Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Global Contract-level Public Procurement Dataset

Mihály Fazekas[a,*], Bence Tóth[b], Aly Abdou[c], Ahmed Al-Shaibani[c]

[a] Central European University, Quellenstraße 51, 1100, Wien, Austria
[b] University College London, Gower Street, London WC1E 6BT, UK
[c] Government Transparency Institute, 6000 Kecskemét, Futár u. 48., Hungary

ARTICLE INFO

Article history:
Received 17 February 2024
Revised 20 March 2024
Accepted 10 April 2024
Available online 15 April 2024

Dataset link: GTI Global Public Procurement Dataset (GPPD) 2/2 (Original data)
Dataset link: GTI Global Public Procurement Dataset (GPPD) 1/2 (Original data)

Keywords:
Public procurement
Contracts
Tenders
Anti-corruption, market structure, public spending efficiency

ABSTRACT

One-third of total government spending across the globe goes to public procurement, amounting to about 10 trillion dollars a year. Despite its vast size and crucial importance for economic and political developments, there is a lack of globally comparable data on contract awards and tenders run. To fill this gap, this article introduces the Global Public Procurement Dataset (GPPD). Using web scraping methods, we collected official public procurement data on over 72 million contracts from 42 countries between 2006 and 2021 (time period covered varies by country due to data availability constraints). To overcome the inconsistency of data publishing formats in each country, we standardized the published information to fit a common data standard. For each country, key information is collected on the buyer(s) and supplier(s), geolocation information, product classification, price information, and details of the contracting process such as contract award date or the procedure type followed. GPPD is a contract-level dataset where specific filters are calculated allowing to reduce the dataset to the successfully awarded contracts if needed. We also add several corruption risk indicators and a composite corruption risk index for each contract which allows for an objective assessment of risks and comparison across time, organizations, or countries. The data can be reused to answer research questions dealing with public procurement spending efficiency among others. Using

[*] Corresponding author.
E-mail address: FazekasM@ceu.edu (M. Fazekas).
Social media: @mihaly_fazekas (M. Fazekas)

https://doi.org/10.1016/j.dib.2024.110412

unique organizational identification numbers or organization names allows connecting the data to company registries to study broader topics such as ownership networks.

Specifications Table

Subject	Management, Monitoring, Policy and Law
Specific subject area	Public Policy, Open Data, Public Procurement
Type of data	Tables, csv files Figures
Data collection	Data were scraped and downloaded from the official websites of the national procurement authorities and the EU's Tenders Electronic Daily portal.
Data source location	Primary data sources: the source data on published contracts and corresponding tenders are available from: • EU's Tenders Electronic Daily: https://ted.europa.eu/TED • National Data portals [Data portal Links by country in Annex]
Data accessibility	Repository name: Global Public Procurement Dataset (GPPD) Data identification number: https://doi.org/10.17632/fwzpywbhgw.3, https://doi.org/10.17632/w9mzf4vswh.3, Direct URL to data: https://data.mendeley.com/datasets/fwzpywbhgw/3, https://data.mendeley.com/datasets/w9mzf4vswh/3 Further data updates are made available at: https://www.govtransparency.eu/category/databases/

1. Value of the Data

- A wide set of researchers and policy analysts studying public spending can benefit from this global, standardized, micro-level dataset. It offers rich, contract-level information on where and how governments spend public funds, accounting for about ⅓ of general government spending in the countries covered by the data.
- Academics, governments, and control bodies (e.g. auditors) can use the data to monitor and analyse public procurement across a wide range of countries, including tracking corruption risks.
- Government contracts data can be linked to other datasets increasing its value. For example, it can be linked to company registry data or politicians' asset declarations in order to gain a more comprehensive insight into public spending quality and good governance.
- This dataset adds value to existing macro-level datasets on public spending, especially public investment, by providing comprehensive contract-level information. Micro-level data on the process and outputs of public procurement spending help analyse market dynamics and spending efficiency.

2. Background

Public procurement is a crucial area of public spending as it amounts to about 1/3rd of general government spending across the world [1]. Such spending is worth around 9.5 trillion USD annually [2]. These large amounts are accompanied by high public interest as key infrastructure and services depend on government contracts. Moreover, public procurement faces high corruption risks due to its complexity and high degrees of discretion.

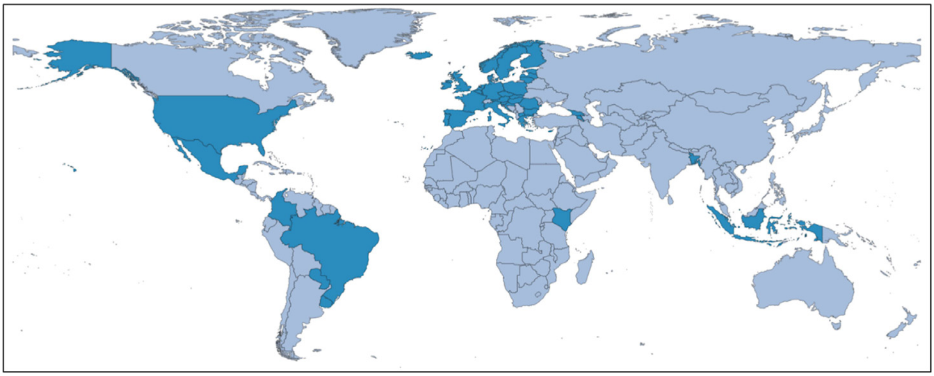

Fig. 1. Countries covered in the Government Transparency Institute Global Public Procurement Dataset. Data is available for countries in dark blue.

Most countries around the world publish large amounts of micro-level information on public procurement. Unfortunately, this information is hard to use because it is usually badly formatted (e.g. individual contracts published in semi-structured html pages) and often fragmented (e.g. information stored in different websites, following formats varying by legal regime) [3]. Such data, as it is typically published by governments, allows for reviewing individual contracts, but limits analysis across large volumes of contracts.

Despite its importance and widely available source data, only a few datasets exist which allow governments, citizens, and researchers to monitor public procurement performance (e.g. [4–6]). These datasets, however, typically include one country and/or sector, lacking the scale and scope our dataset offers.

