Analysis of state register data to identify anomalies and corruption threats in tenders and procurements

Kyrylo Chornyi[1,†], Galina Shilo[1,†] and Anastasiia Lebedieva-Dychko[1,†]

[1] *Zaporizhzhia National University, м, Zhukovs'koho St, 66, Zaporizhzhia, Zaporizhia Oblast, Zaporizhzhia, 69600, Ukraine*

Abstract

Public procurement plays a crucial role in economic development, accounting for a significant portion of government expenditures worldwide. However, the sector remains highly vulnerable to corruption, inefficiencies, and financial mismanagement, leading to severe economic and social consequences. Corruption in public procurement not only undermines fair competition but also results in inflated costs, poor-quality services, and weakened public trust in governmental institutions. The digitization of legal algorithms presents a promising solution to these challenges. By leveraging artificial intelligence (AI), big data, blockchain, and automated decision-making systems, governments can strengthen compliance, detect fraudulent activities, and enhance oversight in public procurement. The system for detecting fraudulent activities in public procurements and tenders is proposed. The system gets data from the open database of companies, their activities, tenders, court cases, as well as tax and legal information. Using an algorithm, the system can detect any activities with corruption threats. After analyzing hundreds of tenders, a dataset was formed. This dataset of tenders, companies and their activity will be used for training a machine-learning-based algorithm to further enhance the analytical capabilities of the system.

Keywords

Procurement, corruption, government, tenders, fraud, cluster analysis, dataset

1. Introduction

Corruption in public procurement is one of the most serious problems of the modern public administration system. It undermines public trust, limits economic development and leads to inefficient spending of budget funds. At the same time, existing control tools do not allow for timely and comprehensive monitoring of possible corruption schemes and connections between procurement participants due to the large volume of data and the complexity of analyzing the relationships. As a result, violations are often identified after the completion of procedures, which significantly reduces the possibility of restoring justice and efficient use of resources.

One of the industries that is particularly susceptible to corruption risks is the construction sector. This is due to the large volumes of funding and the complexity of projects. According to studies by the European Commission and the Organization for Economic Cooperation and Development (OECD), up to a fifth of the European Union's GDP is spent through public procurement, and losses from corrupt practices reach 20-25% of the allocated funds. The most common forms of corruption in procurement include bribery, collusion between participants, conflicts of interest and hidden lobbying. Recent Eurobarometer surveys confirm the prevalence of the problem: a significant share of entrepreneurs consider corruption a serious barrier to participation in tenders [1].

The Second International Conference of Young Scientists on Artificial Intelligence for Sustainable Development (YAISD), May 8-9, 2025, Ternopil, Ukraine

[*] Corresponding author.

[†] These authors contributed equally.

nonchasd@gmail.com (K. Chorniy); shilo.gn@gmail.com (G.Shilo); ldas.1405@gmail.com (A.Lebedieva-Dychko)

ORCID -0009-0009-3724-4707(K. Chorniy); ORCID 0000-0002-5020-6707 (G.Shilo); ORCID 0009-0003-1931-1861 (A.Lebedieva-Dychko)

To address these issues, European countries are actively implementing digital technologies and legislative measures aimed at increasing transparency. Examples of successful initiatives are laws on access to information - the Freedom of Information Act in the UK. The law allows the public to access government data, significantly reducing the opportunities for corruption. In addition, the EU has initiated a large-scale transition from paper to electronic tendering procedures over the past decade. E-procurement platforms such as Prozorro in Ukraine and TED (Tenders Electronic Daily) in the EU provide a full "electronic trail", increasing transparency and accessibility of information for all stakeholders.

The platform that operates in Ukraine provides a unified information space where tender announcements, commercial offers, and contracts are published. At the same time, in Ukraine, there have been significant problems with access to information since the beginning of the full-scale invasion. Limited access to legal registers prevents full monitoring and timely detection of corruption risks. This circumstance significantly reduces the effectiveness of the state's anti-corruption policy.

The digitization of legal algorithms as a tool for mitigating corruption in public procurement is explored [1]. Corruption in procurement arises from inefficiencies, conflicts of interest, and weak oversight, leading to financial mismanagement and security risks. The study highlights how digital technologies, including AI-driven analytics, blockchain, and big data, can enhance transparency, automate oversight, and detect fraudulent activities in procurement processes. Implementing AI-based monitoring systems and aligning national procurement policies with international best practices can significantly reduce misappropriation risks.

