УДК 324:004.8:351.9
DOI https://doi.org/10.20535/2308-5053.2023.1(57).280822

НАПРЯМКИ ВПРОВАДЖЕННЯ ІННОВАЦІЙНИХ АНАЛІТИКО-СТАТИСТИЧНИХ ТЕХНОЛОГІЙ ЯК ІНСТРУМЕНТУ ПРОТИДІЇ КОРУПЦІЇ В ДЕРЖАВІ

Яцина Ю. О.,
голова
ГО «Союз соціальних технологів України»
ORCID ID: 0000-0002-7286-4655
yatsyna.yuliia@gmail.com

Стаття присвячена проблемі впровадження інноваційних аналітико-статистичних технологій як інструменту протидії корупції в державі. Інноваційні аналітико-статистичні технології визначено у широкому значенні як сукупність методів та інструментів, що базуються на використанні математичних та статистичних методів аналізу даних з метою виявлення корисних залежностей та закономірностей в даних, підвищення ефективності прийняття рішень та виявлення аномалій у різних сферах діяльності та, у вузькому, як процес використання найсучасніших методів аналізу даних з метою виявлення складних залежностей та корисних закономірностей в даних. За результатами аналізу змісту тематичних публікацій було визначено 4 напрямки: 1) напрям розробки методологічного інструментарію; 2) напрям аналізу вторинної соціологічної інформації; 3) напрям створення автоматизованих систем аналізу природньої мови та візуалізації просторових даних; 4) напрям впровадження технологій машинного навчання і штучного інтелекту для ідентифікації суб'єктів корупційних відносин та/або отримання статистично обґрунтованих підтверджень наявності / відсутності корупції. Визначено, що найбільшу значимість на сучасному етапі розвитку суспільства мають практико-орієнтовані дослідження, адже саме на цьому рівні відбувається безпосередня апробація відповідних теоретичних моделей і методологічного інструментарію, а також створюються прецеденти для використання результатів аналітико-статистичних досліджень як складової доказової бази тих або інших фактів шахрайства або корупції. Визначено, що безпосереднім прикладом застосування інноваційних аналітико-статистичних технологій як інструменту протидії корупції з політологічної точки зору є електоральна криміналістика (electoral forensic). В рамках електоральної криміналістики виділяються дві групи методів. Перша група веде своє походження від теорії чисел і звертається до частотних характеристик числових даних електоральної статистики. Друга група методів спирається на пошук аномалій у відносинах між різними параметрами електорального процесу, наприклад, рівнем явки та рівнем підтримки кандидатів. Основним критерієм, що використовується для виявлення електоральної корупції, є розбіжність реальних (документально зафіксованих) результатів виборів від нормативних (модельних).

Ключові слова: протидія корупції, інноваційні технології, машинне навчання, штучний інтелект, електоральна криміналістика.

Постановка проблеми. Корупції є бар'єром на шляху прогресивного соціально-економічного розвитку будь-якої держави. Вона спотворює моральні норми, зменшує рівень довіри громадян до інститутів влади та є досить важким і невловимим від ока дослідників явищем. Розмитість проявів корупції обумовлює розбіжність у думках дослідників щодо сутності цього феномену, а також нескінченність дискусій щодо визначення його сутності, причин, наслідків та шляхів запобігання. Одним із рішень в цьому питанні вважається діджиталізація держави. Цифрова трансформація державного управління має серйозне антикорупційне значення та вимір. Розвиток на основі інформаційно-комунікаційних технологій принципово нових механізмів державного управління дозволяє розробити раніше невідомі та зовсім несподівані засоби протидії традиційним негативним явищам в системі державного управління. При цьому розвиток інформаційно-комунікаційних технологій може породжувати і нові корупційні та бюрократичні схеми, що у своєму потенціалі може звестися до електронної бюрократизації, електронної корупції чи навіть цифрового концтабору.

Саме тому формування антикорупційної політики із застосуванням інноваційних технологій потребує серйозного наукового аналізу, формулювання таких напрямків розвитку держави на шляху до цифрового формату роботи, які б виключали можливість розвитку вищезгаданих негативних тенденцій та наслідків цифровізації, і стали б основою принципово нових механізмів регулювання державного управління, публічної влади, правоохоронної діяльності. Одним із таких напрямів дослідження можна вважати аналіз реалізованих науково-практичних проектів, результати яких безпосередньо пов'язані або засновані на активному залученні новітніх інформаційних, в тому числі і аналітико-статистичних, технологій до виявлення корупційних злочинів та відповідно протидії їм на рівні держави.

Мета дослідження – визначити напрями впровадження інноваційних аналітико-статистичних технологій у протидії корупції в сучасній державі.

Виклад основного матеріалу. Почнемо наш аналіз із визначення основних понять.

Під корупцію ми розуміємо протиправне використання посадовою особою наданих їй управлінських ресурсів для особистої чи групової вигоди, що може мати як матеріальну, так і нематеріальну форму. При цьому протиправне використання – це порушення як формальних нормативно-юридичних установ, включаючи норми службової поведінки та етики, так і неформалізованих норм поведінки, етики і моралі. Під інструментами протидії корупції розуміються будь-які засоби щодо створенню перешкод вчиненню корупційних діянь, здійснення опору їх поширенню, а також реагування відносно тих діянь, які вже проявилися у фактично скоєних правопорушеннях.

Інноваційні аналітико-статистичні технології визначаються нами:

– в широкому значенні як сукупність методів та інструментів, що базуються на використанні математичних та статистичних методів аналізу даних з метою виявлення корисних залежностей та закономірностей в даних, підвищення ефективності прийняття рішень та виявлення аномалій у різних сферах діяльності;

– у вузькому – як процес використання найсучасніших методів та технологій аналізу даних, таких як машинне навчання, глибинне навчання, нейронні мережі, обробка природної мови, аналіз графів тощо з метою виявлення складних залежностей та корисних закономірностей в даних. До таких технологій відносяться також методи аналізу даних у режимі реального часу, які дозволяють отримувати швидкі та точні результати аналізу великих обсягів даних.