3. Data Description

Public procurement procedures are highly regulated and tightly structured processes. A typical, open public procurement tender starts with a call for tenders or request for quotations [7]. At this point, the buyer calls for potential suppliers to submit their bids. During the ensuing advertisement period, interested bidders can submit their bids which are evaluated and ranked by the tender evaluation committee composed of officials of the buyer but often including external experts. Then a contract award decision is reached, and a corresponding notice is published in the official gazette. After this, the contract is concluded between the buyer and the supplier. Next, contract implementation takes place. The procurement process is completed by delivering according to the contract or incomplete termination of the contract.

Our dataset includes harmonized public procurement contracts from 42 countries as shown in Fig. 1 (dark blue). The published contracts are mainly from 2006 to 2021 although for some countries we collect data from earlier years depending on data availability and quality. The research team routinely works on adding more recent data and updating the datasets to improve data quality for older tenders if needed (e.g. changes made to older records). Updated datasets published by the Government Transparency Institute at https://www.govtransparency.eu/category/databases/ contain time stamps, indicating the end of data collection period which allows users to track recent additions to the dataset. Updated records can be identified using the variable persistent_id which remains the same across dataset updates. If 2 records in 2 dataset releases with the same persistent_id are different, the record has been updated.

The Global Public Procurement Dataset (GPPD) is comprised of the unfiltered datasets which include the harmonized contract data as well as procurement notices that have failed or are cancelled (i.e. tender information without contract information). The GPPD contains more than

Table 1
Data Description by country.

Country	Years	Observations	Contracts	Buyers	Bidders	Total Value ($ bn)
Armenia	2017 - 2021	204,415	198,466	269	7351	3.4
Austria	2006 - 2021	134,482	73,686	16,692	56,094	71.7
Bangladesh	2011 - 2021	270,180	211,466	4841	29,661	60.2
Belgium	2006 - 2021	188,842	73,003	26,184	53,109	103.2
Bulgaria	2006 - 2021	918,302	272,139	55,628	146,708	150.9
Brazil	2001 - 2021	4942,807	4338,457	5061	77,913	413.5
Colombia	2000 - 2021	3641,726	3639,751	9614	930,045	4153.7
Cyprus	2006 - 2021	19,539	9362	2011	4919	12.4
Czech Republic	2006 - 2021	760,874	301,058	44,928	95,664	331.9
Germany	2006 - 2021	851,024	467,541	87,523	332,911	324.6
Denmark	2006 - 2021	96,250	44,214	12,154	32,770	91.8
Estonia	2006 - 2021	216,053	98,582	8423	25,127	36.1
Spain	2006 - 2021	2779,290	1921,720	430,350	661,439	1582.3
Finland	2006 - 2021	122,276	64,945	10,623	47,733	74.3
France	2005 - 2021	5469,835	1389,859	428,133	1626,042	714.6
Georgia	2010 - 2021	626,785	202,343	3442	28,401	24.5
Greece	2006 - 2021	177,868	62,583	20,090	52,103	65.4
Croatia	2007 - 2021	473,967	249,739	12,742	44,311	65.9
Hungary	2005 - 2021	544,159	214,537	29,574	121,579	279.9
Indonesia	2008 - 2021	3814,693	1070,434	50,444	184,341	533.8
Ireland	2006 - 2021	172,164	18,863	11,525	22,954	28.6
Iceland	2006 - 2021	4413	2038	311	2147	7.0
Italy	2006 - 2021	12,114,318	12,004,113	85,064	2380,891	714.5
Kenya	2009 - 2021	89,612	24,912	463	13,882	0.7
Lithuania	2006 - 2021	458,122	121,626	24,399	23,121	56.0
Luxembourg	2006 - 2021	19,170	9177	2374	10,718	15.0
Latvia	2006 - 2021	492,975	221,913	6623	51,330	82.0
North Macedonia	2000 - 2021	427,237	228,747	3128	128,503	20.7
Malta	2006 - 2021	12,400	5293	1339	2572	7.7
Mexico	2009 - 2021	2094,711	2093,279	5220	279,731	554.6
Netherlands	2006 - 2021	190,128	93,533	24,468	73,975	139.1
Norway	2006 - 2021	379,896	57,857	36,416	56,408	66.4
Poland	2006 - 2021	6541,620	4006,614	130,414	1296,498	1228.3
Portugal	2006 - 2021	2659,390	1282,191	22,692	278,389	139.7
Paraguay	2010 - 2021	785,619	179,842	449	28,379	46.4
Romania	2001 - 2021	1897,636	610,458	42,322	207,135	368.2
Sweden	2001 - 2021	206,214	116,344	15,538	67,309	188.6
Slovenia	2006 - 2021	608,371	244,035	18,940	38,804	68.2
Slovakia	2006 - 2021	541,110	409,619	22,084	83,180	89.5
United Kingdom	2006 - 2021	879,589	356,645	117,510	379,672	742.9
Uruguay	2002 - 2021	1330,397	1139,276	386	48,302	389.2
United States	2007 - 2021	34,447,771	34,447,771	193	120,654	2776.4
Total		**92,606,230**	**72,578,031**	**1830,584**	**10,152,775**	**16,823.9**

72 million contracts from around 1.8 million buyers and more than 10 million suppliers in 42 countries (Table 1). Each contract is concluded between a buyer and supplier which is the most relevant unit of observation in our dataset as it represents commitment to public spending. Nevertheless, the number of observations can be higher than the number of contracts as many tenders lead to no contract award (i.e. failed or cancelled tenders) or administrative records are simply incomplete (e.g. call for tenders are published but no contract award can be linked to it). We also show the total number of unique buyers and bidders in the dataset as these are the key actors concluding transactions (i.e. contracts) with each other. The total contract value represented by the GPPD is more than USD 16.8 trillion representing on average around 1.1 % of global GDP annually. Those contract values are taken into account in this aggregation which are reported for awarded contracts in administrative records of sufficiently high quality (i.e. contract award notices missing the name of the winning bidder are excluded). Table 1 breaks down GPPD by country, based on the government publication portal publishing the information which nearly

