УДК [001.103-022.56+004.77+004.8]:311.3-048.22
JEL Classification: C81, C88, L82 **Doi: 10.31767/su.1(100)2023.01.01**

О. Г. Осауленко,
доктор наук з державного управління, професор,
член-кореспондент НАН України,
ректор,
E-mail: O.Osaulenko@nasoa.edu.ua
ORCID: https://orcid.org/0000-0002-7100-7176;
О. О. Горобець,
кандидат економічних наук,
доцент кафедри статистики, інформаційних технологій
та математичних методів в економіці,
E-mail: babutska@ukr.net
ORCID: https://orcid.org/0000-0001-5433-6448;
Національна академія статистики, обліку та аудиту

Імплементація інструментарію Smart-статистики в офіційну статистику

Розглянуто актуальні питання Smart-статистики. Досліджено ключовий інструментарій Smart-статистики: великі дані, дані штучного інтелекту та Інтернету речей, соціальних медіа та адміністративні дані.

У ході дослідження сформульовано концепцію Smart-статистики та виявлено переваги та загрози для офіційної статистики при використанні даних Smart-статистики. Запропоновано низку принципів роботи з великими даними, згруповано джерела вироблення великих даних за категоріями (відповідно до видів економічної діяльності), які нині затребувані як офіційною статистикою, так і суспільством (сільське господарство, охорона здоров'я, гірнича промисловість, машинобудування, освіта, енергетика). Розроблено систему розумних даних у розумних містах, яка містить такі елементи: розумний будинок, розумне середовище, розумний контроль, розумний трафік, розумне здоров'я та розумний громадянин. Визначено, що метою цієї системи є формування розумного середовища та спрощення способу життя шляхом економії часу, енергії і коштів. Визначено складові персоналізованого збирання та розповсюдження даних.

Розглянуто штучний інтелект (ШІ) як компонент концепції Smart-статистики. У контексті статті автори зазначають, що використання технологій ШІ викликає етичні міркування щодо таких питань, як право власності на дані, їх прозорість і підзвітність. Ці міркування необхідно враховувати, аби переконатися, що використання технологій ШІ в офіційній статистиці є відповідальним і етичним.

За результатами дослідження сформульовано висновок, що наразі необхідно сформувати логічні та адекватні підходи до розроблення методології збирання, опрацювання, групування й аналізу статистичних даних з альтернативних інформаційних джерел. Обґрунтовано, що, незважаючи на усі переваги інструментів Smart-статистики та потенційно позитивні результати їх імплементації в офіційну статистику, з огляду на умови війни й відсутність нормативно-правового поля щодо використання цих інструментів Україна наразі не готова зіштовхнутися із тими загрозами, які вони в собі несуть.

Подальші дослідження за розглянутою у статті тематикою мають спрямовуватися на поглиблення вивчення Smart-статистики та пошуки оптимальних шляхів упровадження її інструментів в офіційну статистику України.

Ключові слова: *Smart-статистика, офіційна статистика, великі дані, штучний інтелект, Інтернет речей, соціальні медіа, адміністративні дані.*

Вступ. Нещодавно Віце-прем'єр-міністр з інновацій, розвитку освіти, науки та технологій України – Міністр цифрової трансформації України М. Федоров запропонував бачення трансформації Державної служби статистики, наголосивши при цьому, що ухвалення управлінських рішень і відновлення України повинні базуватися на якісних даних. Ядро трансформаційної концепції, на його думку, полягає у такому: "У центрі всіх процесів Держстату має бути IT-платформа, яка в режимі реального часу збирає інформацію, аналізує і поширює" [1]. Таке бачення майбутнього офіційної статистики відкриває нові інформаційні горизонти шляхом залучення новітніх інноваційних технологій, водночас провокує появу нових професійних і соціальних загроз, у тому числі у вигляді вивільнення великої кількості фахівців-статистиків.

Зазначене змушує як науковців, так і практиків у сфері статистики до ґрунтовного вивчення й аналізу наявних підходів до трансформації статистичної галузі з метою визначення логічних кроків

у розробленні методології збирання, опрацювання, групування та аналізу даних з альтернативних джерел. Це, своєю чергою, сформувало мету пропонованого дослідження.

Аналіз останніх досліджень і публікацій. В останні десятиліття сукупність новітніх технологічних розробок підштовхнула глобальний процес цифровізації всього суспільства, ключовими віхами якої стала побудова Інтернету і Всесвітньої павутини, поява повсюдних соціальних онлайн-мереж, поширення смартфонів та інших розумних пристроїв, а останнім часом – розробка так званого Інтернету речей (IoT) і безумовно – штучного інтелекту [2].

K. Cukier та V. Mayer-Schoenberger ще у 2013 р. зазначали, що за останні два десятиліття внаслідок бурхливого розвитку цифровізації та інтелектуалізації [3] стає доступною велика кількість нових типів джерел цифрових даних.

З огляду на зазначене доцільно навести слова F. Ricciato та ін., які у своєму дослідженні акцентували увагу на тому, що місія офіційної статистики полягає в забезпеченні кількісного представлення суспільства, економіки та навколишнього середовища для цілей суспільного інтересу, для розробки політики й оцінювання, а також як основи для інформування [4], а отже, офіційна статистика повинна йти пліч-о-пліч із розвитком суспільства, економіки та довкілля.

Ідеї трансформації статистики відповідно до нових вимог суспільства, які залежать від історичних, інституційних та культурних умов, не нові. Статистика як дзеркало суспільства постійно трансформується, тою чи іншою мірою відображаючи зміни, якими характеризується соціум.

У контексті вищезгаданої заяви М. Федорова варто зазначити, що у червні 2018 р. представниками Євростату під час Європейської Конференції з офіційної статистики оприлюднено доповідь "Надійна Smart-статистика: роздуми про майбутнє (офіційної) статистики". Її автори зауважили, що статистику можна розглядати як технологію, яка дозволяє розробляти політику на основі фактів, і в рамках концепції Інтернету речей функціонал статистики можна масштабувати [5]. За підсумками Конференції, для позначення еволюції офіційної статистики Євростатом запропоновано термін Trusted Smart Statistics – надійна Smart-статистика та окреслено майбутню розширену роль офіційної статистики у світі, насиченому інтелектуальними технологіями. Безпосередньо надійну Smart-статистику запропоновано розглядати як послугу, що надається інтелектуальними системами шляхом вбудовування перевірених та прозорих життєвих циклів даних, забезпечуючи при цьому достовірність і точність вихідних даних, поважаючи і захищаючи конфіденційність суб'єктів даних. Зазначалося, що інтелектуальні технології охоплюють автоматизовані інтерактивні технології у режимі реального часу, які оптимізують фізичну роботу приладів і споживчих пристроїв, у зв'язку з цим статистика буде перетворена на розумну технологію, вбудовану в розумні системи, яка перетворюватиме дані на інформацію [6].