The use of quantitative indicators to detect corruption risks in public procurement is researched [2]. The study applies machine learning techniques to analyze roadwork contracts in Italy, identifying new red flags derived from police investigations and judicial practices. It finds that multi-parameter awarding criteria are systematically linked to high corruption risk. However, urgency-based procedures are more obvious red flags, but they are ineffective due to their predictability. The research also demonstrates that corruption risk prediction improves when including unmonitored indicators, highlighting the adaptability of corrupt actors to scrutiny. Furthermore, the study emphasizes that private firm competition and transparent bidding processes help mitigate corruption risks. It concludes that enhancing data collection on public contracts and strengthening coordination between courts and regulatory authorities could improve corruption detection. Finally, the findings suggest that concealing specific monitoring criteria may prevent corrupt actors from adapting their strategies, ultimately aiding enforcement efforts.

The object of the research is the processes of public procurement and tendering within Ukrainian electronic registry systems. The subject of the research is the methods and digital tools for analyzing data from state registries to identify anomalous patterns and detect corruption risks in tenders and procurement activities. To reduce the corruption risks in Ukraine, the operational control over the transparency of procurement is a relevant problem [3]. Current solutions in Ukraine for analyzing public procurement and tenders offer fragmented information. Prozorro contains data on tenders, while YouControl shares s business analytics about companies. To solve the problem, an integrated approach is proposed with the possibility of automated risk assessment based on multidimensional parameters and the use of artificial intelligence. Artificial intelligence technologies will allow identifying additional unknown corruption schemes [4-5]. The paper aims to develop a specialized analytical system as a web application to aggregate and process data from open sources (Zakupivli.pro [6], YouControl [7], etc.), providing a comprehensive analysis of tender purchases by means artificial intelligence technologies. The developed software solution allows for a detailed analysis of winners, identifying links between companies and officials, visualizing data at various stages of procurement, and generating analytical reports available to citizens and experts. The implementation of such an automated system is aimed at increasing transparency, promptly identifying and preventing corruption risks, and strengthening public confidence in public institutions and procurement procedures.

2. The web application for tender analysis

One of the common methods of abuse is to create the appearance of competition when related companies participate in a tender. This can happen through several mechanisms:

- affiliation by managers and beneficiaries (several legal entities belong to the same group of owners);
- using registration of fictitious firms to submit alternative bids and simulate competition;
- price manipulation (preliminary agreements between participants on the distribution of winnings by overstating or understating offers).

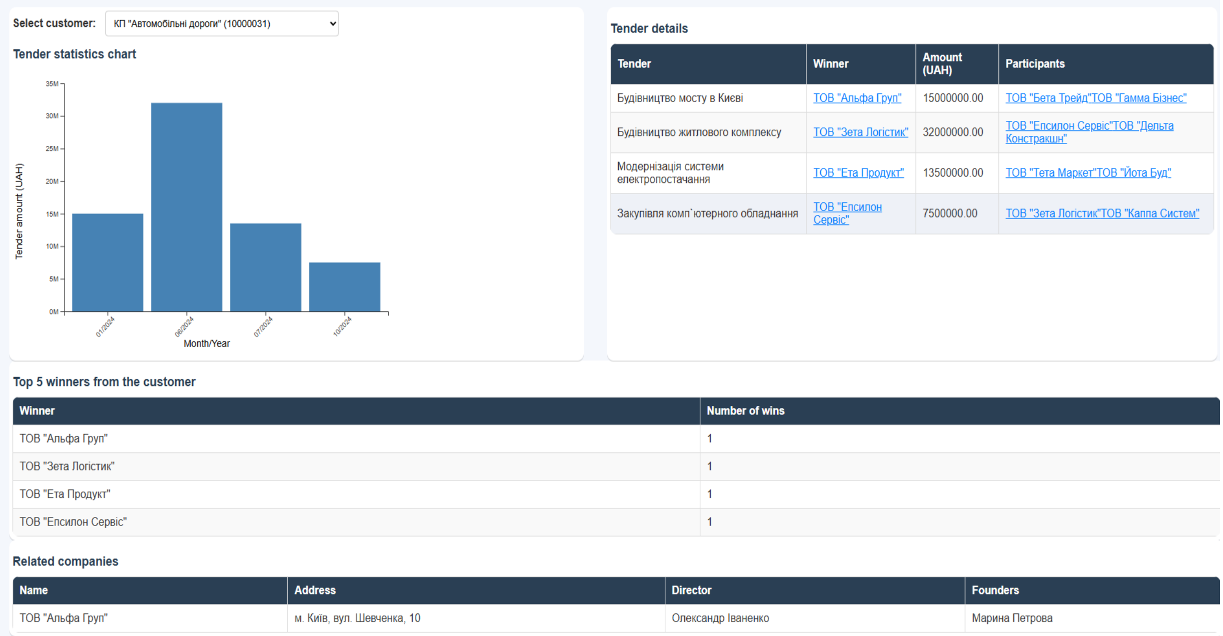
The proposed web application automatically collects data via API on the following information: about tenders, amounts of concluded contracts and data on contractors, subcontractors, including their identification code (EDRPOU), legal addresses, information on directors and ultimate beneficiaries. After collection, the data is consolidated in a single database.