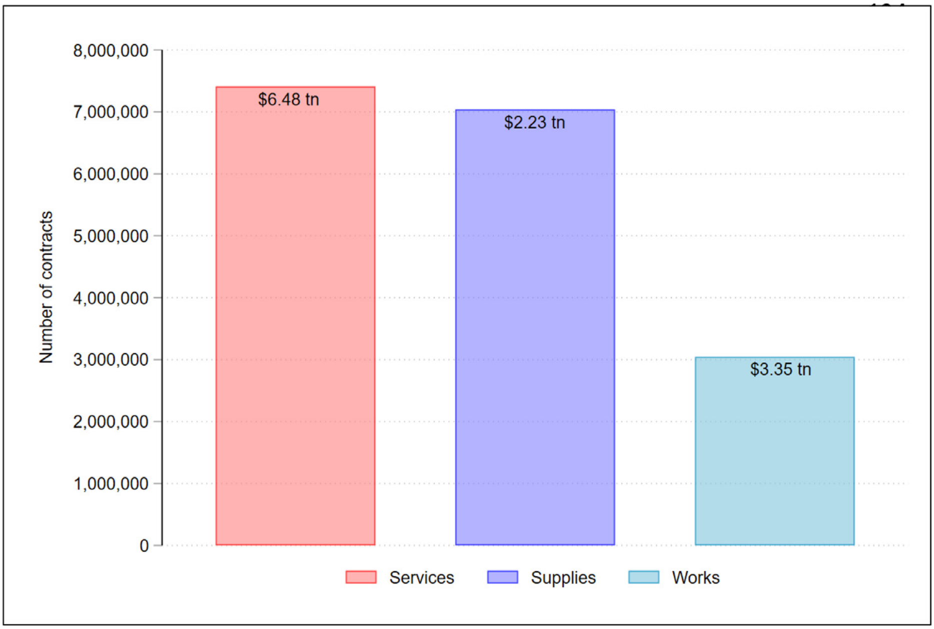

Fig. 2. Total Number of contracts and total contract value by supply type
The total number of contracts is represented on the y-axis whereas the total value of contracts are represented by the numbers inside the bars. The US data is excluded from this figure as the supply type variable is missing from the source data. There are around 11 million contracts missing the supply type in Italy, they are not presented in the figure above for clarity.

always corresponds to the country of contract implementation. The dataset can be downloaded by country from the Mendeley repository.

The dataset is highly diverse in terms of products purchased. The majority of contracts, both in terms of contract numbers and total value, pertain to services, amounting to USD 6.48 trillion. Services encompass diverse economic activities ranging from medical treatments, through waste collection, to education services. Works contracts, although fewer in number, account for a total value of USD 3.35 trillion. Works in public procurement refer to public works or construction works encompassing activities such as road construction, building refurbishment or tree planting. Lastly, supplies contracts represent around USD 2.23 trillion from the total contract value in the dataset (Fig. 2). Supplies, or goods in other words, include ordinary goods such as cars, office supplies, furniture or commodities such as coffee.

The dataset is also diverse in terms of annual country coverage: it spans 33 countries in Europe, 4 in South America, 2 in North America, 2 in Asia, and 1 in Africa. Fig. 3 shows the annual distribution of awarded contracts per continent. Notably, since 2015, the number of published contracts has shown a consistent rise in Europe and North America, peaking at approximately 5 million contracts per year in 2019–2020. The drop in the European contract count in 2021 is due to changing publication practices in some countries (e.g. Italy has switched to a new open data publication format which will have to be retrospectively incorporated in the database during future updates). In contrast, South America has witnessed a steady increase in contract numbers, reaching around 1.5 million contracts by 2021. The data also shows a rise in the number of published contracts, albeit to a lesser extent, in Asia and Africa.

A powerful feature of GPPD is that it makes available a diverse set of contracts, spanning from small transactions such as purchasing rice to larger scale contracting like highway construction. Fig. 4 illustrates the distribution of contract prices (using logarithmic scale) in local currency across the top 10 countries with the highest contract values in the dataset.

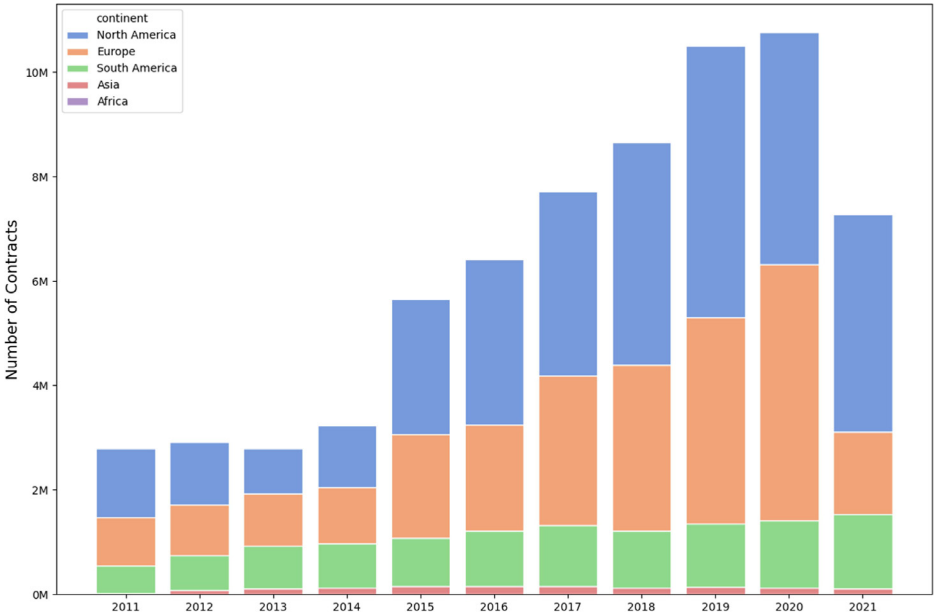

Fig. 3. Annual total number of contracts by continent across time.
Notes: Please note that the drop in the European contract count in 2021 is due to changing publication practices in some countries (e.g. Italy has switched to a new open data publication format which will have to be retrospectively incorporated in the database at future updates).

Some countries display a left-skewed distribution, indicating a higher frequency of lower bid prices, while others exhibit a right-skewed distribution, implying a prevalence of higher-value contracts. Furthermore, certain histograms portray a symmetrical shape closer to a normal distribution. Such differences amply demonstrate different publication practices in different countries driven by regulatory differences among others.

In addition to the variables available on the government sources, we also enrich the dataset by calculating corruption risk indicators. These indicators include single bidding, procedure type risk, publication of call for tender documents, advertisement and decision period risk, tax haven status of the bidder country, and the buyer's spending share. We then compute the composite corruption risk index (CRI) by averaging these indicators for each contract, providing a more encompassing indication of overall risks. While the CRI and its underlying individual risk indicators can be computed at the contract level, aggregating them allows for consistent risk comparisons across countries, markets, time and organizations (buyers and suppliers). We offer a more detailed description of the methodology behind computing the scores in the “Data use” section below.