З метою отримання уявлень про сучасні тенденції використання інноваційних аналітико-статистичних технологій для виявлення корупції в різних країнах було обрано публікації глибиною у 10 років, оскільки дослідження не ставить за мету висвітлення всіх публікацій за проблемною тематикою, а керується принципом збільшення кількості аналізованих джерел до досягнення порогового значення, коли кількість нової інформації про методи та підходи, отримані від кожного наступного джерела, не зменшиться настільки, що подальша анотація стане непрактичною. Тому, проаналізувавши щонайменше 45 тематичних публікацій, було визначено, що достатньою кількістю для нашого аналізу є 15–20 робіт у зв'язку із повторенням використовуваних методів, нечітко визначеною методологією або у зв'язку з надмірно декларативними положеннями, що не підтверджувалися емпіричними даними.

Всю сукупність наявних тематичних публікацій можна поділити за такими напрямками:

– технології аналізу, візуалізації статистичних даних і даних соціологічних та відповідних ним досліджень в сфері корупції (Heritage Foundation, Transparency International, ООН, Міжнародний банк тощо), що можна назвати аналізом вторинних даних;

– технології аналізу та візуалізації документальних даних (новини, звіти, офіційні звернення), що містять інформацію про корупційні (шахрайські) дії та правопорушення – аналіз документів, контент-аналіз;

– методи пошуку аномалій, що мають широкий спектр застосування, в тому числі і в сфері протидії (ідентифікації) корупції як аномальної події – інструментальні дослідження в сфері науки про інформацію та тотожних дисциплін;

– експериментальні розробки в сфері провадження інформаційних технологій як засобу забезпечення належного врядування та протидії корупції. До цього напряму відносяться також публікації щодо застосування аналітико-статистичних технологій (методів, алгоритмів) для пошуку аномалій в політичній сфері чи сфері державного управління (вибори, державні закупівлі тощо) з метою отримання статистичних підтверджень наявності корупції.

1. *Аналіз та візуалізація вторинних даних.*

Дослідження [8] презентує результати використання методів машинного навчання для виявлення фактів корупції на основі розширених нелінійних моделей з високим рівнем точності прогнозування. Автори підсумували, що відкритість уряду (90,47%), захист прав власності (78,84%), належно функціонуюча судова система (77,94%), а також високий індекс освіти (53,93%) є найбільш

впливовими факторами визначення рівня корупції. Ці висновки були зроблені на основі 30 моделей, реалізованих з використанням набору показників 132 країн за 2017-2018 роки: індекс простоти ведення бізнесу, даних фондів Heritage Foundation, Transparency International, а також звітів про розвиток людського капіталу ООН.

Як зазначили дослідники, найбільшу ефективність показав метод Random Forest (точність 85,77%), метод опорного вектору (SVM) та метод нейронних мереж (artificial neural networks) показали меншу точність (76,15% та 73,84% відповідно). Для виконання запропонованої методології було використано ряд програмних засобів, зокрема Microsoft Excel, Tableau, SAS JMP і KNIME.

У наступній статті [2] про дослідження корупції в державних закупівлях було використано метод систематичного огляду літератури Торрес-Каррион. Вчені обрали 102 наукові статті, опубліковані в базах WOS та SCOPUS за 2015–2019 роки, присвячені дослідженню корупції, та здійснили аналіз із застосуванням технології штучного інтелекту. Аналіз вибраних статей дозволив авторам відповісти на наступні 4 питання:

– про методи, що використовуються для дослідження корупції в сфері державних закупівель (What methods are being applied to investigate corruption in public procurement contracts?);

– про характеристики організацій щодо яких проводяться розслідування (What are the characteristics of the organizations in which the research has been carried out?);

– про технологічні інструменти, що використовуються для дослідження виявлення та запобігання корупції в системі державних закупівель (What technological tools are being used to investigate the detection and prevention of corruption?);

– про алгоритми, методології та інструменти аналізу даних, що використовуються для виявлення корупції в системі державних контрактів (What algorithms, methodologies, and data analysis tools are used to detect corruption?).

В рамках нашого дослідження цікавими є 1, 3 та 4 питання. Як виявилося, 87% дослідників працюють із структурованими базами даних, оскільки за методологією дослідження переважають кількісні дослідження (88%). Найбільш часто використовуваний метод – кореляційний аналіз (77%). При цьому 68% досліджень розроблено саме в сфері (field of science) інформаційних технологій. Інші дослідження стосуються таких наукових напрямків як економіка, математика і статистика. Політології і соціологія у переліку не представлені. Половина (50%) досліджень сфокусовано на сфері бізнесу, 31% досліджують публічну (державну) сферу. У 89% досліджуваних статей автори використовують застосунки для персональних комп'ютерів (desktop).

Більшість дослідників (92%) використовувало для аналізу масиви структурованих даних (бази даних) з метою виявлення (descriptive) фактів корупції (79%), решта розробляли прогнозні моделі з метою попередження корупції. Серед конкретних методів аналізу найбільш застосовуваними виявились методи лінійної регресії (16%) та Байєсовські методи (16%). У меншій мірі застосовуються теорія множин, теорія графів, обробки природної мови, метод k-середніх. Серед методів отримання нового знання переважали методи класифікації (54%) та кластеризації (43%). Найпоширеніші технологічні середовища (technological tools) для аналітико-статистичних досліджень – Python Java, Matlab, Weka. Менш згадувані такі назви як R, RapidMiner, Hadoop, Spark, Neo4j, Casandra, Kafta, Visual Studio.

2. *Аналіз документів, контент-аналіз.*

У статті [11] проаналізовано рівень корупції в 52 провінціях Іспанії, які, незважаючи на єдність держави, наділені досить високим ступенем політичної та економічної автономії. Автори дослідження, використовуючи матеріали іспанської щоденної газети El Mundo, створили масив даних про кримінальні справи, пов'язані із корупційними злочинами, в яких фігурували державні службовці за період з 2000 по 2017 роки, і навчали штучну нейронну мережу "SOMs" (self-organizing maps – самоорганізаційні карти) аналізувати подібні публікацій та визначати рівні корумпованості регіонів за допомогою розрахунку відповідного показника (на 100 тисяч жителів). Під час аналізу перевірялася гіпотеза про вплив на рівень корупції рівня оподаткування (real estate taxation), державного боргу на душу населення в регіоні (debt per capita), загально боргу за послуги (debt service), економічного зростання (deposit institution growth), чисельності населення (population growth), динаміки зареєстрованих компаній (variation in the number of registered companies), цін на житло (house price increase), рівня безробіття (unemployment rate та unemployment rate growth), підтримки правлячої партії (governments ruling in majority) та періоду перебування правлячої партії при владі (number of years in government). В результаті дослідження гіпотеза про різницю між корупційними та некорупційними регіонами країни, що спричинені кореляцією корупції із обсягом інвестицій, цінами на нерухомість, кількості депозитних установ, а також тривалістю періоду перебування при владі певної партії, була підтверджена. У той же час, наявність державного боргу та кількість голосів правлячої партії не були пов'язані з рівнем корупції в регіоні.