Згодом, у жовтні 2018 р., керівниками статистичних відомств країн-членів Європейського Союзу (ЄС) було прийнято "Бухарестський меморандум про офіційну статистику в суспільстві, що базується на даних (надійна Smart-статистика)" [7]. Бухарестський меморандум слідував за Схевенінгенським меморандумом від 2013 р. про великі дані та офіційну статистику. І якщо останній ознаменував початок узгодженого вивчення великих даних для використання в офіційній статистиці в рамках Європейської статистичної системи [8], то Бухарестський меморандум оцінив досягнення станом на 2018 р. та визначив пріоритети для її подальшого розвитку, що базуються на даних надійної розумної статистики.

Бухарестським меморандумом визначено такі завдання Генеральних директорів національних статистичних інститутів та Євростату:

– подальше вдосконалення правової бази на європейському та національному рівнях з метою зменшення перешкод для доступу, використання й інтеграції різнорідних даних для створення та розвитку надійної інтелектуальної статистики;

– заохочення створення спеціалізованих статистичних smart-спільнот для забезпечення обміну знаннями та навичками, а також сталої імплементації досягнень і подальших розробок;

– застосування штучного інтелекту в офіційній статистиці та використання інтелектуальних технологій, що вимагає прийняття нових стандартів, які можуть забезпечити сумісність між різними системами даних, і дотримання етичних, правових правил і принципів прозорості та якості.

Також у Меморандумі визначена необхідність європейської і міжнародної координації між національними статистичними інститутами, Євростатом, ООН та іншими міжнародними статистичними органами з метою розробки ефективної платформи співпраці для забезпечення синергії, узгодження, оптимізації діяльності й комунікаційних ініціатив на європейському та світовому рівнях [7].

У 2020 р. W. J. Radermacher визначив своєрідну нову еру офіційної статистики і назвав її "Офіційна статистика 4.0", зважаючи на її трансформації від цифровізації до глобалізації. У своїй роботі науковець зазначає: те, що чекає на нас (статистиків – авт.) у майбутньому, нелегко передбачити. Безумовно, передбачення не закладено в самій природі статистиків, які зазвичай дивляться у дзеркало заднього виду. Оскільки офіційна статистика має характеристики океанського лайнера, курсом і швидкістю якого можна маневрувати лише по-

вільно, усі тенденції слід тлумачити перспективно. Якщо офіційна статистика хоче підтримувати нинішню позицію через п'ять років, то необхідна стратегія має бути встановлена зараз [9].

На думку F. Ricciato зі співавторами, в більшості країн нові джерела даних доповнюватимуть, але не замінюватимуть застарілі, встановлені компоненти статистичних систем не будуть відкинуті, а навпаки, доповнені новими. За аналогією, джерела даних можна розглядати як паливо, а статистичну систему – як двигун: нове паливо не може бути подано в застарілий механізм і статистичним системам потрібно розробити новий тип двигуна з принципами роботи, відмінними від традиційних, адаптований до особливостей нового палива – даних. Майбутня статистична система врешті-решт буде багатопаливною машиною з двома двигунами: надійна Smart-статистика й традиційна статистика [4].

Зазначене обумовлює необхідність формулювання концепції Smart-статистики та виокремлення основних інструментів її функціонування.

Результати дослідження та їх обговорення. Основні джерела даних для надійної Smart-статистики мають задовольняти вимогам, висунутим до даних державою та суспільством у сучасних умовах. Зважаючи на це, авторами сформульовано концепцію Smart-статистики (рис. 1), яка базується на залученні ресурсів інтернет-технологій.

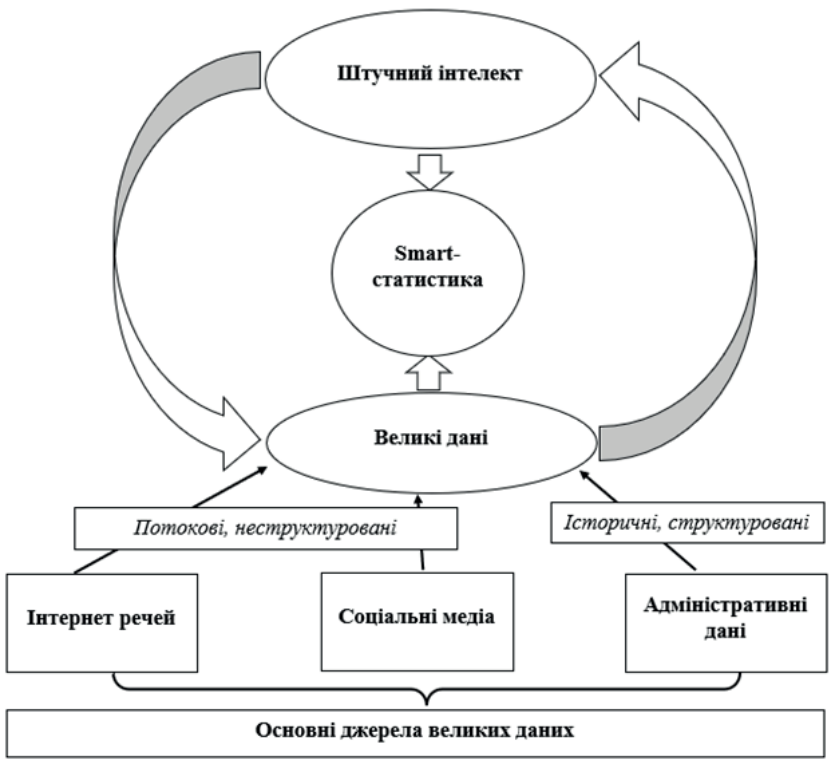

Рис. 1. Концепція Smart-статистики

Передбачається, що Smart-статистика має використовувати:

– дані Інтернету речей (Internet of Things, IoT) і дані соціальних медіа, які утворюють групу потокових даних в екосистемі великих даних, згідно з Класифікацією великих даних, розробленою Європейською економічною комісією ООН;

– адміністративні дані й дані, які виникають унаслідок роботи штучного інтелекту (ШІ, Artificial Intelligence, AI), алгоритми якого функціонують на основі великих даних.