4. Experimental Design, Materials and Methods

4.1. Data collection

Dataset creation consists of 3 main stages that can be further broken down into smaller steps. First, a data collection stage takes place where the data sources are scraped or downloaded. Second, in the data standardization stage, the collected data is parsed into a standard data structure, and related records are matched with each other. Third, a data validation stage ensues which is composed of several rounds of cross-checking the created dataset against the official source data and corrections if needed.

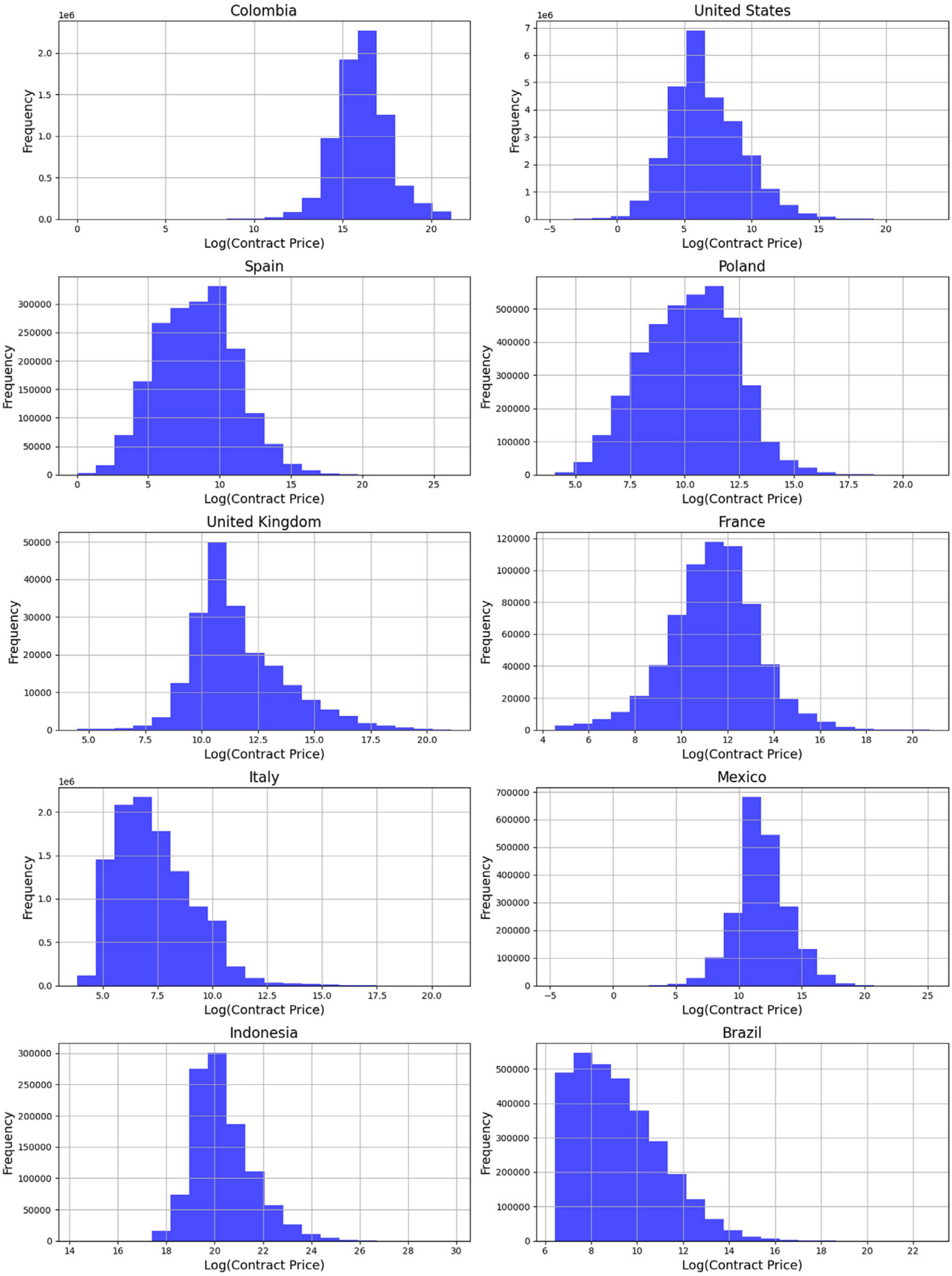

Fig. 4. Histograms of the contract price for largest countries.

First, the data collection stage starts by identifying official government publication portals for public procurement information. This may be an online tender journal with each tender having its own page or a structured database of procurement data (i.e. a data dump or API). Then, an automated web crawler is developed to scrape data from the publication portals of the countries covered in the dataset, with some countries having more than one portal. The full list of data sources can be found in Table A.1.1. HTML, XML, JSON and CSV files are downloaded or scraped from official government sources. Data can be collected only from countries that publish semi-structured online data on their public procurement procedures (i.e. public procurement publications follow more or less standard structure and content defined by legally binding rules).

Second, the data standardization stage includes parsing, cleaning, and mastering steps. During parsing, each publication is transposed from its original format to a uniform structured data template. This process involves mapping each data field with its corresponding values into our data standard. The cleaning step converts structured text to standard data types such as numbers, dates, and enumeration values - e.g. mapping the national procedure types, and supply types to our set of standardized types. The mastering step creates the most accurate and comprehensive record of each public procurement process by combining all publications relating to the same process or tender (e.g. linking the call for tenders to the contract award). To understand this task, recall that the procurement cycle is divided into the following stages: (i) tender preparation and the tendering process when bidders prepare their bids; (ii) bid evaluation, contract award, and contract signing stage; and (iii) contract execution and completion. Therefore, we have to link all the information that describes the same procurement process, from the call for tenders (one or more) to the contract award (one or more), and completed by a series of payments (or a contract completion announcement). We also take into account if any modifications or cancellations occur at any point during the process, i.e. considering changes to information relating to a tender. Once all published publications referring to the same procurement process are linked, the available information can be reconciled to create a single best and most up-to-date image covering the entire procurement process. This includes reconciling conflicting information or filling in empty fields if available in a related notice (e.g. if the buyer address is missing in one publication, it can be filled in from a linked publication that includes that information). The single best image of a tender is organized at the contract level.