Однієї з переваг роботи є розроблена модель підвищення ефективності заходів антикорупційної боротьби. В умовах обмежених ресурсів органи влади можуть використовувати систему раннього попередження про корупцію, яка класифікує кожний регіон країни відповідно до її профілю корупції, що дозволяє здійснювати «адресну» профілактичну та кориговану політику протидії корупції. Також модель дозволяє прогнозувати корупційні випадки задовго до їх виявлення, що посилює першочергові заходи.

Аналогічним можна вважати дослідження, в якому автори [14] представили розроблений веб-застосунок із функцією збору інформацію з новинних сайтів провінцій Індонезії, уточненням, класифікацією змісту залежно від кількості та ступеня корупційних правопорушень та візуалізації результатів на географічній мапі. Веб-застосунок було розроблено з використанням фреймворку Laravel та Google Maps API. Під час дослідження була зібрана інформація з 7 новинних сайтів Індонезії (liputan6.com, tribunnews.com, merdeka.com, kompas.com, detik.com та tampo.co) за останні останніх 7 років (2010–2017). З 900 000 статей новин близько 2 000 статей (0,2%) відносилося до категорії статей про корупцію. Рівень корупції на мапі визначався за допомогою кольору: від синього (незначні випадки корупції) до яскраво-рожевого (значні випадки корупції).

Для візуалізації збільшення або зменшення кількості випадків для кожної провінції наявна можливість розкрити хронологічну схему, яка показує щомісячну або щорічну динаміку корупційних правопорушень. В цілому, за допомогою карти корупції, за свідченнями авторів публікації, можна чітко і детально відобразити візуальну мапу корупції в Індонезії з урахуванням географічного розташування провінції та хронологічної ретроспективи Це дозволяє більш об'єктивні приймати державні рішення та розробляти ефективні програми зменшення корупції.

Аналогічну методологію було використано в роботі [5], що представила концепцію міждержавного індексу новин про корупцію (NIC) та індексу новин про боротьбу з корупцією (anti-NIC), з подальшою оцінкою впливу цих індексів. Назва індексу найбільше відображає його вміст – це карта новин про корупцію, а не про корупцію, оскільки не завжди інформація про корупцію потрапляє до новинних сайтів. Перевагою підходу, на думку авторів, є можливість відстежувати частоту змін у країнах за порівняний період, здатність охоплювати корупційні скандали, а також відсутність потреби покладатися на суб'єктивні та часто дуже ангажовані думки місцевих експертів. Однак суб'єктивність при виборі інформації для публікації на сайті та надійність опублікованої інформації не дозволяє говорити про щось більше, ніж про новини про корупційні правопорушення.

Висновки дослідження підтверджують вплив новин на корупційні акти та антикорупційні заходи, особливо в країнах перехідної економіки. В результаті дослідження було підтверджено гіпотезу про стабільний негативний вплив NIC на економічне зростання в тривалому періоді. Крім того, автори роблять висновок, що для успішних антикорупційних зусиль недостатньо використовувати лише індекс anti-NIC, проте зафіксовано позитивний вплив індексу anti-NIC на індекс NIC. Інформація про використані інформаційні технології в першоджерелі відсутня.

3. *Інструментальні дослідження.*

Дана категорія робіт суттєво відрізняється від попередніх категорій декількома параметрами.

По-перше, метою подібних публікацій є представлення власної інформаційної системи штучного інтелекту чи методу машинного навчання під час вирішення задач пошуку аномалій. Тому, ймовірно, навмисно, для збільшення аудиторії потенційних читачів, автори намагаються охопити максимальну чисельність зацікавлених осіб і декларують, що розроблені та представлені у роботі методи, алгоритми чи технології мають широке використання, починаючи від біології і закінчуючи сферою боротьби з шахрайством (вважай, корупцією).

По-друге, на відміну від публікацій попередніх категорій, роботи в даній категорії представлені у двох форматах: по-перше, як наукова (в більшості випадків у вигляді тексту статті, тез за результатами конференції) публікація; по-друге, як набір коді у соціальних мережах для розробників (наприклад, github або kaggle), які дозволяють об’єднувати зусилля різних фахівців, спілкуватися, коментувати, редагувати коди один одного з функцією слідкування за версіями коду, а також можливістю відтворювати запропоновані методи самостійно з використанням навчальних чи власних наборів даних.

В цьому плані нам достатньо буде зробити огляд найбільш популярного репозиторію на www.github.com ADBench [3]. ADBentch є спільним проектом дослідників Шанхайського університету фінансів та економіки (SUFE) і Університету Карнегі-Меллона (CMU). Проект розроблено авторами найбільш популярних бібліотек виявлення аномалій, включаючи виявлення аномалій для табличних даних чи баз даних (PyOD), часових рядів (TODS) та графів (PyGOD).

За результатами проекту було проведена оцінка продуктивності 30 алгоритмів виявлення аномалій в масивах даних з використанням 57 наборів даних в кількості 98 436 експериментів за трьома параметрами:

– типу методу машинного навчання (supervision): тести працездатності алгоритмів включають 14 алгоритмів контрольованого, 7 напівконтрольованого і 9 неконтрольованого навчання;

– характеру аномалії, що досліджується: локальні, глобальні, кластерні, залежні;

– стійкості і стабільності алгоритму в умовах наявності інформаційного шуму чи неповних даних.

Проєкт поєднав в собі розробки за трьома напрямами – виявлення аномалій в табличних даних, часових та графових.