The analysis is carried out according to the following key parameters:

- intersection of beneficiaries, directors and legal addresses of companies;
- frequency of victories of one company in tenders from one customer;
- relationship of participants through third structures (affiliation of companies);
- unusual changes in the cost of tender offers.

It helps to identify companies that systematically win tenders from the same subcontractors, and find the connections between the winners of purchases through founders, directors and registration addresses.

The program implements the analysis of tender processes, including the selection of the contractors, viewing detailed information on purchases and analyzing the winners. The system displays data on tenders, including the name of the purchase, the winner, the contract amount and the list of participants. Information on the winner is interactive, which allows for immediate in-depth analysis (Figure 1).



Select customer: КП "Автомобільні дороги" (10000031)

Tender statistics chart

Tender details

Tender	Winner	Amount (UAH)	Participants
Будівництво мосту в Києві	ТОВ "Альфа Груп"	15000000.00	ТОВ "Бета Трейд"ТОВ "Гамма Бізнес"
Будівництво житлового комплексу	ТОВ "Зета Логістик"	32000000.00	ТОВ "Епсилон Сервіс"ТОВ "Дельта Констракшн"
Модернізація системи електропостачання	ТОВ "Ета Продукт"	13500000.00	ТОВ "Тета Маркет"ТОВ "Йота Буд"
Закупівля комп`ютерного обладнання	ТОВ "Епсилон Сервіс"	7500000.00	ТОВ "Зета Логістик"ТОВ "Каппа Систем"

Top 5 winners from the customer

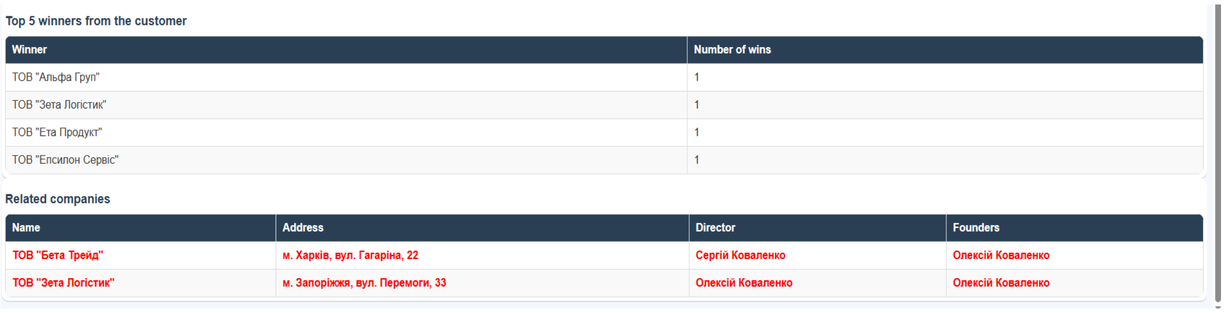
Winner	Number of wins
ТОВ "Альфа Груп"	1
ТОВ "Зета Логістик"	1
ТОВ "Ета Продукт"	1
ТОВ "Епсилон Сервіс"	1

Related companies

Name	Address	Director	Founders
ТОВ "Альфа Груп"	м. Київ, вул. Шевченка, 10	Олександр Іваненко	Марина Петрова

Figure 1. General interface

The web application generates a list of companies that most often win tenders from a specific contractor, which helps identify possible violations. Additionally, a check is made of the connections between companies by the coincidence of directors, founders or legal address (Figure 2). This allows for finding hidden connections between procurement participants.



Top 5 winners from the customer

Winner	Number of wins
ТОВ "Альфа Груп"	1
ТОВ "Зета Логістик"	1
ТОВ "Ета Продукт"	1
ТОВ "Епсилон Сервіс"	1

Related companies

Name	Address	Director	Founders
ТОВ "Бета Трейд"	м. Харків, вул. Гагаріна, 22	Сергій Коваленко	Олексій Коваленко
ТОВ "Зета Логістик"	м. Запоріжжя, вул. Перемоги, 33	Олексій Коваленко	Олексій Коваленко

Figure 2. Displaying related companies

This tool is effective for monitoring public procurement, helping journalists and anti-corruption agencies quickly identify possible violations. By using open data and modern analysis algorithms, the web application promotes transparency and fairness of procurement processes.