Third, the data validation stage follows data collection and standardization. Given the complex processes needed to create the datasets and the diversity of source publication formats and content, we put considerable effort into data verification and correcting uncovered errors in our scraping and parsing algorithms. Data validation starts by drawing a random sample of the data to be cross-checked manually with the publications' official source. This step verifies that our data accurately captures the full set of information published by the government portal. If errors are found a data verification log is created which then is fed back into updating the scraping and parsing algorithms. This stage may imply several rounds of data validation and fixes to ensure that all the annotated data fields on the source are correctly captured in the dataset.

The technical challenges of constructing public procurement datasets are discussed in more detail in [3].

4.2. Data enhancements

Once the dataset as accurately as possible represents the official government source, a series of further data cleaning, standardization and extension steps are taken to improve the data and make it more useful for analysis. These steps include variable transformations, such as setting correct date formats to DDMMYYYY; data enhancements, such as filling in missing cells from other available information; and error corrections, such as dropping implausible values. The main data improvements are the following:

Tender year: Missing values of the tender year are imputed using the contract award publication year, contract award year, or call for tender publication year.

Procedure type enumeration: Based on our research, we harmonize procedure types from various sources to follow a common set of categories. This enumeration is crucial for maintaining data integrity and ensuring that the information is structured and organized for meaningful analysis.

Procurement entities and firm names: Contracts with incorrect organization names which are composed of only non-alphabetic symbols and/or only punctuation marks are removed, as these names are just erroneous data. We also perform basic cleaning steps such as removing any odd characters, back slashes, and extra whitespaces.

NUTS codes: The restructuring of NUTS codes is undertaken to ensure a standardized presentation that includes three hierarchical levels for European countries. This modification involves organizing and formatting NUTS codes to precisely represent the geographical divisions at three distinct levels. The objective is to enhance the clarity and usability of region information, enabling more effective analysis and cross-country comparisons within Europe.

Framework Agreements: To address variations in the publication methods of framework agreements across different countries, a harmonization process was implemented. This involved the creation of a "filter_framework" flag designed to identify the initial contract derived from the respective framework agreement. By employing this flagging mechanism, we establish a standardized approach to highlight and distinguish the primary contract associated with each framework agreement, thereby enhancing the consistency and comparability of this information across diverse contexts.

Price and PPP adjustments: Using the local currency variable, we create the PPP (Purchasing power parity) adjusted price data to allow for cross country comparisons. We use the PPP conversion factor, GDP (LCU per international $) indicator (code: PA.NUS.PPP) available from the World Bank open data portal (https://data.worldbank.org/). This indicator serves as a reliable source for obtaining the necessary exchange rate data, enabling the transformation of raw prices into a standardized metric that reflects the actual purchasing power in international terms.

The cleaned, standardized data contains the variables listed in Table 2.

Following the harmonization of the variables' names and formats, we add several filters to make the data easier to use (Table 3). The user can (i) filter out cancelled contracts using *filter_cancelled*, (ii) filter out observations with missing buyer or bidders name using *filter_buyer|filter_bidder*, (iii) filter out losing bids using *filter_losingbids*, (iv) accurately handle framework agreements using *filter_framework*, (v) filter out data duplicates published in different sources based on the reporting thresholds using *filter_opentender,* (vi) filter out years where data is not reliable using *filter_year*. Finally, we added a combined filter (*filter_ok*) that narrows down the sample of contracts to successfully completed tenders making the identification of the most relevant records for analysis easier.

4.3. Data use: the example of corruption risk assessment

Given the high risk of corruption in public procurement, even in otherwise non-corrupt countries, we also develop and validate context-specific corruption risk indicators following [8]. These corruption risk indicators capture strategies of corruption that are specific to public procurement and detectable with open public procurement data. These strategies represent deviations from principles of open and fair competition in public procurement, thus benefiting connected bidders to the detriment of others. One simple way to approximate the presence of these types of corrupt behaviours is to track the prevalence of single bidding (one bid submitted in a tender) in otherwise competitive markets, as it indicates the exclusion of bidders from competition. Another example is the use of non-competitive tendering conditions for bidders (for example, the selection of non-open procedure types or the shortening of advertising periods) which directly enables the exclusion of non-connected companies. A host of such indicators have been validity tested exploiting co-variation among them as well as against external indicators of corruption coming from surveys and other administrative datasets [8,9]. In addition, they also predict overpricing in public tenders across a wide set of countries [10]. While extensive validity tests are confirmatory, these indicators only capture corruption risks and do not per se signal wrongdoing or deliberate unethical behavior. They help to understand risk trends in public procurement and to point out tenders or markets where further investigation is warranted. The full list of indicators in the dataset is outlined along with conceptual definitions in Table 4.

Table 2
List of columns in country datasets.

Column name	Definition
persistent_id	*Internal persistent tender ID hashed from the earliest publication URL related to the given tender.*
tender_id	*Internal tender ID generated during the data processing.*
tender_title	*Tender title.*
tender_proceduretype	*Procedure type mapped to DIGIWHIST standard. It is based on the original procedure type published on the source publication that we recategorized to a standard enumeration. The DIGIWHIST categories are the following: Open, Restricted, Restricted with publication, Negotiated without publication, Competitive dialog, Design contest, Minitender, DPS purchase, Outright award, Approaching bidders, Public contest, Negotiated, Innovation Partnership, Concession, Other (national type)*
tender_nationalproceduretype	*Procedure type as it is published in the source publication. It contains jurisdiction specific procedure types that might not be possible to relate to the tender_procedureType categories.*
tender_isawarded	*Whether the tender is awarded or not.*
tender_supplytype	*The type of the purchase. It can have the following values: supplies, services, public works.*
tender_biddeadline	*The final deadline until when companies can submit a bid. It is based on the latest call for tender documents published.*
tender_isjointprocurement	*Whether the purchase is a joint procurement (when multiple public bodies purchase something jointly,* e.g. *because of economies of scale)*
tender_lotscount	*Number of lots of a given tender.*
tender_recordedbidscount	*Number of recorded bids - based on unique bids recorded in the source publication,* i.e. *it differs from lot_bidscount.*
tender_isframeworkagreement	*Whether the tender is a framework agreement.*
tender_isdps	*Whether the tender is a dynamic purchasing system (a tendering mode similar to framework agreements).*
tender_contractsignaturedate	*The date of contract signature if the tender only has one lot or all lots have the same signature date.*
tender_cpvs	*List of product codes purchased in the tender. It is based on the Common Procurement Vocabulary (CPV) codes published on the source publication - https://simap.ted.europa.eu/cpv*
tender_maincpv	*Main product code of the tender. It is based on the Common Procurement Vocabulary (CPV) codes published on the source publication - https://simap.ted.europa.eu/cpv*
tender_iseufunded	*Whether the tender has EU funding.*
tender_selectionmethod	*Whether the winning supplier is the lowest priced tender or the most economically advantageous tender ('MEAT'). In case of MEAT, the contracting authorities can qualify their awarding criteria (quality, technical details or sustainability etc.).*
tender_awardcriteria_count	*Number of award criteria used in evaluating the bids.*
tender_cancellationdate	*The date of cancellation of the tender.*
cancellation_reason	*Reason for tender/contract cancellation.*
tender_awarddecisiondate	*The award decision date.*
tender_estimatedprice	*Estimated price of the tender.*
tender_finalprice	*Final price of the tender.*
lot_estimatedprice	*Estimated price of the given lot.*
bid_price	*The bid price.*
tender_corrections_count	*Number of corrections related to the tender.*
lot_row_nr	*Unique lot identifier within a given tender.*
lot_title	*Lot title.*
lot_status	*Whether the lot was awarded*
lot_bidscount	*Total number of bids submitted for a given lot.*
lot_validbidscount	*Total number of valid bids (those that were not excluded) submitted for a given lot.*
lot_electronicbidscount	*Total number of bids submitted by electronic means for a given lot.*
lot_smebidscount	*Total number of bids submitted by SMEs for a given lot.*
lot_updateddurationdays	*Latest duration (in days) of a given lot/contract.*
buyer_id	*Main Identifier of the buyer from the source documents.*