Отже, основними складовими концепції Smart-статистики є IoT, соціальні медіа й адміністративні дані як глобальні групи екосистеми великих даних в цілому, а також ШІ. При цьому необхідно зважати на потенційно сумнівну якість даних, отриманих із зазначених вище джерел, що може знижувати їхню вартість як для офіційної статистики, так і для суспільства. Очевидно, що не усі згенеровані дані є цінними та необхідними для статистики. Зазначимо, що вартість даних формується у результаті їх генерації джерелами, збирання, зберігання й аналізу (рис. 2, адаптовано авторами за даними [10]).

Розглянемо детальніше кожен із елементів концепції Smart-статистики (див. рис. 1), зосередившись при цьому на їхніх перевагах та загрозах (або недоліках). Щодо одного з елементів – великих даних, – то дещо повільна імплементація в офіційну статистику пояснюється їх глобальністю та в цілому низькою якістю. Водночас Євростатом досі не оприлюднено чітких вказівок щодо збирання, обробки, групування чи аналізу великих даних, хоча їх джерел щороку стає дедалі більше.

Рис. 2. Формування вартості даних: обов'язкові етапи та можливі складові

Схевенінгенським меморандумом 2013 р. про великі дані та офіційну статистику визначено, що великі дані – це явище, яке впливає на багато сфер політики [8]. У документі зазначається, що використання великих даних в офіційній статистиці вимагає нових розробок у методології, оцінці якості та питаннях, пов'язаних з ІТ. З-поміж іншого йдеться про те, що використання великих даних для статистичних цілей кидає виклик європейській статистиці, разом з тим великі дані слугують своєрідною системою для ефективного вирішення різноманітних проблем. Офіційна статистика повинна включати якомога більше всіх потенційних джерел даних, включаючи великі дані [8].

На сьогодні існує велика кількість розробок щодо використання великих даних в офіційні статистиці. У процесі їх імплементації в офіційну статистику обов'язково слід дотримуватися Фундаментальних принципів офіційної статистики. При цьому корисними стануть принципи аналітики великих даних [11] та 12 правил Кодда [12], які вже вважаються класичними при побудові систем керування базами даних.

З огляду на зазначене та враховуючи нагальну необхідність використання великих даних офіційною статистикою, автори пропонують такі принципи роботи з великими даними:

1. Процес збирання великих даних повинен гарантувати безпеку та збереження їх конфіденційності, а також забезпечувати перевірку на надійність та якість цих даних.

2. Необхідно забезпечити інтелектуальний супровід збирання й аналізу великих даних, оприлюднення та поширення висновків, що ґрунтуються на результатах їх оброблення.

3. Норми та закони щодо збирання, оброблення й аналізу великих даних, а також оприлюднення та поширення висновків на основі результатів їх аналізу мають бути узгоджені й оприлюднені у форматі чотирьохсторонньої групи "держава – наука – бізнес – статистика".

4. Для членів групи "держава – наука – бізнес – статистика" має бути гарантовано вільний доступ до державних національних і приватних сховищ великих даних.

5. Напрацювання статистичними відомствами методологій та рекомендацій щодо збирання, опрацювання й аналізу великих даних має відбуватися з урахуванням специфіки кожної їхньої окремої групи.

6. Має бути забезпечено узгодження й оприлюднення алгоритмізації основних процесів збирання, оброблення та аналізу великих даних [13].

Дотримання вищезазначених принципів дозволить напрацьовувати та надалі дотримуватися єдиних підходів на шляху від неординарних пропозицій щодо використання великих даних до адекватних управлінських рішень щодо імплементації великих даних у різні сфери життєдіяльності.

З огляду на зазначене, необхідно згрупувати різні джерела вироблення великих даних за категоріями (за видами економічної діяльності), які нині затребувані як офіційною статистикою так і суспільством у цілому (табл. 1, згруповано авторами).

Аналіз табл. 1 свідчить, що дані з IoT можуть ефективно використовуватись у статистиці різних видів економічної діяльності. Вже зараз можна стверджувати, що великі дані є рушійною силою інноваційних рішень, а їх застосування зростає завдяки новітнім технологіям, комунікаційним мережам та обчислювальним потужностям, що своєю чергою стимулює розвиток технологічних екосистем. Зазначене підтверджується звітом Cisco, згідно з яким у 2023 р. у світі підключені додатки для дому матимуть 48% домогосподарств, а автомобільні додатки щороку зростатимуть на 30% протягом періоду прогнозування (2018–2023 рр.). Щоденний обсяг генерованих з допомогою IoT даних у 2023 р. зросте до понад 2,5 квінтильйонів байт (потокових, історичних, масивних) [22], які необхідно обробляти, перетворювати й аналізувати на постійній основі.

Прискорений розвиток IoT вважається одним із головних проривів у інформаційно-комунікаційних технологіях, ключовим фактором, що сприяє

Таблиця 1

Можливості імплементації великих даних в офіційну статистику за видами економічної діяльності