1. PyOD представляє собою бібліотеку Python для виявлення аномальних об'єктів у багатовимірних даних. Оригінальний PyOD включає понад 40 алгоритмів виявлення, починаючи від класичного LOF до найсвіжішого ECOD [16; 19].

2. TODS – це універсальна система автоматичного машинного навчання для виявлення викидів (аномалій) в даних багатовимірних часових рядів. TODS включає модулі для побудови систем виявлення аномалій на основі машинного навчання, включаючи: обробку даних, обробку часових рядів, аналіз ознак (добування), алгоритми виявлення і модуль посилення. Функціональні можливості, надані через ці модулі, включають попередню обробку даних для загальних цілей, згладжування/перетворення даних часових рядів, витягування ознак з часових/частотних блоків даних, різні алгоритми виявлення, включаючи алгоритми залучення експертів (людського досвіду) для калібрування системи [7; 18].

3. PyGOD – бібліотека Python для виявлення викидів (аномалій) в графах. PyGOD включає в себе понад 10 алгоритмів виявлення аномалій в графах, таких як DOMINANT або GUIDE [10; 15]. Однією з переваг даного комплексу алгоритмів є простота їх застосування в сенсі обсягу використання коду – для запуску більшості алгоритмів достатньо 5 строчок коду.

4. *Експериментальні розробки.*

До цієї категорії публікації відносяться ті, що містять в собі результати безпосереднього застосування інноваційних аналітико-статистичних технологій як інструменту для виявлення суб'єктів корупційних відносин або оцінювання безпосереднього рівня корупції.

Так, в дослідженні [17] Бразильського університету та CGU (Controladoria-Gereldaunião (CGU) – орган федерального уряду Бразилії, який був створений у 2001 році, що відповідає за безпосередню допомогу президенту в питаннях щодо захисту державної власності, проведення аудиту, забезпечити прозорості та протидії корупції) описується проблема отримання корисної інформації з метою протидії корупції з федеральної бази закупівель у Бразилії, яка використовується державними аудиторами для виявлення та запобігання картельної корупції (змов).

Основними перешкодами виявлення корупції, на думку авторів, є велика кількість даних, що використовуються для вивчення відносин, а також динамічні та диверсифіковані стратегії, що використовуються компаніями для приховування їх шахрайських схем та операцій. Для вирішення цих проблем було розроблено аналітико-статистичну технологію ідентифікації картельних змов в системі державних закупівель під назвою AGMI (від "agent-mining"), де було використано такі методи як розподілений аналіз даних (distributed data mining – DDM), виявлення знань у базах даних (knowledge discovery in data bases – KDD), багатоагентні системи (multi agent systems – MAS) та інші технології розподільних обчислювань (сітка, хмара тощо), які включені в концепцію інтелектуального програмного агента (intelligent software agent – ISA).

У дослідженні автори поєднали два методи аналізу даних – DDM/KDD та MAS. За допомогою першого інструменту дослідники виявили необхідну первинну інформацію, а завдяки другого групували взаємодіючих агентів, що беруть участь у закупівлі, що дозволило авторам виявити приховані форми корупційних схем. Так, у дев'яти різних державних закупівлях, здійснених в одному і тому же штаті для однієї державної установи, алгоритм виявив процес закупівель за участю двох компаній, одна з яких вигравала усі тендерні конкурси. Дослідники звернули увагу на той факт, що програшна компанія більше не брала участь у жодних покупках, але діяла лише свого роду спаринг-партнера для переможця. Це було свідченням можливого моделювання конкуренції для маскування створення картелю. Підставна компанія, ймовірно, була створена лише для того, щоб імітувати конкуренцію в державних закупівлях, в яких конкуренція є обов'язковою. Автори дослідження показали, як система дозволяє не лише ефективно групувати дані про державні закупівлі, але й виявляти більше прихованих механізмів, що імітують конкуренцію на користь картельних змов.

До цієї категорії також можна віднести дослідження [9]. Як стверджують автори цієї публікації, організована злочинність в США змінила стратегію своєї діяльності щодо продажу наркотиків, а саме відійшла від практик нелегального розповсюдження і почала займатися менш ризикованими видами діяльності в сфері охорони здоров'я. Національна асоціація охорони здоров'я з боротьби з шахрайством (NHCAA) повідомила, що лише у Флориді державні та приватні програми медичної допомоги

втратили сотні мільйонів доларів у зв'язку із діяльністю організованої злочинності, в ході якого деякі співробітники медичних установ стають учасниками шахрайських та корупційних схем, включаючи виписування непотрібних рецептів, отримання відкатів тощо. Також учасниками змов інколи стають і пацієнти. Дослідження авторів презентує результати розробки аналітико-статистичного методу аналізу даних із використанням методу аналізу графів в сфері охорону здоров'я з метою пошуку шахрайської (корупційної) чи зловмисної поведінки.

Автори методу, за підтримки фахівців консалтингової та аудиторської компанії Xerox Services, розробили програма перевірки доброчесності – Xerox Program Integrity Validator (XPIV), що була застосована в роботі аналітиків компанії Xerox під час проведення аудитів та слідчої діяльності. Програма дає можливість: 1) автоматизувати процес створення переліку (звуження кола) підозрюваних контрагентів (automated screening); 2) отримати початкові дані про подію чи особу, які викликали у системи підозру, з можливістю прослідкувати взаємозв'язок та зібрати докази для побудови слідчої справи (interactive drill down). Автоматизований скринінг (1) фокусується на визначенні оптимального алгоритмі (методі) виявлення аномалій, а інтерактивне буріння (2) фокусується на індексації баз даних для забезпечення швидкого пошуку даних та створенні інтерфейсу користувача для інтуїтивної взаємодії користувача з інформаційною системи.