(*continued on next page*)

Table 2 (*continued*)

Column name	Definition
buyer_masterid	*Unique identifier of the buyer assigned during the data processing based on name, source identifiers, address fields. Note that these identifiers are assigned by source, not by country, hence the same company appearing in different data sources is expected to get different identifiers.*
buyer_name	*Name of the buyer.*
buyer_nuts	*Regional code of the buyer. (These are published NUTS codes from the source publication - https://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistics)*
buyer_city	*City of the buyer.*
buyer_country	*Country of the buyer.*
buyer_mainactivities	*Main activity of the buyer. It can have the following values: GENERAL_PUBLIC_SERVICES, SOCIAL_PROTECTION, EDUCATION, HEALTH, ENVIRONMENT, PUBLIC_ORDER_AND_SAFETY, HOUSING_AND_COMMUNITY_AMENITIES, DEFENCE, ECONOMIC_AND_FINANCIAL_AFFAIRS, RECREATION_CULTURE_AND_RELIGION, GAS_AND_HEAT_PRODUCTION, GAS_AND_OIL_EXTRACTION, COAL_AND_OTHER_EXTRACTION, ELECTRICITY, WATER, POSTAL, RAILWAY, URBAN TRANSPORT, PORT, AIRPORT, OTHER, and the national raw terms that could not be categorized.*
buyer_buyertype	*Type of the buyer. It can have the following values: NATIONAL_AUTHORITY, NATIONAL_AGENCY, REGIONAL_AUTHORITY, REGIONAL_AGENCY, PUBLIC_BODY, EUROPEAN_AGENCY, UTILITIES, OTHER.*
buyer_postcode	*Postcode of the buyer.*
buyer_nuts_1	*Buyer's First-level NUTS*
buyer_nuts_2	*Buyer's Second-level NUTS*
buyer_nuts_3	*Buyer's Third-level NUTS*
buyer_street	*Street address of the buyer from the source documents*
buyer_url	*Buyer's website from the source documents*
buyer_email	*Buyer's email from the source documents*
buyer_phone	*Buyer's phone from the source documents*
buyer_contactName	*Buyer's contact person's name from the source documents*
buyer_extra_source_id	*Other Buyer identifiers from the source documents*
buyer_sourceid_type	*Type of other Buyer identifiers from the source documents*
bidder_id	*Main Identifier of the bidder company from the source documents .*
bidder_masterid	*Unique identifier of the bidder company assigned during the data processing based on name, source identifiers, address fields. Note that these identifiers are assigned by source, not by country, hence the same company appearing in different data sources is expected to get different identifiers.*
bidder_name	*Name of the bidder company.*
bidder_nuts	*Regional code of the bidder company. (These are published NUTS codes from the source publication - https://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistic)*
bidder_city	*City of the bidder company.*
bidder_country	*Country of the bidder company.*
bidder_postcode	*Postcode of the bidder company's from the source documents*
bidder_street	*Street address of the bidder company's from the source documents*
bidder_email	*Bidder company's Email from the source documents*
bidder_phone	*Bidder company's phone from the source documents*
bidder_extra_source_id	*Other Bidder company's identifiers from the source documents*
bidder_sourceid_type	*Type of other Bidder company's identifiers from the source documents*
bidder_url	*Bidder company's website from the source documents*
bidder_contactName	*Bidder company contact person's name from the source documents*
bidder_nuts_3	*Bidder company's Third-level NUTS codes*
bidder_nuts_2	*Bidder company's Second-level NUTS codes*
bidder_nuts_1	*Bidder company's First-level NUTS codes*
bid_iswinning	*Whether it was a winning bid.*
bid_issubcontracted	*Whether part of the contract is planned to be subcontracted.*
bid_subcontractedproportion	*Share of the contract that is expected to be subcontracted.*
bid_isconsortium	*Whether the bid is submitted by a consortium.*
source	*Source of the tender.*

(*continued on next page*)

Table 2 (*continued*)