Вид економічної діяльності	Джерела великих даних	Отримання та використання даних
1	2	3
Сільське господарство	IoT (розумні трактори, вантажівки, дрони, оснащені спеціальними датчиками для збирання агрономічних даних), супутникові дані	Агрономічні дані, зібрані з ґрунтів, рослин та ін. допоможуть визначати існуючі проблеми, зокрема щодо ефективності роботи, якості ґрунту, відстежувати процес посівів протягом року, у результаті чого можна буде побудувати прогнозні алгоритми й мобільні додатки[1], які дозволятимуть попередити й уникнути проблем ще до того, як вони виникнуть, зокрема налагодити ефективне використання добрив і пестицидів [2]. Важливою є також можливість застосування блокчейн-рішень для відстеження шляху продуктів від поля до полиць супермаркетів
Охорона здоров'я	Медичні картки, дані клінічних випробувань, носимі пристрої	Найпоширеніші джерела великих даних про здоров'я – електронні медичні картки (або записи), прогностична аналітика, результати клінічних випробувань, дані спостережень за станом здоров'я, дані, отримані з носимих пристроїв (wearables devices). Використання цих даних сприятиме удосконаленню спостережень за станом здоров'я населення у частині моніторингу й відстеження поширення інфекційних захворювань, даючи змогу працівникам сфери охорони здоров'я своєчасно виявляти спалахи пандемій та вживати заходів для запобігання поширенню хвороб
Гірнича промисловість	GPS (Global Positioning System, або система глобального позиціонування), геодезичні вимірювання, буропідривні операції, обслуговування залізничних колій	Дані датчиків, які використовуються для моніторингу різних аспектів процесу видобутку, таких як продуктивність обладнання, геофізичні параметри, умови навколишнього середовища та параметри безпеки; геопросторові дані, включаючи дані супутникових зображень, дані LiDAR (Light Identification, Detection and Ranging – технологія отримання й обробки інформації про віддалені об'єкти з допомогою активних оптичних систем), геологічні дані, критично важливі для гірничодобувних компаній для ідентифікації та картографування родовищ корисних копалин, планування й оптимізації макетів шахт, моніторингу змін у землекористуванні та відстеженні впливу на довкілля [16]; операційні дані – дані, отримані в результаті операційних процесів, наприклад дані про виробництво, технічне обслуговування, ланцюги поставок і робочу силу. Аналіз цих даних може надати уявлення про ефективність виробництва, використання обладнання, графіки технічного обслуговування та продуктивність праці, що в результат має сприяти покращенню умов роботи та економії коштів
Виробництво автотранспортних засобів	IoT, VR, AR, датчики, напівпровідникові та інтегральні схеми [17], робототехнічні програми, радіочастотні ідентифікатори, пристрої для зчитування штрих-кодів	Ці джерела генерують величезні обсяги даних, пов'язаних із продуктивністю автомобіля, поведінкою водія, місцем розташування тощо. Такі дані можна збирати, обробляти й аналізувати в режимі реального часу з метою дистанційної діагностики, прогнозного обслуговування та надання персоналізованих послуг. Дані, отримані завдяки розумним автомобілям також можна використовувати для покращення керування дорожнім рухом, зменшення викидів і підвищення безпеки на дорогах. Такі дані використовують для моделювання поведінки споживачів та планування виробництва на їх основі, підвищення ефективності маркетингу, контролю продуктивності, підтримки якості й безпеки продукції [18], автоматичного оновлення автомобільних карт, аналізу ефективності двигуна

1 Так, компанія Soufflet Agriculture Farmi пропонує повну сервісну платформу, доступну в будь-який час. Фермер має єдину програму на телефоні, щоб керувати своєю фермою й отримувати користь від інформації, вигідних пропозицій продуктів і постійного доступу до найкращих можливостей продажу. Ця програма надає широкий спектр даних: агрометеорологія, агрономічні консультації, котирування, новини ринку, онлайн-продажі врожаю тощо [14].

2 Digital Transmission network (DTN) надає своїм клієнтам інформаційні рішення для сільського господарства та аналізу ринків. Використовуючи DTN, фермери та трейдери можуть отримати доступ до актуальних даних про погоду та ціни, щоб краще керувати своїм бізнесом [15].

Продовження табл. 1

1	2	3
Освіта	Системи управління навчанням (Learning management system, LMS), інформаційні системи для студентів (SIS), освітні програми та інструменти, соціальні медіа та онлайн-спільноти, відкриті освітні ресурси; IoT	Платформи LMS, такі як Blackboard, Canvas і Moodle, GoogleClassroom та ін. збирають і зберігають дані, пов'язані з опануванням студентами змісту курсів, оцінюванням і обговореннями, надаючи дані про залученість студентів, результативність освітнього процесу і навчальну поведінку. Платформи SIS, такі як PowerSchool, Infinite Campus і Skyward, збирають і зберігають дані, що характеризують студентів, процес зарахування, відвідуваність, оцінки й іншу адміністративну інформацію, яку можна використовувати для різноманітних аналітичних цілей [19]. Освітні програми й інструменти (адаптивні навчальні платформи, онлайн-системи репетиторства та віртуальні навчальні середовища) генерують дані про взаємодію, успішність і вподобання студентів, доцільні для персоналізації навчального досвіду та покращення результатів освіти. Соціальні медіа-платформи, онлайн-спільноти та дискусійні форуми генерують дані про взаємодію студентів, співпрацю та поведінку в соціальному навчанні, що дозволить сформувати уявлення про залучення студентів, соціальну динаміку та результати навчання. Цифрові підручники, відео та інтерактивне моделювання у стають джерелами даних про залученість студентів, використання вмісту навчальних матеріалів та результати навчання й потрібні для розробки персоналізованих підходів до навчання. Розумні класи та розумні кампуси генерують дані про фізичну активність студентів, умови довкілля та інші контекстні фактори, здатні впливати на навчання [20]
Енергетика	IoT, відновлювані джерела енергії, системи енергоменеджменту, системи накопичення енергії	Розумні лічильники та датчики мережі збирають і передають великі обсяги даних про споживання електроенергії, напругу, струм, коефіцієнт потужності й інші параметри мережі. Дані з цих джерел можна використовувати для прогнозування навантаження, реагування на попит, оптимізації мережі та прогнозного обслуговування енергетичної інфраструктури. Пристрої та устаткування, що застосовують відновлювані джерела енергії (сонячні батареї, вітряні турбіни, гідроелектростанції), дозволяють отримувати дані про виробництво енергії, продуктивність та ефективність, які можна застосовувати для оптимізації енерговиробництва, моніторингу справності обладнання та покращення операційних стратегій. Системи енергоменеджменту, які використовуються в комерційних і промислових будівлях, наприклад системи керування будівлями (BMS), промислові системи керування (ICS), системи HVAC (опалення, вентиляції та кондиціонування повітря), освітлення, інші будівельні системи генерують дані про споживання енергії. Ці дані можна використовувати для оптимізації енергоспоживання, зменшення втрат енергії, підвищення ефективності будівлі [21]. Системи накопичення енергії, такі як батареї та гідроакумулятори з насосом, дозволяють отримувати дані про ємність накопичення енергії, швидкість заряджання, розряджання та ефективність цих пристроїв. Ці дані також можна використовувати для оптимізації операцій накопичення енергії, підвищення продуктивності системи накопичення енергії та підтримки стабільності мережі

створенню технологічних рішень для суспільства, громадян, підприємств та урядів, а також використовує переваги великих даних і методів машинного навчання [23]. Так, використання засобів IoT створило своєрідну систему розумних даних у розумних містах (рис. 3, побудовано авторами), метою якої є формування розумного середовища та спрощеного способу життя шляхом економії часу, енергії та коштів. З огляду на стрімкий розвиток засобів IoT можна передбачити, що згодом ця система буде доповнена розумною культурою, розумною освітою, розумним житлово-комунальним господарством та іншими розумними сферами життєдіяльності міста.