За твердженням авторів, XPIV є першою інформаційною системою, що дозволяє аналітикам в сфері шахрайства та корупції виявляти мережеве шахрайство (network-based fraud) або коротше кажучи, корупцію, атрибутивною характеристикою якої є не лише участь посадової особи в шахрайській схемі, а наявність змови з іншою (третьою) стороною. Кожен набір даних представлений як великий, гетерогенний граф, де вузли представляють мільйони пацієнтів і сотні тисяч провайдерів (лікарів, лікарні, аптеки), а ребра представляють мільярди затребуваних послуг, медикаментів і поставок, пов'язаних з множинами відносин між ними. Система шукає чотири типи аномалій в графах: 1) підозрілі особи; 2) підозрілі відносини; 3) аномальні тимчасові зміни та геопросторові характеристики, і 4) структури.

Існує два підходи до аналізу графів. Перший, відомий як егоцентричний підхід (ego-net), фокусується на окремих вузлах і особливостях розподілу зв'язків з іншими вузлами. Цей підхід дозволяє досліджувати відносин, пов'язані із отриманням наркотичних засобів, отримувати просторово-часові характеристики потоку пацієнтів між аптеками і лікарями чи лікарнями. Другий підхід аналізує глобальну структуру мережі охорони здоров'я і шукає спільноти, що мають спільні ознаки підозрілої (шахрайської) поведінки, або ті об'єднані спільноти, що набувають рис аномальності за умов аналізу агрегованих даних. Саме структурний підхід дозволяє ідентифікувати корупційні (шахрайські) мережі, такі як мережі змов або організовану злочинність.

Аномалії діляться на три категорії: аномалії індивідуального рівня, аномалії зв'язки (edge) і аномалії з недослідженою медичною поведінкою. Аналіз індивідуальних аномалій передбачає: аналіз осіб, віднесених до категорії «тяжкі споживачі наркотичних речовин» та джерел отримання ліків; лікарів, що призначають багато наркотичних речовин і кому саме; аптеки, що продають надто багато наркотичних речовин і кому саме.

Аномальні відносини можуть включати незвично сфокусовані відносини, коли: продаж наркотичних засобів аптеками здійснюється обмеженій кількості пацієнтів чи згідно рецептів обмеженого кола лікарів; лікар направляє виписує рецепт на придбання важких наркотиків в певні аптеки; і лікар призначає наркотики лише декільком пацієнтам. Висока концентрація між вузлами може бути інтерпретована як потенційна змова.

Наслідком такого аналізу є здатність швидко виявляти шахрайські схеми, які представляють інтерес для певних слідчих органів. Наприклад, система ідентифікує так званих «торгових пацієнтів», тобто пацієнтів, що відвідують велику кількість лікарень, лікарів з метою отримання рецептів на ліки із наркотичними речовинами.

Поведінкові аномалії – це ті патерни поведінки, які важко пояснити в рамках медичної практики. До них відносяться пацієнти, що споживають нічого, окрім наркотиків; або відносини пацієнт-лікар зосереджується на виписуванні рецептів лише на наркотичні ліки. Для кількісної оцінки показників здійснюється додатковий аналіз даних наданих лікарем пацієнту рецептів, що не включають наркотичні речовини, з метою визначення співвідношення та можливого факту приховування змови.

Окремим напрямом публікацій можна вважати дослідження в сфері електоральної корупції. В англомовній літературі статистичні методи, що покликані вирішити перелічені вище проблеми, називаються «electoral forensics», що можна перекласти як «електоральна криміналістика». При цьому електоральну криміналістику можна розуміти у двох значення. В широкому значенні, електоральна криміналістика включає в себе такі методи, як паралельний підрахунок голосів, спостереження за ходом голосування або перерахунок вибірки з бюлетенів після голосування, які

важко назвати суто аналітичними чи статистичними методами. У вузькому значенні електоральна криміналістика передбачає орієнтацію на аналітико-статистичні методи з мінімізацією людського чинника та оперування усіма даними загалом. Використання аналітико-статистичних методів в електоральній криміналістиці засновано на використанні принципу нормальності [4].

В рамках вузького значення електоральної криміналістики виділяються дві групи методів. Перша група веде своє походження від теорії чисел і звертається до частотних характеристик числових даних електоральної статистики. Друга група методів спирається на пошук аномалій у відносинах між різними параметрами електорального процесу, наприклад, рівнем явки та рівнем підтримки кандидатів. Основним критерієм, що використовується для виявлення фальсифікацій, є розбіжність реальних (документально зафіксованих) результатів виборів від нормативних (модельних).

У першому випадку у якості нормативної моделі виступає певний розподіл цифр, що очікується від «спонтанного» фіксування волевиявлення виборців, у другому випадку – певні відносини між загальними параметрами виборів (зазвичай – явкою) та приватними (зазвичай – частками голосів, поданих за кандидатів чи партії). В основі першої групи методів були спроби застосувати закон Бенфорда до аналізу електоральних даних. Цей закон стосується розподілу перших кількох цифр великих номерів. Виходячи з цього закону, чисел, що починаються з одиниці, завжди зустрічається більше, ніж чисел, що починаються з двійки. Кількість номерів, що починаються на цифру два, завжди більше чисел, що починаються на з цифру три, і т.д.: 3>4, 4>5, 5>6, 6>7, 7>8, 8>9. Оскільки люди не можуть ефективно створювати випадкові числа, існує значна ймовірність, що цифри, які людина вигадує сама, не будуть відповідати закону Бенфорда. Логіка застосування закону Бенфорда полягає в тому, що при появі цифр у протоколах, обраних людиною як «випадкові», ймовірність призначити останньою цифрою «круглу», наприклад, 5 або 0, буде вищою, ніж «незручну», наприклад, 8. І, аналогічно при штучному отриманні електоральних даних ймовірність зустріти парні цифри у молодших розрядах (11, 22, 33...) відрізнятиметься від очікуваної ймовірності у 1/10. Цей метод зазвичай називають методом Бебера і Скакко [1].

Друга група методів здійснює пошук аномальних залежностей між загальними (явкою) і приватними (успіх конкретного партійного списку або кандидата) показниками електоральної статистики із застосування методу моделювання. Якщо звернутися до результатів науково-дослідної роботи Мічиганського університету [4], існує три напрямки розвитку моделей оцінки результатів виборів, що дозволяють отримати статистично обґрунтовані висновки щодо електоральних фальсифікацій.