Based on the analysis of several tenders by the system, the characteristic patterns indicating possible corruption risks are proposed and represented in Tables 1 and 2.

Table 1
Contractors table schema

Category	Attributes	Patterns of corruption risks
Tender activity	- Number of tenders held during the period; - Average tender amount; - number of tenders without competition (single participant); - Share of cancelled tenders.	Contractors with abnormal activity (inflated amounts, frequent cancellations)
Financial flows	- Total amount of tenders; - Average contract price; - The share of the budget going to one subcontractor - Average difference between the forecast and actual contract price	Analysing the distribution of financial flows allows not only to determine the share of the budget attributable to each contractor, but also to identify anomalies - systematic and disproportionate allocation of funds to one bidder. Such behaviour may indicate possible affiliation with the contracting authority, restriction of competition or the existence of corrupt agreements.
Relationships with subcontractors	- Number of unique winners - Percentage of wins for one subcontractor in total and for the period - Number of subcontractors connected (coincidence of owners, addresses) - Number of tenders won by new companies (young firms opened no more than 1 year ago) - Analysis of whether the second in the tender was a young company	Contractors would be more interested in some specific subcontractors based on personal gains, but on professional requirements
Geographical factor	- Subcontractor's region - Share of local subcontractors - Average contract amount by region	Regional schemes and differences in procurement

Category	Attributes	Patterns of corruption risks
Contract terms	- Distribution of tenders by month (are there any spikes in November/December?) - The share of tenders conducted urgently (less than 10 days between the announcement and the submission of applications)	During wartime, a minimum of 7 days is allowed for conducting a tender, usually, it is 15, and if it is often 7, then it is suspicious; it may be a tender for one person.

Table 2
Subcontractors table schema

Category	Attributes	Comments
The number of tender wins	- Number of won tenders - The number of submitted commercial offers - Share of wins from the total number of participations	Companies that win too often
Contract amounts	- Total amount of won tenders - Average amount of one contract - Maximum contract amount	Identify abnormal wins. For example, large amounts (over 10 million) for young companies opened less than 1 year ago.
Relationship with other subcontractors	- Coincidence of owners - Coincidence of legal addresses - General Directors - Participation in tenders with the same contractors	Cartel schemes, collusion between companies
Time characteristics	- Dates of wins - Bursts of activity (e.g. November/December) - Frequency of participation	Analysis of the temporal characteristics of participation in tenders allows us to identify typical and anomalous patterns of contractor behaviour - for example, a concentration of wins in certain periods (usually at the end of the budget year), bursts of activity in November-December, and irregular or selective frequency of participation. Such temporal anomalies may indicate attempts to target budget allocation in favour of a particular bidder, participation in pre-agreed procedures, or fictitious competition.
Participants	- Number of participants - Frequency of participation of the same companies in one tender - Number of refusals to participate	The subcontractors that participated in the previous tenders with the same contractor.

3. Cluster analysis to anomaly detection in tenders

To further enhance the system for tender analysis, we can utilize a machine-learning-based approach. The cluster analysis is used to identify groups of companies exhibiting anomalous behavior in tenders. This approach enables the segmentation of companies into clusters based on a multidimensional feature space, such as encompassing tender activity, behavioral patterns, affiliations, geography, interactions with contracting authorities, and financial indicators. Additionally, it aids in identifying companies that significantly deviate from typical behavioral profiles. By utilizing unsupervised machine learning techniques such as k-Means clustering, DBSCAN, and Isolation Forest, it is possible to detect individual companies or small groups exhibiting unusual patterns. To do so, first, we need a proper dataset to train our model with.

There are many types of datasets based on the type of data they contain. In our case, we have a tabular dataset, which is a database dump into CSV.

The dataset consists of 3 CSV files, each represents a specific table in the database schema. Contractors.csv file represents information about the companies that put tenders, as shown in Table 3.

Table 3
Contractor table schema

Column name	Description
id	Unique identification number
name	Company name
short_name	Short company name
address	Company address
director	Company CEO
status	Company status, either alive, paused, bankruptcy
economic_activity	Company primary economic direction
founders	Company founders
actual_date	Date on which information about the company is relevant
registration_date	Date when the company was registered
created_at	When the company was first added to the database
edrpou	Unique Ukrainian identification number

The file tenders.csv represents information about tenders that are submitted by the companies, as shown in Table 4.

The file subcontractors.csv represents information about the customers who win/participate in those tenders, as shown in Table 5.