У цій статті увагу акцентовано здебільшого на перевагах використання альтернативних джерел даних, що є беззаперечним. Однак у процесі імплементації цих віртуальних джерел даних у процеси офіційної статистики або у сферу управління необхідно також мати на увазі їх недоліки, які мо-

Рис. 3. Система розумних даних у розумних містах

жуть призвести до появи цілком реальних загроз. Так, A. Bhardwaj у роботі [24] описав низку загроз, які можуть бути пов'язані безпосередньо з IoT:

– програми-вимагачі, запущені через шкідливі програми;

– атаки сканування та відображення у дротових і бездротових мережах;

– атаки на мережеві протоколи;

– зараження інтелектуальних систем промислового управління, систем диспетчерського контролю та збирання даних;

– атаки на криптографічні алгоритми й управління ключами;

– атаки з пошкодженням даних;

– атаки задля порушення цілісності операційної системи та програм;

– відмова в обслуговуванні та його блокування.

Варто зазначити, що ЄС позиціонує IoT як пріоритет єдиного цифрового ринку. Також IoT є предметом низки горизонтальних політик, серед яких захист даних та кібербезпека. Зокрема у Робочому документі ЄС "Європейська стратегія щодо даних" зазначається, що потенційною перешкодою для досягнення єдиного ринку IoT є необхідність працювати з дуже великими обсягами найрізноманітніших підключених пристроїв, надійно ідентифікувати їх (пристрої – авт.) та за можливості – виявляти, щоб надалі підключати до відповідних систем IoT [25].

Ще однією складовою Концепції Smart-статистики є соціальні медіа, які нині глобально впроваджені. Платформи Facebook (2,9 млрд активних користувачів), Instagram (2 млрд), Snapchat (750 млн), TikTok (1 млрд) і YouTube (2,5 млрд) набули значної популярності та мають величезну кількість користувачів у різних демографічних групах та регіонах [26]. Станом на квітень 2023 р. кількість активних ідентифікаторів соціальних медіа у світі досягла 4,8 млрд [27].

Соціальні медіа можуть бути цінним ресурсом для статистики у частині розуміння поведінки населення (завдяки величезній кількості даних, створених безпосередньо користувачами). Платформи соціальних медіа, забезпечуючи середовище для публічного висловлення своїх думок і настроїв, створюють величезний обсяг контенту – публікації, коментарі, оцінки, профілі, хештеги та ін. Ці дані можна збирати й аналізувати з метою отримання цінної статистичної інформації про демографічні показники, уподобання, тенденції та поведінку користувачів. Так, з допомогою аналізу мови та характеру висловлювань у публікаціях і коментарях у соціальних мережах можна оцінити настрої громадськості щодо певної теми, бренду чи події.

Варто зазначити, що обробка природної мови (Natural language processing – NLP) є одним із найперспективніших напрямів оброблення даних у соціальних мережах [28]. Відстежуючи хештеги, згадки та розмови на різних платформах, статистики можуть отримувати певні уявлення про популярні теми, інтереси та моделі поведінки користувачів соціальних мереж. А оскільки дані соціальних медіа супроводжуються метаданими, які містять інформацію про час, геолокацію та ін., то цілком реальним є отримання додаткового контексту для аналізу, що дозволятиме зрозуміти часові та просторові аспекти даних.

З метою представлення об'єктивної характеристики даних соціальних медіа варто розглянути сильні та слабкі сторони, а також можливості та загрози використання цих даних для потреб статистики (рис. 4, авторська розробка).

Сильні сторони	Слабкі сторони
1. Соціальні медіа – велике джерело даних для аналізу. 2. Генерування даних у режимі реального часу дозволяє відстежувати й аналізувати тенденції, події та суспільні настрої по мірі їх появи. 3. Дані соціальних мереж дають цінну інформацію про поведінку, уподобання, інтереси та взаємодію користувачів	1. Наявність у даних соціальних медіа неточностей, упередження, а також облікових записів спаму, що впливає на якість отриманих результатів. 2. Загроза конфіденційності та безпеці особистої інформації. 3. Ймовірність недостатньо репрезентативної вибірки, оскільки певні демографічні групи чи окремі особи можуть бути надмірно або недостатньо представлені на платформах соціальних мереж
Можливості	**Загрози**
1. Забезпечення у рамках статистичного аналізу даних соціальних медіа глибокого розуміння поведінки, уподобань і потреб споживачів, що дозволить покращити процес управління та прийняття рішень. 2. Залучення у статистику нових методів аналізу. 3. Відстежування подій, тенденцій, суспільних настроїв у режимі реального часу	1. Негативний вплив на доступність і можливості використання даних соціальних мереж, зумовлений певними обмеженнями, що спрямовані на забезпечення конфіденційності даних. 2. Уразливість даних соціальних мереж до поширення дезінформації та фейкових новин, що може призводити до недостовірних результатів аналізу. 3. Швидкість поширення негативних настроїв та реакцій громадськості

Рис. 4. SWOT-аналіз використання даних соціальних медіа для потреб статистики

Наступною складовою Концепції Smart-статистики є адміністративні дані, які генеруються органами загальнодержавної і місцевої влади в процесі виконання ними своїх владних повноважень, накопичуються та зберігаються в системі державних реєстрів. Ці дані, які, на відміну від даних з IoT і соціальних медіа, є структурованими, а не потоковими, вже зараз стали одним із головних джерел інформації для офіційної статистики, і питання їх використання як таких достатньо вивчено науковцями та практиками. Водночас адміністративні дані мають також слугувати важливим джерелом інформації для Smart-статистики. Ця проблематика потребує окремого детального вивчення і не розглядалася авторами у цій статті, оскільки адміністративні дані не є ключовими в екосистемі великих даних. Незважаючи на це, при подальшому збільшенні використання великих даних у статистиці важливість адміністративних даних зростатиме. На підтвердження цього можна навести приклад Великої Британії, де у 2013 р. створено спеціальну мережу дослідження адміністративних даних з метою забезпечення нових досліджень для суспільної користі [29].