Column name	Definition
tender_publications_ lastcontractawardurl	*URL of the last contract award announcement.*
tender_publications_ firstdcontractawarddate	*Publication date of the first contract award announcement.*
notice_url	*URL of the last call for tenders (or contract notice) publication related to a given tender.*
tender_publications_ firstcallfortenderdate	*Publication date of the first call for tender announcement.*
tender_year	*Year of the tender.*
tender_addressofimplementation _nuts	*Regional code of the tender implementation. (These are published NUTS codes from the source publication - https://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistic)*
tender_description_length	*Length of the tender description (number of characters).*
lot_description_length	*Length of the lot description (number of characters).*
tender_personalrequirements _length	*Length of the personal requirements set out for participation (number of characters).*
tender_technicalrequirements _length	*Length of the technical requirements set out for participation (number of characters).*
tender_economicrequirements _length	*Length of the economic requirements set out for participation (number of characters).*
currency	*Currency of prices.*
tender_digiwhist_price	*Estimation of the tender level final price, that equals the a) tender_finalprice if available, b) tender_estimatedprice if (a) is missing, c) sum of bid_prices per unique tender if (a) and (b) are missing, d) sum of lot_estimatedprice if (a), (b) and (c) are missing.*
bid_digiwhist_price	*Estimation of the contract price that equals a) the bid_price, or b) the lot_estimatedprice if (a) is missing.*
lot_id	*Unique identifier of a given lot - assigned during data processing.*
bid_id	*Unique identifier of a bid - assigned during data processing.*
bid_priceUsd	*Equals to bid_price but converted to International USD.*
lot_estimatedpriceUsd	*Equals to lot_estimatedprice but converted to International USD.*
tender_estimatedpriceUsd	*Equals to tender_estimatedprice but converted to International USD.*
tender_finalpriceUsd	*Equals to tender_finalprice but converted to International USD.*

A key advantage of such indicators is that they directly stem from micro data on public procurement contracts, but the scale of the datasets allows for macro-level analysis too. The indicators are calculated so that they are not dependent on any particular regulatory regime, hence they allow for tracking the impact of specific regulatory changes too. Moreover, as the risk indicators proxy specific corrupt behaviours, they can help policymakers pinpoint practices that are exploited and hence allow for targeted policy interventions. Nevertheless, we also offer a composite corruption risk index as an average of the individual risk indicators which lead to a more reliable risk assessment. This composite score proxies corrupt behaviours on the contract level, irrespective of the specific corrupt strategy employed. In sum, our indicators offer both a specific measurement for corrupt behaviours and an overall assessment of corruption prevalence. In Fig. 5, we provide a ranking of countries, illustrating their overall composite risk scores and highlighting the contribution of each corruption risk component. The United States shows the lowest average risk score stemming mainly from single bidding risk, procedure type risk and call for tender documents not being published. On the other hand, Portugal has the highest average corruption risk score driven by elevated single bidding and procedure type risks. The figure also shows the type of procurement risk components that can be calculated for each country based on data availability.

While we consider measuring corruption risks in public procurement as one of the main applications of the GPPD, it is by far not its only use. Many scholars have developed methods to measure public spending efficiency using public procurement data [11] or estimated inter-bidder

Table 3
List of filters in country datasets.

Column name	Definition
filter_framework	*The variable filters out framework agreements in a way that if the resulting contract of consecutive minitenders are included in the dataset, those are kept, while only the first stage of the framework agreement awarding process is published, the prospective suppliers are kept with an estimated total value of the framework agreement.*
filter_buyer/filter_bidder	*The variable filters out rows where the buyer and bidder names are either missing or contain erroneous data.*
filter_cancelled	*The variable filters out tender for which a cancellation date or cancellation reason is stored in the data.*
filter_opentender	*This variable deduplicates tenders from overlapping data sources. As a given country can have multiple data sources that publish data on the same tenders, some of them can be present in a country dataset multiple times. This variable is a simple way of tackling these overlaps by keeping only one tender (*i.e. *having the value 'true' for those contracts that are deduplicated). It is based on tender value (*e.g. *above a certain value threshold, only tenders from one source have 'true' values, whereas below it only tender from the (an) other source have 'true' values), supply type (*i.e. *different value thresholds are in use for supplies/services/works). For example, if there is a national source and an EU source (TED), this variable will be 'true' for all supply tenders that have a value more than EUR 135 thousand and published in TED, while it will have 'true' values for supply tenders below the EUR 135 thousand threshold and published on the national portal.*
filter_year	*The variable filters for the years where we think data quality is good and consistent*
filter_losingbids	*The variable filters out rows referring to the losing bids*
filter_ok	*GTI specific filter which applies a combination of the above filters to work with non-duplicated awarded tenders/lots.*

Table 4
List of procurement corruption risk indicators available in the datasets.

Variable name	Definition
corr_singleb	*The indicator is 0 if the lot received more than one bid during the tendering process, 1 otherwise.*
corr_proc	*The indicator is 0 if the tender has an open procedure type (*i.e. *one that is not associated with higher likelihood of single bidding), 1 otherwise.*
submission_period	*Number of days between the first call for tenders publication date and the bidding deadline.*
corr_subm	*The indicator is 0 if the contract's submission period length is not significantly related to higher probability of single bidding, 1 otherwise.*
corr_nocft	*The indicator is 0 if the tender does have a call for tenders publication, 1 otherwise.*
decision_period	*Number of days between the bidding deadline and award decision date.*
corr_decp	*The indicator is 0 if the contract's decision period length is not significantly related to higher probability of single bidding, 1 otherwise.*
corr_tax_haven	*The indicator is 0 if the supplier is not from a high financial risk country, 1 otherwise.*
corr_spending_concentration	*The indicator is the share of the total amount (based on bid_price) won by a specific supplier from a given buyer (*i.e. *higher the values refer to bigger spending concentration).*
cri (Composite Risk score)	*GTI Composite Risk score - Average of the above risk scores*

collusion or bid rigging [12]. Moreover, the large weight of public procurement in government spending has also made it into a key field to study for political science, for example looking at distributive politics, or electoral accountability [13].

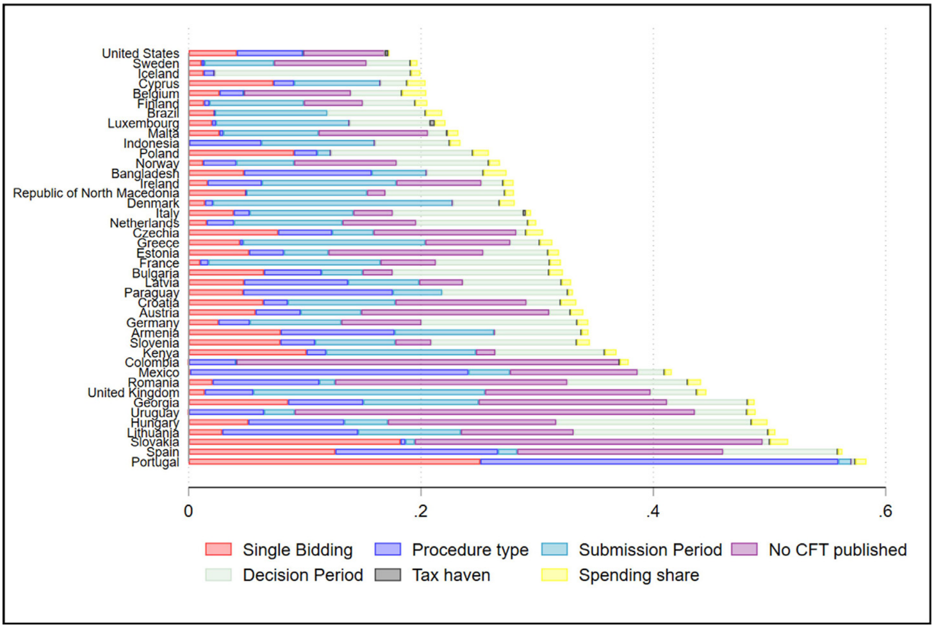

Fig. 5. Composite CRI and its constitutive risk indicators by country.

