

УДК 004.85:336.5

**ВИЯВЛЕННЯ АНОМАЛЬНИХ ЦІНОВИХ ПРОПОЗИЦІЙ
У СИСТЕМІ ПУБЛІЧНИХ ЗАКУПІВЕЛЬ
З ВИКОРИСТАННЯМ АЛГОРИТМІВ КЛАСТЕРИЗАЦІЇ**

Бабарицький Д. О.

e-mail: dmytro.babarytskyi@nure.ua

Науковий керівник – к.т.н., доц. Білова Т. Г.

Харківський національний університет радіоелектроніки, каф. КМІТ
м. Харків, Україна

This study proposes a machine learning model for detecting price anomalies in public procurement. Focused on data measured in kilograms, the approach uses NLP (TF-IDF) and K-Means for semantic categorization of lots. Outlier detection is implemented via Principal Component Analysis (PCA) and the DBSCAN algorithm. Results demonstrate that this multidimensional model identifies price deviations more effectively than standard statistical methods. It facilitates automated monitoring of budget expenditures and supports corruption risk mitigation. The integration of semantic and numerical analysis ensures high accuracy in detecting suspicious procurement patterns and enhancing financial transparency.

Ефективність цифрових публічних закупівель безпосередньо залежить від прозорості ціноутворення та своєчасного виявлення ризикових тендерів. В умовах системи «Prozorro» суттєвою проблемою залишається значний обсяг неструктурованих даних і довільність формулювання назв товарів, що ускладнює порівняння цін. Дослідження застосування штучного інтелекту у сфері публічних закупівель підтверджують доцільність автоматизованого аналізу таких даних [1]. Традиційні статистичні методи, зокрема правило трьох сигм або IQR, не враховують багатовимірні залежності між ціною, обсягом, сезонністю та характеристиками товару, що обмежує їхню ефективність [2]. У роботі запропоновано модель машинного навчання для кластеризації лотів і виявлення аномальних цін на основі алгоритмів навчання без учителя.

Об'єктом дослідження є закупівлі з одиницею виміру «кілограм». Модель враховує ціну за одиницю, логарифмований обсяг і очікувану вартість, наявність ПДВ та місяць оголошення тендеру; для зменшення впливу екстремальних значень застосовано логарифмування та стандартизацію. Семантичну уніфікацію назв забезпечено через лематизацію та TF-IDF-векторизацію [1]. Кластеризацію виконано алгоритмом K-Means, зниження розмірності – методом головних компонент, а виявлення аномалій – алгоритмом DBSCAN, що належить до щільнісних методів кластеризації та ефективно працює з шумом у багатовимірному просторі [3, 4].

Як показано в таблиці 1, перші дві головні компоненти пояснюють 79% загальної дисперсії, що свідчить про збереження основної інформації після зниження розмірності. Аналіз матриці ваг засвідчив, що PC2 переважно визначається ціною за одиницю, тоді як PC1 відображає масштаб закупівлі через обсяг і загальну вартість; сезонний фактор має помірний вплив, а ознака ПДВ – мінімальний.

Таблиця 1 – Вплив ознак на головні компоненти

Ознака	PC1	PC2
Ціна за одиницю	-0.027	0.960
Обсяг кг log	0.651	-0.128
Вартість log	0.661	0.028
is_VAT	0.000	0.000
Month	-0.371	-0.246

Результати багатовимірного аналізу закупівель представлено на рисунку 1.