1. Першою можна вважати модель мультимодального шахрайства [6]. Згідно з концепцією авторів, базове припущення полягає в тому, що голоси на виборах без шахрайства формуються через взаємодію процесів, ефекти яких можна узагальнити за допомогою двох нормальних розподілів: одного розподілу для часток явки та іншого, незалежного розподілу для частки голосів на користь «переможця» (тобто партії з найбільшою кількістю голосів). Автори припускають, що електоральне шахрайство – це ситуація, під час якої здійснюється збільшення кількості голосів за переможця із порушенням офіційних процедур голосування. Деякі голоси переносяться до переможця від опозиції, а деякі – від тих, хто не з'явився на виборчу дільницю. При цьому автори виділяються два види електорального шахрайства: помірковане ("incremental") та жадібне ("extreme"). В першому випадку перенос голосів здійснюється обережно; в другому – перенос здійснюється для всієї загальної чисельності виборчої дільниці без додаткових розрахунків щодо тих голосів, які дійсно відбулися. Було розраховано критичні значення параметрів, що визначають ймовірність того чи іншого варіанту шахрайства: f_i – це ймовірність помірковане шахрайство, а f_e – ймовірність жадібного шахрайства. Інші параметри повністю описують бімодальні та тримодальні розподіли, які модель характеризує як наслідки електоральних шахрайств. Похідною від даної моделі є її модифікація із визначенням показника ймовірності шахрайства на рівні вибіркової дільниці. В даному випадку увага акцентується на 1) статистичних тестах щодо наявності шахрайства та 2) оцінках ймовірності того, що кожна спостережувана одиниця агрегації голосів – наприклад, кожна дільниця – є шахрайською [13].

2. Для вирішення недоліків попередньої моделі виникла модель географічної кластеризації. Індикатори чи явища, які географічно кластеризовані, заслуговують особливої уваги. Географічна кластеризація може показати, де відбувається співпраця або змова під час виборчого процесу. Географічна кластеризація також може натякнути тим, хто має відповідні експертні знання, на інші фактори, які можуть сприяти спостережуваним шаблонам у виборчих результатах. Ці інші фактори можуть бути пов'язані або непов'язані з можливістю шахрайства. Наприклад, кластер може збігатися з домашньою базою політичного лідера або з територією, на якій переважає політична партія або етнічна група лідера (або меншості). Виходячи з цього методу, використання географічних

координат дозволяє отримувати додаткові підтвердження про втручання в електоральний процес. У найбільш простому вигляді як така функція передбачається залежність від відстані між географічними точками: чим ближче розташовані виборчі дільниці (і, отже, чим ближче один до одного живуть виборці, що голосують на них), тим менше має бути відмінність у результатах голосування на цих дільницях.

На відміну від попередніх моделей, робота з географічними даними передбачає набагато більші трудовитрати при неясних перспективах дослідження. Без чіткого уявлення про географічні властивості електорату складно виправдати витрати на з'ясування адрес виборчих дільниць і тим більше зіставлення їх з географічними координатами. З урахуванням можливостей новітніх інформаційних технологій, цей метод може мати значний розвиток в разі впровадження технологій електронного голосування, що не за горами.

3. Третя модель є поєднанням попередніх двох і представлена не лише у вигляді публікацій, а у вигляді онлайн-сервісу «Інструментарій електоральної криміналісти» (Election Forensics Toolkit) [12], на якому дозволяє кожній зацікавленій особі здійснити статистичне оцінювання власних масивів електоральних даних, або подивитися на роботу із наявними в системі даними. Інформаційна система оцінює масиви за такими тестами: 2BL, LastC, C05s, P05s, Skew, Kurt, DipT. 2BL – тест середнього значення другої цифри. «Друга цифра» відноситься до другої значущої цифри в кожному підрахунку, до якого застосовується тест (наприклад, якщо підрахунок становить «1234», то «2» є другою значущою цифрою). LastC – тест середнього значення останньої цифри. «Остання цифра» відноситься до останньої цифри в кожному підрахунку, до якого застосовується тест (наприклад, якщо підрахунок становить «1234», то «4» є останньою цифрою). C05s – тест середнього значення бінарної змінної, яка вказує, чи є остання цифра підрахунку голосів для відповідної партії або кандидата нулем або п'ятіркою. P05s – тест середнього значення бінарної змінної, яка вказує, чи є остання цифра округленого відсотка голосів для відповідної партії або кандидата нулем або п'ятіркою. Skew – тест на асиметрію. Kurt – тест на ексцес. DipT – тест на унімодальність.

Слід зазначити, що жоден із проілюстрованих методів неспроможний однозначно підтвердити факт фальсифікації виборів. Вони можуть лише вказувати на наявність аномалій даних і порушувати питання про подальше розслідування. Рішення про те факт фальсифікації виборів приймається відповідними органами влади і має ґрунтуватися на додаткових дослідженнях та доказах. Звичайно, якщо ці фальсифікації не є дією привладних структур, оскільки в таких випадках країну очікує лише два ймовірних варіанти розвитку подій – Майдан або захоплення держави.

Загалом, використання алгоритмів пошуку аномалій в даних електоральної статистики може бути важливим інструментом виявлення дійсних фактів фальсифікації виборів. Однак, як зазначають дослідники, необхідно застосовувати їх з обережністю та аналізувати отримані результати у поєднанні з іншими джерелами інформації, оскільки, як показує сумнівна практика, статистичні розрахунки, що не підкріплені додатковими доказами, можуть розцінюватися зацікавленими особами, наприклад в ситуації національних виборів, не як інструмент протидії корупції, а лише як засіб делегітимації результатів виборів із відповідними наслідками і реакцією влади/опозиції.

Висновки

Отже, впровадження інноваційних аналітико-статистичних технологій як інструменту протидії корупції в державі можна розділити на 4 напрямки:

– методологічний напрям – в якому дослідники розробляють універсальні методи (алгоритми) аналізу даних (пошуку аномалій) із визначенням можливих сфер їх практичного застосування, в тому числі і в сфері протидії корупції;

– напрям вторинного аналізу даних – в якому автори намагаються розробити власні показники рівня корупції або описати стан розвитку корупційних відносин у певний період часу на базі існуючих показників;

– напрям створення автоматизованих систем аналізу текстових даних – в якому відображаються результати впровадження технологій NLP (обробки природньої мови) у поєднанні з технологіями просторової візуалізації даних;

– роботи, що присвячені результатам практичного впровадження технологій машинного навчання і штучного інтелекту для ідентифікації суб'єктів корупційних відносин та/або отримання статистично обґрунтованих підтверджень наявності / відсутності корупції.