Limitations

Despite the exceptionally wide scope and detail of the data, it suffers from a number of limitations which users must be aware of. Crucially, both the scope and quality of the datasets vary across countries and over time within the same country limiting analytical uses. First, countries differ with regards to the range of contracts they publish based on regulatory requirements. Usually, contracts are published above a certain value threshold or there are sectoral exceptions such as defence contracts, making some public procurement datasets more or less closely reflecting the actual full population of government contracts. Second, the quality of the data varies across countries depending on the quality of the official government data source. Among others, missing values make some comparisons challenging and limit the analytical uses of some variables. For example, if only call for tenders are published and not the actual final price and/or winner of the contract then the use of those tender records is limited for market analytics. Furthermore, users should assess if missing data in a country is randomly distributed or there is a systemic bias in data publication. The data collection process is usually a decentralized effort where central procurement authorities depend on local authorities to feed the e-procurement system with complete and accurate information. This is an important nuance to determine if the unpublished data is deliberately left out or is just the result of lack of capacity for data collection. These issues may limit the validity of indicators that can be calculated for each country.

Ethics Statement

The data were obtained from the official websites of the EU's Tenders Electronics daily and each country's national public procurement data portal (See Annex) which publish the data with the aim of advancing transparency, market efficiency and government transparency. The data includes information on organizations and formal tenders and contracts, hence do not fall under personal data protection regulations in Europe or elsewhere (i.e. no personal information is processed).

Data Availability

GTI Global Public Procurement Dataset (GPPD) 2/2 (Original data) (Mendeley Data)
GTI Global Public Procurement Dataset (GPPD) 1/2 (Original data) (Mendeley Data)

CRediT Author Statement

Mihály Fazekas: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Writing – original draft; **Bence Tóth:** Conceptualization, Investigation, Methodology, Resources, Software, Supervision, Writing – original draft; **Aly Abdou:** Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft; **Ahmed Al-Shaibani:** Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft.

Acknowledgements

The authors gratefully acknowledge the generous and long-standing support from a range of funders, in particular the Open Society University Network, the Open Society Foundations, the EEA and Norway Grants Fund for Regional Cooperation, the EU's Horizon-2020 funding, and the UK Foreign, Commonwealth and Development Office's Anti-corruption Evidence Programme. In addition, the Government Transparency Institute has devoted its own funds to the creation, maintenance and improvement of the datasets over the last 6 years. The authors are also greatly indebted to the numerous research assistants and programmers working on these datasets in the last 10 years, in particular Government Transparency Institute staff and programmers at Datlab and Precognox companies. This publication is the result of research conducted for Central European University, Private University. It was made possible by the CEU Open Access Fund.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Supplementary Materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2024.110412.

References

[1] OECDGovernment at a Glance 2021, OECD, Paris, 2021, doi:10.1787/1c258f55-en.
[2] United Nations Office on Drugs and Crime (UNODC), Guidebook on Anti-Corruption in Public Procurement and the Management of Public Finances, UNDC, Vienna, 2013.
[3] Á. Czibik, B. Tóth, M. Fazekas, How to Construct a Public Procurement Database from Administrative Records, Government Transparency Institute, Budapest, 2015 GTI-R/2015:02.
[4] L.J. de Sousa, J.P. Martins, L. Sanhudo, Portuguese public procurement data for construction (2015–2022), Data Brief. 48 (2023).
[5] L. Potin, V. Labatut, P.H. Morand, C. Largeron, FOPPA: an open database of French public procurement award notices from 2010 to 2020, Sci. Data 10 (2023) 303, doi:10.1038/s41597-023-02213-z.

[6] T. Stout, A. Teston, B. Langhals, J. Delorit, C. Hendrix, S Schuldt, United states department of defense (DoD) real property repair, alterations, maintenance, and construction project contract data: 2009–2020, Data Brief. 32 (2020).

[7] M. Fazekas, A. Abdou, Y. Kazmina, N. Regős, Development aid contracts database: world Bank, Inter-American Development Bank, and EuropeAid, Data Brief. 42 (2022) 108121.

[8] M. Fazekas, G. Kocsis, Uncovering high-level corruption: cross-national objective corruption risk indicators using public procurement data, Br. J. Polit. Sci. 50 (1) (2020) 155–164.

[9] M. Fazekas, I.J. Tóth, L.P. King, An objective corruption risk index using public procurement data, Eur. J. Crim. Pol. Res. 22 (3) (2016) 369–397.

[10] A. Abdou, O. Basdevant, E. David-Barrett, M. Fazekas, Assessing Vulnerabilities to Corruption in Public Procurement and Their Price Impact. IMF Working Papers: WP/22/94, IMF, Washington, D.C., 2022.

[11] Fazekas, M., Borges de Oliveira, A., & Regos, N. (2021). Lowering Prices of Pharmaceuticals, Medical Supplies, and Equipment. Insights from Big Data for Better Procurement Strategies in Latin America. Policy Research Working Paper: WPS 9689, Washington, D.C.: The World Bank.

[12] Adam, I., Fazekas, M., Kazmina, Y., Teremy, Z., Tóth, B., Villamil, I.R., & Wachs, J. (2022). Public Procurement Cartels: a Systematic Testing of Old and New Screens. GTI-WP/2022:01, Budapest: Government Transparency Institute.

[13] R. Broms, C. Dahlström, M. Fazekas, Political competition and public procurement outcomes, Comp. Polit. Stud. 52 (9) (2019) 1259–1292.