators. The improvement of open datasets and quantitative methods is expected to continue in the coming years. Crucially, the increased use of data analytics should be placed in the broader context of societal and administrative accountability mechanisms where it can strengthen existing mechanisms rather than supplant them.

References

1. Abdou, A. ; Basdevant, O.; David-Barrett, E. and Fazekas, M. (2022) Assessing Vulnerabilities to Corruption in Public Procurement and Their Price Impact. IMF Working Papers: WP/22/94. Washington, D.C.: IMF.
2. Abdou, Aly, Ágnes Czibik, Bence Tóth, and Mihály Fazekas. "COVID-19 emergency public procurement in Romania: Corruption risks and market behavior." Government Transparency Institute 3 (2021): 1-6.
3. Adam, I. and Fazekas, M., 2021. Are emerging technologies helping win the fight against corruption? A review of the state of evidence. *Information Economics and Policy, 57,* p.100950
4. Akkoyunlu, Sule, and Debora Ramella. "Corruption and economic development." Journal of Economic Development 45, no. 2 (2020): 63-94.
5. Bănărescu, A., 2015. Detecting and preventing fraud with data analytics. *Procedia economics and finance, 32,* pp.1827-1836
6. Basdevant, O. and Fazekas, M., 2022. An Online Tool to Assess Corruption Risks in Public Procurement: The Corruption Cost Tracker (CCT). Technical Guidance Note, International Monetary Fund, Washington DC.
7. Bauhr, M. and Grimes, M., 2017. Transparency to curb corruption? Concepts, measures and empirical merit. *Crime, Law and Social Change*, *68*(4), pp.431-458
8. Bauhr, M., 2017. Need or greed? Conditions for collective action against corruption. *Governance, 30*(4), pp.561-581
9. Bauhr, Monika, Czibik, Agnes, Fazekas, Mihály & de Fine Licht, Jenny, (2020), Lights on the Shadows of Public Procurement. Transparency as an antidote to corruption. Governance. 33(3).
10. Czibik, Á. – Fazekas, M. – Tóth, B. (2015): How to Construct a Public Procurement Database from Administrative Records? GTI-R/2015:02, Budapest: Government Transparency Institute.

11. Decarolis, F., and Giorgiantonio, C. 2022. Corruption red flags in public procurement: new evidence from Italian calls for tenders. EPJ Data Science, 11, 16
12. Dobrolyubova, E., Alexandrov, O. and Yefremov, A., 2017, June. Is Russia ready for digital transformation?. In *International Conference on Digital Transformation and Global Society* (pp. 431-444). Springer, Cham
13. Dufek, L., 2015. Public procurement: a panel data approach. Procedia Economics and Finance, 25, pp.535-542.
14. EIB (2022) European Investment Bank Financial Report 2021. Luxembourg: EIB.
15. Evans, P., 1998. Transferable lessons? Re-examining the institutional prerequisites of East Asian economic policies. *The Journal of Development Studies*, *34*(6), pp.66-86
16. Fazekas, M. and Sanchez, A.H. 2021. Emergency Procurement: The Role of Big Open Data. In S. Arrowsmith, L. Butler, A. L. Chimia and C. Yukins (eds.) Public Procurement in (a) Crisis: global lessons from the COVID-19 pandemic. Hart Publishing. chapter 23.
17. Fazekas, M. and Blum, J.R. 2021. Improving Public Procurement Outcomes : Review of Tools and the State of the Evidence Base. Policy Research Working Paper;No. 9690. World Bank, Washington, DC.
18. Fazekas, M. and Tóth, I.J., 2016. From corruption to state capture: A new analytical framework with empirical applications from Hungary. *Political Research Quarterly*, *69*(2), pp.320-334.
19. Fazekas, Mihály, and Kocsis, Gábor, 2020. Uncovering High-Level Corruption: Cross-National Corruption Proxies Using Public Procurement Data. British Journal of Political Science, 50(1).
20. Geddes, B., 1994. Politician's dilemma. In *Politician's Dilemma*. University of California Press
21. Gründler, K. and Potrafke, N., 2019. Corruption and economic growth: New empirical evidence. European Journal of Political Economy, 60, p.101810.
22. Harrison, TM, Guerrero, S, Burke, GB, et al. 2012. Open government and e-government: democratic challenges from a public value perspective. Information Polity 17(2): 83–97
23. Heald DA(2006a) Varieties of transparency. In: Hood C and Heald DA (eds) Transparency: The Key to Better Governance? Proceedings of the British Academy 135. Oxford: Oxford University Press, 25–43.
24. Heald, D., 2012. Why is transparency about public expenditure so elusive?. *International review of administrative sciences*, *78*(1), pp.30-49.
25. Homburg, V., & Snellen, I. (2007). Will ICTs finally reinvent government? – The mutual shaping of institutions and ICTs. In C. Pollitt, S. van Thiel, & V. Homburg (Eds.), New public management in Europe (pp. 135–148). Palgrave Macmillan.
26. Hood C (2006) Transparency in historical perspective. In: Hood C and Heald DA (eds) Transparency: The Key to Better Governance? Proceedings of the British Academy 135. Oxford: Oxford University Press, 3–23.
27. Jiménez, A., Hanoteau, J. and Barkemeyer, R., 2022. E-procurement and firm corruption to secure public contracts: The moderating role of governance institutions and supranational support. *Journal of Business Research*, *149*, pp.640-650.
28. Knack, S., 2007. Measuring corruption: A critique of indicators in Eastern Europe and Central Asia. *Journal of Public Policy*, *27*(3), pp.255-291
29. Köbis, N., Starke, C. & Rahwan, I. 2022. The promise and perils of using artificial intelligence to fight corruption. *Nature Mach Intelligence*, 4, pp.418–424.
30. Lagunes, P. (2021) The eye and the whip. Corruption Control in the Americas. Oxford University Press, Oxford, UK.