Найбільшу значимість в даному випадку мають практико-орієнтовані дослідження, адже саме на цьому рівні відбувається безпосередня апробація відповідних теоретичних моделей і методологічного інструментарію, а також створюються прецеденти для використання результатів аналітико-статистичних досліджень як складової доказової бази тих або інших фактів шахрайства або корупції.

Yatsyna Yu. Directions for innovative analytical and statistical technologies implementation as a tool for corruption counteraction in the state

The article is dedicated to the problem of implementing innovative analytical and statistical technologies as a tool to counteract corruption in the state. Innovative analytical and statistical technologies are broadly defined as a set of methods and tools based on the use of mathematical and statistical methods of data analysis to detect useful dependencies and regularities in data, improve decision-making efficiency, and detect anomalies in various fields of activity. More narrowly, they are defined as the process of using the most advanced data analysis methods to detect complex dependencies and useful regularities in data. Based on the content analysis of thematic publications, 4 directions were identified: 1) the direction of developing methodological tools; 2) the direction of analyzing secondary sociological data; 3) the direction of creating automated systems for natural language analysis and visualization of spatial data; 4) the direction of implementing machine learning and artificial intelligence technologies for identifying subjects of corruption relations and/or obtaining statistically substantiated confirmations of the presence/absence of corruption. It is determined that the most significant direction is practice-oriented studies, where testing of the relevant theoretical models and methodological tools takes place, and precedents are created for the use of the results of analytical and statistical studies as part of the evidence base of fraud or corruption. It is determined that a direct example of applying innovative analytical and statistical technologies as a tool for corruption counteraction from a political science perspective is electoral forensics. Within electoral forensics, two groups of methods are distinguished. The first group originates from the theory of numbers and refers to the frequency characteristics of numerical data of electoral statistics. The second group of methods relies on the search for anomalies in the relationships between different parameters of the electoral process, for example, turnout level and winning candidate support level. The main criterion used to detect electoral corruption is the discrepancy between the real (documented) election results and the normative (model) ones.

Key words: corruption counteraction, innovative technologies, machine learning, artificial intelligence, electoral forensics.

Література:

1. Beber B., Scacco A. What the Numbers Say: A Digit-Based Test for Election Fraud. *Political Analysis*. 2012. #20(2). PP. 211–234. URL: https://doi.org/10.1093/pan/mps003 (дата звернення: 10.04.2023)
2. Berru Y.T., Batista V.F.L., Torres-Carrión P., Jimenez M.G. Artificial Intelligence Techniques to Detect and Prevent Corruption in Procurement: A Systematic Literature Review. *ICAT 2019, CCIS*. 2020. #1194. PP. 254–268. URL: https://doi.org/10.1007/978-3-030-42520-3_21 (дата звернення: 10.04.2023)
3. Han S., Hu X., Huang H., Jiang M., Zhao Y. ADBench: Anomaly Detection Benchmark. *NeurIPS 2022*. 2022. #45. URL: https://doi.org/10.48550/arXiv.2206.09426 (дата звернення: 10.04.2023)
4. Hicken A., Mebane W.R.J. A Guide to Elections Forensics: Research and Innovation Grants Working Papers Series. 2015. URL: https://electionforensics.cps.isr.umich.edu/pdf/guide.pdf (дата звернення: 10.04.2023)
5. Hlatshwayo S., Oeking A., Ghazanchyan M., Corvino D., Shukla A., Leigh L. The Measurement and Macro-Relevance of Corruption: A Big Data Approach. Washington, D.C.: International Monetary Fund, 2018. 73 p.
6. Klimek P., Yegorov Y., Hanel R., Thurner S. Statistical detection of systematic election irregularities. *Proc Natl Acad Sci USA*. 2012. #109(41). PP. 16469–16473. URL: https://doi.org/10.1073/pnas.1210722109 (дата звернення: 10.04.2023)
7. Lai K.-H., Zha D., Wang G., Xu J., Zhao Y., Kumar D., ... Hu X. TODS: An Automated Time Series Outlier Detection System. 2021. URL: https://doi.org/10.48550/arXiv.2009.09822 (дата звернення: 10.04.2023)
8. Lima M.S.M., Delen D. Predicting and explaining corruption across countries: A machine learning approach. *Government Information Quarterly*. 2020. #37(1). URL: https://doi.org/10.1016/j.giq.2019.101407 (дата звернення: 10.04.2023)
9. Liu J., Bier E., Wilson A., Guerra-Gomez J.A., Honda T., Sricharan K., ... Davies D. Graph Analysis for Detecting Fraud, Waste, and Abuse in Health-Care Data. *AI MAGAZINE*. 2016. PP. 33–46.

10. Liu K., Dou Y., Zhao Y., Ding X., Hu X., Ding R.Z.K., ... Yu P.S. BOND: Benchmarking Unsupervised Outlier Node Detection on Static Attributed Graphs. *NeurIPS 2022*. 2022. URL: https://doi.org/10.48550/arXiv.2206.10071 (дата звернення: 10.04.2023)
11. López-Iturriaga F.J., Sanz I.P. Predicting Public Corruption with Neural Networks: An Analysis of Spanish Provinces. *Social Indicators Research*. 2017. #140(3). PP. 975–998. URL: https://doi.org/10.1007/s11205-017-1802-2 (дата звернення: 10.04.2023)
12. Mebane W.R.J., Kalinin K. Guide to Election Forensics Toolkit. URL: https://electionforensics.cps.isr.umich.edu/election (дата звернення: 10.04.2023)
13. Mebane W.R.J., Wall J. Election Frauds, Postelection Legal Challenges and Geography in Mexico. URL: http://www.umich.edu/~wmebane/apsa15.pdf (дата звернення: 10.04.2023)
14. Noerlina, Dewanti R., Mursitama T.N., Fairianti S.P., Kristin D.M., Sasmoko, ... Makalew B.A. Development of a Web Based Corruption Case Mapping using Machine Learning with Artificial Neural Network. *International Conference on Information Management and Technology (ICIMTech)*. 2018. PP. 400–405.
15. PyGOD. *Github.com.* URL: https://github.com/pygod-team/pygod (дата звернення: 10.04.2023)
16. Python Outlier Detection (PyOD). *Github.com*. URL: https://github.com/yzhao062/pyod (дата звернення: 10.04.2023)
17. Ralha C.G., Silva C.V.S. A multi-agent data mining system for cartel detection in Brazilian government procurement. *Expert Systems with Applications*. 2012. #39(14). PP. 11642–11656. URL: https://doi.org/10.1016/j.eswa.2012.04.037 (дата звернення: 10.04.2023)
18. TODS: Automated Time-series Outlier Detection System. *Github.com*. URL: https://github.com/datamllab/tods (дата звернення: 10.04.2023)
19. Zhao Y., Nasrullah Z., Li Z. PyOD: A Python Toolbox for Scalable Outlier Detection. *Journal of Machine Learning Research*. #20. PP. 1–7. URL: https://doi.org/10.48550/arXiv.1901.01588 (дата звернення: 10.04.2023)