Справжньою революцією в інформаційному середовищі стала поява штучного інтелекту і впровадження машинного навчання. На початку 2022 р. Європейська економічна комісія ООН (ЄЕК ООН, UNECE) оприлюднила публікацію "Машинне навчання для офіційної статистики" [30], яка базувалася на результатах реалізації двох міжнародних ініціатив: Проєкту машинного навчання (2019–2020 рр.) Групи високого рівня ЄЕК ООН з модернізації офіційної статистики (HLG-MOS) та Групи машинного навчання 2021 Офісу національної статистики Великої Британії (ONS) – UNECE і була затверджена HLG-MOS [31]. У цій публікації, з-поміж іншого, зазначалося, що машинне навчання має великий потенціал для статистичних організацій. Безпосередньо ШІ може зробити виробництво статистики ефективнішим, автоматизувавши певні процеси або допомагаючи фахівцям виконувати традиційні процеси. Це також дозволяє статистичним організаціям використовувати нові типи даних, такі як дані соціальних мереж і зображення. На сторінках документу представлено приклади практичного застосування машинного навчання у статистичних організаціях, подана структура якості як набір процедур і процесів забезпечення останньої, виклики, які виникають під час інтеграції такого навчання у статистичне виробництво, і ключові кроки для його переходу від експериментальної стадії до стадії виробництва та ін.

Технології ШІ можна, зокрема, використовувати в офіційній статистиці для підвищення точності й ефективності збирання, оброблення, аналізу та поширення даних. Також є можливість навчити алгоритми ШІ розпізнавати й отримувати інформацію з неструктурованих джерел даних, таких як текст, зображення та аудіо. У рамках процесу оброблення даних ШІ дозволяє автоматизувати такі завдання обробки даних, як їх очищення, перевірка

та стандартизація, що сприяє зменшенню кількості помилок і невідповідностей у даних і забезпечує їх високу якість. ШІ забезпечує швидкий та ефективний аналіз великих обсягів даних, визначаючи закономірності, тенденції та кореляції, які можуть бути пропущені у рамках традиційного аналізу, що сприятиме отриманню точнішої, надійнішої та глибшої статистики. Крім того, з допомогою ШІ можна розширити спектр джерел даних, які використовуються в офіційній статистиці, інтегруючи дані з різноманітних нових джерел, таких як соціальні мережі, супутникові зображення та дані датчиків. Це дозволить сформувати нове уявлення та погляди на соціальні й економічні тенденції та допоможе створити повнішу статистику для органів управління.

Малодослідженим для офіційної статистики, водночас цілком можливим для Smart-статистики є персоналізоване збирання та розповсюдження даних з допомогою технологій ШІ (рис. 5, авторська розробка) з метою адаптації цих даних і з огляду на конкретні потреби і вподобання окремих користувачів (або їх груп). Завдяки цьому підвищується доречність і зручність використання офіційної статистики та покращується допомога користувачам у частині прийняття більш обґрунтованих рішень на основі наявних даних.

Рис. 5. Складові персоналізованого збирання та розповсюдження даних з допомогою алгоритмів машинного навчання

Створення профілів користувачів відбувається з допомогою алгоритмів машинного навчання (налаштованих на кластеризацію або ж класифікацію), які збирають, аналізують та інтегрують демографічні дані (вік, стать, освіта, професія), дані про вподобання (відгуки, оцінки), поведінкові дані (історія веб-переглядів, активність у соціальних мережах). Такі алгоритми можуть допомогти виявляти шаблони у даних, а отже, опрацьовувати значний їх масив за невеликий проміжок часу.

Цінність складової "Механізми формулювання рекомендацій" полягає у тому, що ШІ може використовувати такі механізми, щоб пропонувати відповідні дані та ідеї на основі минулої поведінки та вподобань користувачів.

Складова "Налаштування інформаційних панелей" – це звична і невід'ємна частина сайта, адже зараз кожен сайт (урядовий, корпоративний чи персональний) спрямований на забезпечення зручності користувачам. Алгоритми ШІ можуть удосконалити функціонування інформаційних панелей шляхом оптимізації процесу налаштування та сортування даних відповідно до споживацьких вимог.

Адаптація інтерфейсів передбачає їх створення, налаштування та використання для представлення та візуалізації даних відповідно до вподобань користувачів, що своєю чергою допоможе останнім краще зрозуміти/ інтерпретувати дані й аналітику.

Складова "Обробка природної мови" здебільшого залучається для розуміння запитів користувачів та надання відповідних даних у розмовному форматі. Це може допомогти користувачам швидко знайти потрібну інформацію, не потребуючи навігації через складні інтерфейси даних. Методи обробки природної мови (Natural language processing, NLP) можуть ґрунтуватися на аналізі настроїв (Sentiment analysis), тобто розумінні емоцій з допомогою програмного забезпечення для визначення загального настрою відповідей [28].

Разом із перевагами, які технології ШІ можуть запропонувати офіційній статистиці, існують також потенційні загрози (або ж виклики), які необ-

хідно враховувати, щоб забезпечити точність, надійність та етичне використання цих технологій. Так, алгоритми ШІ можуть ухвалювати рішення лише на основі тих даних, на яких відбувається навчання. Тому якщо ці дані є упередженими, ШІ також прийматиме упереджені рішення або зберігатиме існуючі упередження в даних, що в подальшому може призвести до неточної чи дискримінаційної офіційної статистики. Ще одна група загроз пов'язана з безпекою та конфіденційністю, особливо під час роботи з конфіденційними даними. При цьому цілком реальним є ризики витоку даних, злому або несанкціонованого доступу до конфіденційної інформації.