31. Lederman, D., Loayza, N.V. and Soares, R.R., 2005. Accountability and corruption: Political institutions matter. *Economics & politics*, *17*(1), pp.1-35.
32. Lima, M.S.M. and Delen, D., 2020. Predicting and explaining corruption across countries: A machine learning approach. *Government Information Quarterly*, *37*(1), p.101407.
33. Liu, J., Li, J., Li, W. and Wu, J., 2016. Rethinking big data: A review on the data quality and usage issues. *ISPRS journal of photogrammetry and remote sensing*, *115*, pp.134-142.
34. Máchová, R., & Lnénicka, M. (2017). Evaluating the quality of open data portals on the national level. Journal of Theoretical and Applied Electronic Commerce Research, 12, 21–41.
35. Marquette, H. and Peiffer, C., 2015. Corruption and collective action. *DLP Research Paper*.
36. Marquette, H. and Peiffer, C., 2018. Grappling with the "real politics" of systemic corruption: Theoretical debates versus "real world" functions. *Governance*, *31*(3), pp.499-514
37. Mironov, M. and Zhuravskaya, E., 2016. Corruption in procurement and the political cycle in tunneling: Evidence from financial transactions data. *American Economic Journal: Economic Policy*, *8*(2), pp.287-321
38. Mo, P.H., 2001. Corruption and economic growth. Journal of comparative economics, 29(1), pp.66-79.
39. Modrušan, N., Rabuzin, K. and Mršic, L., 2021. Review of Public Procurement Fraud Detection Techniques Powered by Emerging Technologies. *International Journal of Advanced Computer Science and Applications*, *12*(2)
40. Moe, T.M., 2005. Power and political institutions. *Perspectives on politics*, *3*(2), pp.215-233
41. Mungiu-Pippidi, A., 2013. Controlling corruption through collective action. *Journal of Democracy*, *24*(1), pp.101-115
42. Mylovanov, Timofiy. 2019. "Za chotyry roky roboty ProZorro." Censor. net. https://censor.net/ua/news/3167275/za_chotyry_roky_roboty_pro zorro_zaoschadyla_ukrayinskym_platnykam_podatkiv_100_mlrd_ grn_mylovanov
43. Nizhnikau, R., 2022. Love the tender: Prozorro and anti-corruption reforms after the euromaidan revolution. *Problems of Post-Communism*, *69*(2), pp.192-205

44. OECD (2019) Analytica for Integrity: Data driven Approaches for Enhancing Corruption and Fraud-Risk Assessment.
45. Ostrom, E., 2004. *Understanding collective action* (No. 569-2016-39044).
46. Park, C.H. and Kim, K., 2020. E-government as an anti-corruption tool: Panel data analysis across countries. *International Review of Administrative Sciences*, *86*(4), pp.691-707.
47. Pepinsky, T., 2020. Authoritarian innovations: theoretical foundations and practical implications. *Democratization*, *27*(6), pp.1092-1101.
48. Persson, A., Rothstein, B. and Teorell, J., 2013. Why anticorruption reforms fail—systemic corruption as a collective action problem. *Governance*, *26*(3), pp.449-471
49. Ponti, B., Cerrillo-i-Martínez, A. and Di Mascio, F., 2021. Transparency, digitalization and corruption. In *Understanding and Fighting Corruption in Europe* (pp. 97-126). Springer, Cham.
50. Preventing Corruption in Public Procurement, OECD, 2016.

51. Rodrigues, U.M., 2014. Social media's impact on journalism: a study of media's coverage of anti-corruption protests in India. *Global media journal: Australian edition*, *8*(1), pp.1-10
52. Rose-Ackerman, S., 1975. The economics of corruption. *Journal of public economics*, *4*(2), pp.187-203
53. Rothstein, B., 2014. What is the opposite of corruption?. *Third World Quarterly*, *35*(5), pp.737-752
54. Soudijn, M.R.J. and de Been, W.H.J., 2020. Law enforcement and money laundering: Big data is coming. *Criminal defiance in Europe and beyond*, pp.399-426
55. Soylu, A., Corcho, Ó., Elvesæter, B., Badenes-Olmedo, C., Yedro-Martínez, F., Kovacic, M., Posinkovic, M., Medveščček, M., Makgill, I., Taggart, C. and Simperl, E., 2022. Data Quality Barriers for Transparency in Public Procurement. Information, 13(2), p.99.
56. Szakonyi, D., 2018. Businesspeople in elected office: Identifying private benefits from firm-level returns. *American Political Science Review*, *112*(2), pp.322-338
57. Szakonyi, D., 2020. *Politics for profit: business, elections, and policymaking in Russia*. Cambridge University Press
58. World Bank (2000) Anticorruption in Transition: a Contribution to the Policy Debate. Washington DC: World Bank
59. World Bank (2016) World Development Report 2016: Digital Dividends. Washington DC: World Bank.
60. Wright, J., 2008. Do authoritarian institutions constrain? How legislatures affect economic growth and investment. *American Journal of Political Science, 52*(2), pp.322-343.
61. Yang, Y. and Wu, M., 2020, July. Supervised and Unsupervised Learning for Fraud and Money Laundering Detection using Behavior Measuring Distance. In *2020 IEEE 18th International Conference on Industrial Informatics (INDIN)* (Vol. 1, pp. 446-451). IEEE