References:

1. Beber B., Scacco A. What the Numbers Say: A Digit-Based Test for Election Fraud. *Political Analysis*. 2012. #20(2). PP. 211–234. URL: https://doi.org/10.1093/pan/mps003 (retrieval date: 10.04.2023)
2. Berru Y.T., Batista V.F.L., Torres-Carrión P., Jimenez M.G. Artificial Intelligence Techniques to Detect and Prevent Corruption in Procurement: A Systematic Literature Review. *ICAT 2019, CCIS*. 2020. #1194. PP. 254-268. URL: https://doi.org/10.1007/978-3-030-42520-3_21 (retrieval date: 10.04.2023)
3. Han S., Hu X., Huang H., Jiang M., Zhao Y. ADBench: Anomaly Detection Benchmark. *NeurIPS 2022*. 2022. #45. URL: https://doi.org/10.48550/arXiv.2206.09426 (retrieval date: 10.04.2023)
4. Hicken A., Mebane W.R.J. A Guide to Elections Forensics: Research and Innovation Grants Working Papers Series. 2015. URL: https://electionforensics.cps.isr.umich.edu/pdf/guide.pdf (retrieval date: 10.04.2023)
5. Hlatshwayo S., Oeking A., Ghazanchyan M., Corvino D., Shukla A., Leigh L. The Measurement and Macro-Relevance of Corruption: A Big Data Approach. Washington, D.C.: International Monetary Fund, 2018. 73 p.
6. Klimek P., Yegorov Y., Hanel R., Thurner S. Statistical detection of systematic election irregularities. *Proc Natl Acad Sci USA*. 2012. #109(41). PP. 16469–16473. URL: https://doi.org/10.1073/pnas.1210722109 (retrieval date: 10.04.2023)
7. Lai K.-H., Zha D., Wang G., Xu J., Zhao Y., Kumar D., ... Hu X. TODS: An Automated Time Series Outlier Detection System. 2021. URL: https://doi.org/10.48550/arXiv.2009.09822 (retrieval date: 10.04.2023)
8. Lima M.S.M., Delen D. Predicting and explaining corruption across countries: A machine learning approach. *Government Information Quarterly*. 2020. #37(1). URL: https://doi.org/10.1016/j.giq.2019.101407 (retrieval date: 10.04.2023)
9. Liu J., Bier E., Wilson A., Guerra-Gomez J.A., Honda T., Sricharan K., ... Davies D. Graph Analysis for Detecting Fraud, Waste, and Abuse in Health-Care Data. *AI MAGAZINE*. 2016. PP. 33–46.
10. Liu K., Dou Y., Zhao Y., Ding X., Hu X., Ding R.Z.K., ... Yu P.S. BOND: Benchmarking Unsupervised Outlier Node Detection on Static Attributed Graphs. *NeurIPS 2022*. 2022. URL: https://doi.org/10.48550/arXiv.2206.10071 (retrieval date: 10.04.2023)
11. López-Iturriaga F.J., Sanz I.P. Predicting Public Corruption with Neural Networks: An Analysis of Spanish Provinces. *Social Indicators Research*. 2017. #140(3). PP. 975–998. URL: https://doi.org/10.1007/s11205-017-1802-2 (retrieval date: 10.04.2023)
12. Mebane W.R.J., Kalinin K. Guide to Election Forensics Toolkit. URL: https://electionforensics.cps.isr.umich.edu/election (retrieval date: 10.04.2023)
13. Mebane W.R.J., Wall J. Election Frauds, Postelection Legal Challenges and Geography in Mexico. URL: http://www.umich.edu/~wmebane/apsa15.pdf (retrieval date: 10.04.2023)

14. Noerlina, Dewanti R., Mursitama T.N., Fairianti S.P., Kristin D.M., Sasmoko, ... Makalew B.A. Development of a Web Based Corruption Case Mapping using Machine Learning with Artificial Neural Network. *International Conference on Information Management and Technology (ICIMTech)*. 2018. PP. 400–405.
15. PyGOD. *Github.com.* URL: https://github.com/pygod-team/pygod (retrieval date: 10.04.2023)
16. Python Outlier Detection (PyOD). *Github.com*. URL: https://github.com/yzhao062/pyod (retrieval date: 10.04.2023)
17. Ralha C.G., Silva C.V.S. A multi-agent data mining system for cartel detection in Brazilian government procurement. *Expert Systems with Applications*. 2012. #39(14). PP. 11642–11656. URL: https://doi.org/10.1016/j.eswa.2012.04.037
18. TODS: Automated Time-series Outlier Detection System. *Github.com*. URL: https://github.com/datamllab/tods (retrieval date: 10.04.2023)
19. Zhao Y., Nasrullah Z., Li Z. PyOD: A Python Toolbox for Scalable Outlier Detection. *Journal of Machine Learning Research*. #20. PP. 1–7. URL: https://doi.org/10.48550/arXiv.1901.01588 (retrieval date: 10.04.2023)

Стаття надійшла до редакції 28.03.2023
Стаття рекомендована до друку 13.04.2023