Можливості офіційних статистичних організацій щодо ефективного використання технологій ШІ обмежуються потребами у досвіді їх упровадження та наявності спеціальних навичок. Вагомою перепоною є також вартість упровадження цих технологій, адже воно вимагає значних інвестицій в апаратне забезпечення, програмне забезпечення та персонал. Це, своєю чергою, може стати проблемою для офіційних статистичних організацій в умовах обмеженого фінансування зазначених робіт. Також алгоритми ШІ здебільшого складні для розуміння, що утруднює користувачам оцінку точності та надійності офіційної статистики, створеної за цими алгоритмами, а відтак, призводить до зниження довіри до статистики, викликає скептицизм щодо достовірності даних.

І наостанок, використання технологій ШІ викликає етичні міркування щодо таких питань, як право власності на дані, прозорість і підзвітність. Ці міркування необхідно враховувати, аби переконатися, що використання технологій ШІ в офіційній статистиці є відповідальним і етичним.

Висновки. За результатами проведеного дослідження автори дійшли висновку, що наразі, зважаючи на необхідність забезпечення якості та конфіденційності, а також мінімізації вартості даних, необхідно сформувати логічні та адекватні підходи до розроблення методології збирання, опрацювання, групування та аналізу даних із джерел, які є ключовими для Smart-статистики. Незважаючи на усі переваги інструментів Smart-статистики та потенційно позитивні результати їх імплементації в офіційну статистику, з огляду на умови війни та відсутність нормативно-правового підґрунтя щодо ефективного використання цих інструментів Україна наразі не готова зіштовхнутися з тими загрозами, які вони у собі несуть.

Напрями подальших досліджень пов'язані з пошуком та поглибленим висвітленням оптимальних шляхів упровадження інструментів Smart-статистики в офіційну статистику України.

References

1. State Statistics Service as an IT company: Mykhailo Fedorov tells about the digital transformation of the main statistical agency. (2023). *Ministry of Digital Transformation of Ukraine*. Retrieved from https://www.kmu.gov.ua/en/news/derzhstat-iak-it-kompaniia-mykhailo-fedorov-rozpoviv-pro-tsyfrovu-transformatsiiu-holovnoho-statystychnoho-orhanu
2. Ricciato, F., Wirthmann, A., Giannakouris, K., Reis, F., & Skaliotis, M. (2019). Trusted smart statistics: Motivations and principles. *Statistical Journal of the IAOS, 35 (4),* 589–603. Retrieved from https://www.researchgate.net/publication/337549354_Trusted_smart_statistics_Motivations_and_principles
3. Cukier, K., & Mayer-Schoenberger, V. (2013). The Rise of Big Data. How It's Changing the Way We Think About the World. *Foreign Affairs, 92, 3*. Retrieved from https://www.dimt.it/wp-content/uploads/2017/08/www.foreignaffairs.com_system_files_pdf_articles_2013_92305.pdf
4. Ricciato, F., Wirthmann, A. & Hahn, M. (2020). Trusted Smart Statistics: How new data will change official statistics. *Data & Policy, 2, E7*. doi:10.1017/dap.2020.7
5. Giannakouris, K., Reis, F., Skaliotis, M., Wirthmann, A., & Ricciato, F. (2018). Trusted Smart Statistics: A reflection on the future of (Official) Statistics. *European Conference on Quality in Official Statistics. Session 27. (26–29 June, Krakow)*. Retrieved from https://www.q2018.pl/papers-presentations/?drawer=Sessions*Session%2027
6. Trusted Smart Statistics in a nutshell. (n.d.). European Commission. Collaboration in Research and Methodology for Official Statistics. *cros-legacy.ec.europa.eu*. Retrieved from https://cros-legacy.ec.europa.eu/content/trusted-smart-statistics-nutshell_en
7. Bucharest Memorandum on Official Statistics in a Datafied Society (Trusted Smart Statistics). (2018). 104th DGINS Conference 10–11 October, Bucharest). Eurostat. *ec.europa.eu*. Retrieved from https://ec.europa.eu/eurostat/web/european-statistical-system/-/dgins2018-bucharest-memorandum-adopted
8. Scheveningen Memorandum. Big Data and Official Statistics. (2014). *cros-legacy.ec.europa.eu*. Retrieved from https://cros-legacy.ec.europa.eu/system/files/SCHEVENINGEN_MEMORANDUM%20Final%20version.pdf
9. Radermacher, W. J. (2020). Official Statistics 4.0: The Era of Digitisation and Globalisation. *Official Statistics 4.0*. (pp. 119–156). Springer, Cham. Retrieved from https://link.springer.com/chapter/10.1007/978-3-030-31492-7_4

10. Taleb, I., Serhani, M., Bouhaddioui, C., & Dssouli, R. (2021). Big data quality framework: a holistic approach to continuous quality management. *Journal of Big Data, 8,* 76. Retrieved from https://doi.org/10.1186/s40537-021-00468-0

11. Chen, C. L. Philip, & Zhang, Ch.-Ya. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Science, 275,* 314–347. Retrieved from https://www.scinapse.io/papers/2109574129

12. Codd, E. F., Codd, S. B. & Salley, C. T. (1993). Providing OLAP to User-Analysts: An IT Mandate. E. F. Codd Associates. *web.archive.org*. Retrieved from https://web.archive.org/web/20170808214004/https://www.minet.uni-jena.de/dbis/lehre/ss2005/sem_dwh/lit/Cod93.pdf

13. Osaulenko, O., & Horobets, O. (2023). Using Big Data by Ukrainian official statistics when martial law applies: problems and solutions. *Statistics in Transition new series and Statistics of Ukraine. Joint Special Issue: A New Role for Statistics, 24, 1,* 27–41. Doi: 10.59170/stattrans-2023-003

14. InVivo Digital Factory. Digital Solutions. *www.invivo-group.com*. Retrieved March 6, 2023 from https://www.invivo-group.com/en/innovation-digital/digital-solutions

15. Big Data and Agriculture: A Complete Guide. *ua.talend.com*. Retrieved March 6, 2023 from https://ua.talend.com/resources/big-data-agriculture/

16. Chong-Chong, Qi. (2020). Big data management in the mining industry, *International Journal of Minerals, Metallurgy and Materials, 27, 2,* 131–139. Retrieved from http://ijmmm.ustb.edu.cn/en/article/doi/10.1007/s12613-019-1937-z

17. Wang, L., & Alexander, Ch. A. (2015). Big Data in Design and Manufacturing Engineering. *American Journal of Engineering and Applied Sciences, 8 (2)*. Retrieved from https://thescipub.com/pdf/ajeassp.2015.223.232.pdf