The current dataset schema is not complete since right now some of the government data sources are closed due to the war. Because of this, we can only use those that are in the public domain. The data in the public domain is quite good, however, in some cases, the information is incomplete or not enough for analysis.

In particular, the following categories of information are missing:

- history of changes in founders, , and ultimate beneficiaries, which is critical for identifying hidden company affiliations;
- data on the company's turnover, debts, and VAT payments, which is necessary to separate real businesses from fictitious firms;
- data from declarations of civil servants is important for establishing connections between officials and companies that win tenders;
- data on the execution of contracts - including the involvement of subcontractors, certificates of completion of work, and payment, which is often not published;
- historical tender data, including for defense procurement and critical infrastructure facilities, which are currently closed.

Table 4
Tenders table schema

Column name	Description
id	Unique identification number
id_tenders	Unique tender code
name_tenders	Tender name

contractor_edrpou	Unique Ukrainian identification number of the tender customer
contractor	Customer name
winner_edrpou	Unique Ukrainian identification number of the tender winner
winner	Subcontractor’s company name
amount	Amount of money of the tender
date	Date when the tender was published
subcontractors	Subcontractors
created_at	When the tender was created
url_tenders	Tender url

Table 5
Subcontractor table schema

Column name	Description
id	Unique identification number
edrpou	Unique Ukrainian identification number
name	Customer’s company name
created_at	When the company was added to the database

The absence of these data limits the capabilities of machine analysis and does not allow for the full identification of affiliation schemes, price gouging, fictitious competition and other signs of corruption. Gaining access to such registers, at least in a limited form for scientific and analytical purposes, is critically important for building an effective public procurement monitoring system in Ukraine.

Another difficult in tender data analysis is the different terms in text type attributes may represent the same concept. Appling Natural Language Processing technologies during the preliminary data preparation phase to identify synonyms in text data is an effective approach that enhances the quality of subsequent analyses.

4. Conclusion

Thus, the use of modern technologies in public procurement is an important step in the fight against corruption. Digital platforms improve the transparency of procedures and provide information for analysis and identification of possible corruption schemes.

Using web applications for monitoring tenders helps to effectively identify fictitious competition schemes and contract manipulation. The system for identifying relationships between companies helps to minimize the influence of affiliated structures on procurement results.

However, significant challenges remain, such as insufficient access to data, the complexity of integrating different sources of information and the limited capabilities of law enforcement agencies. To achieve maximum effect, it is necessary to further develop the legislative framework, ensure open access to state registers and improve data analysis methods.

To improve the accuracy of analysis and further automated detection of corruption schemes, we propose the formation of a dataset that will include:

- historical data on tenders and their winners;
- information on company connections, directors and beneficiaries;
- financial indicators and amounts of concluded contracts;
- tags of previously identified corruption cases and investigations;
- data on unnatural price jumps and discrepancies with market conditions.

This dataset will be used to train machine learning models that can automatically find suspicious anomalies and offer analytical reports for experts and government agencies.

This direction of the system's development is aimed at use by both government anti-corruption agencies and independent researchers and journalists.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] European Union. *Eurobarometer site.* URL: https://europa.eu/eurobarometer/surveys/detail/3180 (accessed: 14.02.2025).

[2] F. Decarolis and C. Giorgiantonio. Corruption red flags in public procurement: new evidence from Italian calls for tenders. *EPJ Data Science*, 11(1), 2022. doi: 10.1140/epjds/s13688-022-00325-x (accessed: 27.03.2025).

[3] O. Makarenkov. Digitisation of legal algorithms to prevent public procurement corruption. *Baltic Journal of Economic Studies*, 10(5):254–265, 2024. doi: 10.30525/2256-0742/2024-10-5-254-265 (accessed: 27.03.2025).

[4] Applied machine learning to anomaly detection in enterprise purchase processes: a hybrid approach using clustering and isolation forest / A. Herreros-Martínez et al. Information. 2025. Vol. 16, no. 3. P. 177. URL: https://doi.org/10.3390/info16030177 (date of access: 25.04.2025).

[5] Busu M., Busu C. Detecting bid-rigging in public procurement. A cluster analysis approach. Administrative sciences. 2021. Vol. 11, no. 1. P. 13. URL: https://doi.org/10.3390/admsci11010013 (date of access: 25.04.2025).

[6] Prozorro. *Zakupivli Platform.* URL: https://zakupivli.pro/ (accessed: 06.02.2025).

[7] YouControl. *Business Analytics Platform.* URL: http://youcontrol.com.ua (accessed: 18.01.2025).