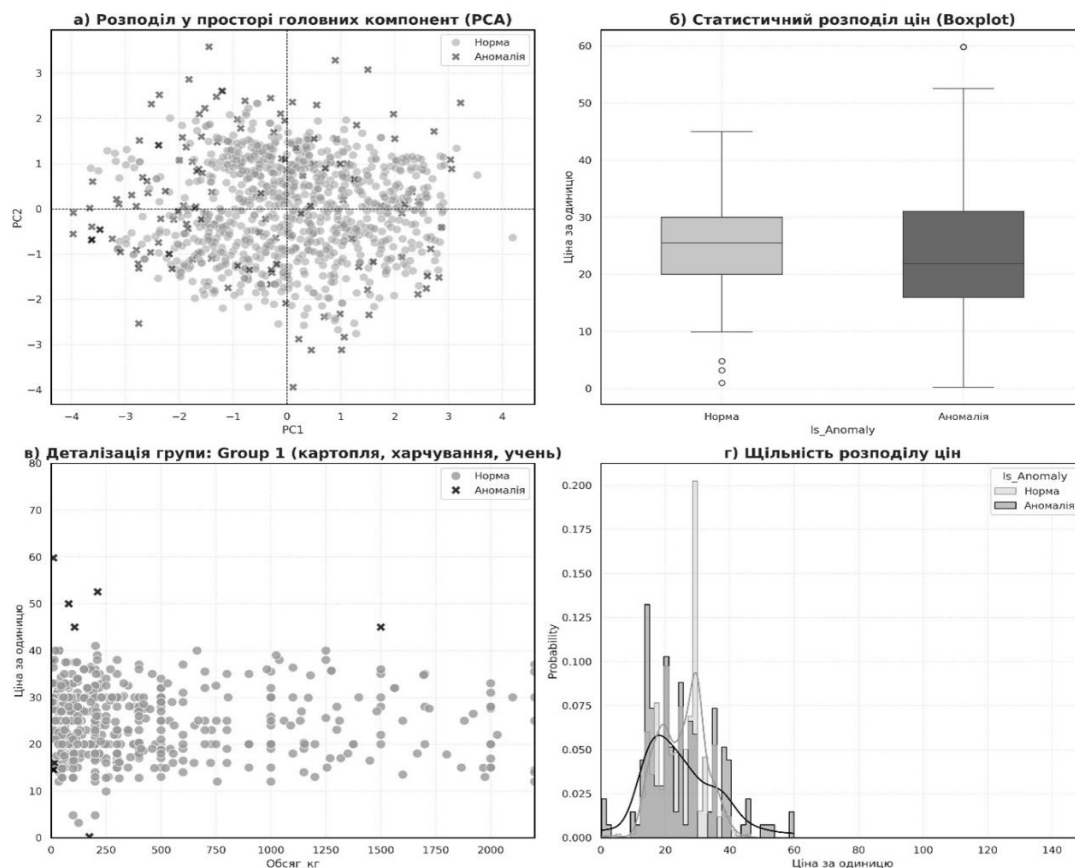


Рисунок 1 – Результати багатовимірного аналізу закупівель:
а – проєкція PCA; б – розподіл цін типових та аномальних лотів;
в – аналіз групи «Group 1»; г – щільність розподілу цін

На рисунку 1, а зображено проєкцію PCA, де 12,95% лотів визначено як аномальні. Більшість із них тяжіє до країв простору, однак частина розташована всередині основної хмари даних, оскільки DBSCAN застосовувався окремо в межах підкатегорій і виявляє локальні відхилення, нетипові для конкретної групи, хоча й не екстремальні для ринку загалом.

Статистичний розподіл цін (рис. 1, б) показує, що попри близькість медіан розкид цін у групі аномалій суттєво більший: їхня дисперсія перевищує типову у 1,6 рази, що підтверджує ширший міжквартильний розмах.

Деталізація найбільшої групи «Group 1» (рис. 1, в) демонструє практичну цінність моделі: діапазон 10–40 грн/кг визначено як ринкову норму, тоді як завищені та демпінгові або помилкові ціни, близькі до 0 грн/кг, маркуються як аномалії.

Аналіз щільності розподілу (рис. 1, г) підтверджує концентрацію типових закупівель навколо модального значення, тоді як аномалії характеризуються більшою дисперсією та правосторонньою асиметрією.

Таким чином, розроблено модель автоматизованого виявлення аномальних цін у публічних закупівлях, що поєднує семантичну уніфікацію описів і багатовимірну кластеризацію. Емпірично встановлено, що аномальні лоти характеризуються підвищеною волатильністю цін, а результати підтверджують доцільність багатовимірного аналізу для моніторингу закупівель. Підхід дає змогу виявляти не лише цінові піки, а й нетипові співвідношення «обсяг–вартість», забезпечуючи пріоритезацію лотів для аудиту. Подальші дослідження доцільно спрямувати на інтеграцію моделі в аналітичні модулі системи «Prozorro» та розширення набору ознак за рахунок контрактних і посттендерних показників для верифікації ризиків.

Список використаних джерел:

1. Fazekas M., Tóth I. J., King L. P. An objective corruption risk index using public procurement data // *European Journal on Criminal Policy and Research*. 2016. Vol. 22, No. 3. P. 369–397. DOI: <https://doi.org/10.1007/s10610-016-9308-z>
2. Білова Т. Г., Ярута В. О., Побіженко І. О. Формування ієрархічної структури показників моніторингу складної інформаційної системи // *Системи обробки інформації*. 2015. № 12(137). С. 78–80.
3. Chandola V., Banerjee A., Kumar V. Anomaly Detection: A Survey // *ACM Computing Surveys*. 2009. Vol. 41, No. 3. Article 15. DOI: <https://doi.org/10.1145/1541880.1541882>
4. Ester M., Kriegel H.-P., Sander J., Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise // *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. Portland, 1996. P. 226–231.