18. Big Data Market in the Automotive Industry – Growth, Trends, COVID-19 Impact, and Forecasts (2022–2027). (2022). *Globe Newswire*. Retrieved from https://www.globenewswire.com/news-release/2022/07/25/2485150/0/en/Big-Data-Market-in-the-Automotive-Industry-Growth-Trends-COVID-19-Impact-and-Forecasts-2022-2027.html

19. Drigas, A., & Leliopoulos, P. (2014). The Use of Big Data in Education. *International Journal of Computer Science Issues, 11, 5,* 1. Retrieved from https://www.researchgate.net/publication/274890131_The_Use_of_Big_Data_in_Education

20. Williamson, B. (2017). *Big Data in Education. The digital future of learning, policy and practice*. London: SAGE Publication Ltd. Retrieved from https://uk.sagepub.com/en-gb/eur/big-data-in-education/book249086#preview

21. Blanco, J. M. (2022). Big Data and Energy: A Combination for Success. *www.plainconcepts.com*. Retrieved from https://www.plainconcepts.com/big-data-energy/

22. Cisco Annual Internet Report (2018–2023). White Paper. (2020). *www.cisco.com*. Retrieved from https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

23. Yang, Li, & Shami, A. (2022). IoT data analytics in dynamic environments: From an automated machine learning perspective. *Engineering Applications of Artificial Intelligence, 116*. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0952197622003803

24. Bhardwaj, A., Kaushik, K., Bharany, S. Rehman, A. Ur, Hu, Yu-Ch., Eldin, E. T. et all. (2022). IIoT: Traffic Data Flow Analysis and Modeling Experiment for Smart IoT Devices. *Sustainability, 14 (21),* 14645. Retrieved from https://www.mdpi.com/2071-1050/14/21/14645

25. European Commission. (2020). Communication from the Commission to the European Parliament, the Council, the European economic and social committee and the committee of the Regions. Chemicals Strategy for Sustainability Towards a Toxic-Free Environment. SWD(2020) 249 final. *www.prevencionintegral.com*. Retrieved from https://www.prevencionintegral.com/sites/default/files/noticia/512669/field_adjuntos/swdpfas.pdf

26. Global Social Media Statistics. *datareportal.com*. Retrieved March 6, 2023 from https://datareportal.com/social-media-users

27. Digital 2023 April Global Statshot Report. *datareportal.com*. Retrieved April 16, 2023 from https://datareportal.com/reports/digital-2023-april-global-statshot

28. Farzindar, A., & Inkpen, D. (2015). *Natural Language Processing for Social Media*. Springer Cham. Retrieved from https://link.springer.com/book/10.1007/978-3-031-02157-2

29. UK Statistics Authority. Administrative Data Research Network (Archived). *uksa.statisticsauthority.gov.uk*. Retrieved March 6, 2023 from https://uksa.statisticsauthority.gov.uk/better-useofdata-statistics-and-research/adrn/

30. *UNECE*. (2021). Machine Learning for Official Statistics. Retrieved from https://unece.org/sites/default/files/2022-09/ECECESSTAT20216.pdf

31. *UNECE*. High-level Group for the Modernisation of Statistical Production and Services. unece.org. Retrieved March 6, 2023 from https://unece.org/statistics/networks-of-experts/high-level-group-modernisation-statistical-production-and-services

O. H. Osaulenko,
DSc in Public Administration, Professor,
Correspondent Member of the NAS of Ukraine,
Rector,
E-mail: O.Osaulenko@nasoa.edu.ua
ORCID: https://orcid.org/0000-0002-7100-7176;
O. O. Horobets,
PhD in Economics,
Associate Professor of Department for Statistics, Information Technology and Econometric Methods,
E-mail: babutska@ukr.net
ORCID: https://orcid.org/0000-0001-5433-6448;
National Academy of Statistics, Accounting and Audit

Implementing Smart Statistics Toolkit in the Official Statistics

Important issues of Smart statistics are addressed. The Smart statistics toolkit is analyzed: big data, data of artificial intellect and Internet of things, social media, and administrative data.

The study involves conceptualization of the Smart statistics and identification of advantages and threats for the official statistics in using Smart statistics data. A set of principles for operating big data are proposed, with categorizing the sources of big data generation (in conformity with the economic activities), presently demanded by the official statistics and the society (agriculture, health protection, mining industry, mechanical engineering, education, power industry). A smart data system in smart cities is developed, containing the following components: smart house, smart environment, smart control, smart traffic, smart health, and smart citizen. It is determined that the system's objective is to create the smart environment and simplify the way of life through saving time, energy and money. The components of personalized collection and dissemination of data are determined.

The artificial intelligence (AI) is considered as a component of the Smart statistics conception. In this article's context, the authors observe that using AI causes ethical discourses on issues such as property right for data, transparence and accountability of data. These discourses need to accounted for, in order to assure that AI technologies are used by the official statistics in a responsible and ethical manner.

Based on the results of the study, it is concluded that now it is necessary to create logical and adequate approaches to elaborating a methodology for collection, processing, grouping and analysis of statistical data from alternative information sources. It is argued that in spite of the advantages of the smart statistics toolkit and potentially positive results of its implementation in the official statistics, present-day Ukraine is not ready to face the threats involved in it given the war conditions and lack of a regulatory framework for its use.

Further studies on this theme have to focus on in-depth analysis of Smart statistics and search for optimal ways for implementing its toolkit in the official statistics of Ukraine.

Key words: *Smart statistics, official statistics, big data, artificial intelligence, Internet of things, social media, administrative data.*

Бібліографічний опис для цитування:

Осауленко О. Г., Горобець О. О. Імплементація інструментарію Smart-статистики в офіційну статистику. *Статистика України*. 2023. № 1. С. 7–18. Doi: 10.31767/su.1(100)2023.01.01

Bibliographic description for quoting:

Osaulenko, O. H., & Horobets, O. O. (2023). Implementatsiia instrumentariiu Smart-statystyky v ofitsiinu statystyku [Implementing Smart Statistics Toolkit in the Official Statistics]. *Statystyka Ukrainy – Statistics of Ukraine, 1,* 7–18. Doi: 10.31767/su.1(100)2023.01.01