3. Програма розвитку державної статистики до 2023 року: Постанова Кабінету Міністрів України від 27.02.2019 р. № 222. URL: https://zakon.rada.gov.ua/laws/show/222-2019-%D0%BF

4. Технологии Big Data: как использовать большие данные в маркетинге. URL: https://www.uplab.ru/blog/big-data-technologies/

Про державну статистику: Закон України від 17.09.1992 р. № 2614-XII, станом на 19.04.2014 р. URL: https://zakon.rada.gov.ua/laws/show/2614-12

ВИКОРИСТАННЯ ВЕЛИКИХ ДАНИХ У ОФІЦІЙНІЙ СТАТИСТИЦІ

Осауленко Олександр Григорович,
доктор наук з державного управління,
професор, член-кореспондент НАН України,
ректор,
Національна академія статистики, обліку та аудиту

Питання перспектив переходу до використання великих даних у статистичних цілях активно дискутуються у розвинених країнах і на міжнародному рівні. Вирішальною у цьому відношенні стала Конференція європейських статистиків 2013 року, на якій Європейська комісія ООН представила свою доповідь з указаної проблематики [1]. У 2014 році в Римі на конференції Європейської статистичної системи з проблематики великих даних питання її статистичного використання було визначено як ключове для офіційної статистики [2].

Як відомо, поняття великих даних виникло у зв'язку з невпинно зростаючою кількістю електронних джерел, що продукують інформацію у вебресурсах, а також частотою виникнення такої інформації. Великі дані характеризуються як дані зростаючого обсягу та швидкості продукування, а також їх різноманітності. Це так звана тривекторність великих даних. Іншою важливою рисою великих даних є те, що вони зазвичай не є структурованими, тобто формуються без заздалегідь визначеної концепції чи моделі, через що не можуть бути інтегровані безпосередньо (тобто без попередньої обробки) до традиційних баз даних [3].

Відтак, відповідно до ролі й функцій офіційної статистики великі дані можуть бути визначені на основі відповідного глосарію, розробленого компанією Gartner: великі дані є джерелом інформації, яка характеризується великими обсягами, постійно зростаючими швидкістю утворення та розмаїттям, що потребує, своєю чергою, ефективних з погляду витрат і можливості автоматизації інноваційних форм їх обробки для потреб забезпечення процесу прийняття рішень [4].

Існують очевидні перспективи застосування великих даних для цілей офіційної статистики або у так би мовити чистому вигляді, або в комбінації з даними традиційних статистичних спостережень та адміністративними

даними. Разом з тим отримання інформації з великих даних і подальше її інтегрування до статистичного виробничого процесу не є простим завданням, його реалізація спирається передусім на вирішення таких фундаментальних питань:

1) на які саме великі дані має орієнтуватись офіційна статистика, беручи до уваги її завдання й функції у суспільстві?

2) у який спосіб офіційна статистика може використовувати великі дані та що вона потребує для цього?

3) яким чином великі дані можуть допомогти здійснювати більш точне й своєчасне (порівняно з сучасним) статистичне оцінювання соціальних, економічних та екологічних явищ?

Донедавна офіційна статистика базувалася виключно на традиційних формах статистичних спостережень або адміністративних джерелах інформації. При цьому у багатьох країнах право на отримання такого роду інформації регламентовано законодавством у сфері статистики та інформації. Великі дані потенційно є джерелом більш релевантної і вчасної статистичної інформації порівняно з традиційними її джерелами. Їх природа не вкладається у традиційну схему отримання статистичних даних, оскільки вони або безпосередньо доступні широкому загалу в мережі Інтернет, або мають власників (і, відповідно, авторське право) у вигляді приватних структур, а частіше за все наявні обидві ці обставини. У результаті, приватні структури можуть отримувати певні переваги у використанні великих даних, виробляючи все більшу кількість саме статистичної інформації, і здатні, своєю чергою, конкурувати з офіційною статистикою в плані її оперативності й відповідності потребам користувачів.

Разом з тим офіційна статистика має таку очевидну перевагу перед приватним сектором, як наявність інфраструктури та тривалого досвіду роботи над різними аспектами якості статистичної інформації, включно з дотриманням вимог до захисту конфіденційної інформації. Отже, за умови своєчасного «вбудовування» великих даних у статистичний виробничий процес офіційна статистика отримує очевидні переваги з погляду якості кінцевої інформації.

Загальновідомі джерела великих даних можуть бути класифіковані за такими групами:

1) адміністративні дані (електронні медичні картки, дані страхування, банківська інформація та ін.);

2) транзакції або бізнес-інформація (транзакції за кредитними та депозитними картками, онлайн транзакції тощо);

3) дані сенсорних уловлювачів (дані сателітних зйомок, дорожніх радарів, кліматичних пристроїв та ін.);

4) дані мобільних сенсорних пристроїв (інформація з мобільних телефонів, GPS та ін.);

5) поведінкові дані (інформація Інтернет-пошукачів тощо);

6) інформація щодо індивідуальної та суспільної думки (коментарі стосовно різних подій у соціальних мережах, реакція на інформацію медіа-ресурсів тощо).

Головним потенційним джерелом даних для офіційної статистики є великі дані адміністративного походження, тобто першої вищезазначеної групи. Такі дані збираються зазвичай на регулярній основі і можуть бути структуровані відповідно до традиційних статистичних баз даних. Щодо можливості використання в офіційній статистиці інших видів великих даних, то це питання, хоча й активно досліджується, залишається на сьогодні поки що відкритим.

Найбільш проблемними моментами з погляду потенційного використання великих даних у статистичних цілях є складність керування їх постійно зростаючими обсягами й періодичністю, відсутність бази вибірки, слабка або відсутня структурованість, відсутність контролю якості та незахищеність індивідуальних даних.

Загалом у світлі перспектив активного використання великих даних перед офіційною статистикою постають численні глобальні виклики. Головні з них такі:

1) юридичні (унормування доступу та використання даних);

2) захист персональних даних громадян та індивідуальних даних підприємств;

3) фінансові (оптимізація співвідношення витрат і виграшів);

4) управлінські (розробка політики та принципів управління й захисту даних);

5) методологічні (забезпечення якості даних та надійності статистичних методів);

6) технологічні (використання відповідних інформаційних технологій) [1].

Юридичні питання пов’язані з правом офіційної статистики на отримання різного роду інформації, у тому числі й адміністративної. Таке право зазвичай визначено законодавством кожної окремої країни у сфері статистики. Але у будь-якому випадку, навіть за умови наявності зазначеного права, офіційна статистика має довести необхідність та продемонструвати доцільність отримання доступу до певної адміністративної інформації у вигляді великих даних.

Захист персональних даних та індивідуальних даних підприємств базується на праві кожного громадянина контролювати та / або впливати на те, яка саме інформація щодо нього може бути розкрита й оприлюднена, та на відповідному праві підприємств (компаній) узгоджувати розкриття індивідуальної інформації про їх діяльність або клієнтів. Останнє стосується випадків, коли підприємства мають на меті захистити свою конкурентоздатність або споживачів власної продукції. Проблемою великих даних є те, що користувачі послуг та пристроїв, які генерують такі дані, у переважній своїй більшості не усвідомлюють цього факту, а також не володіють інформацією щодо того, для яких цілей їх персональні дані

можуть бути використані. Крім того, існує серйозна проблема забезпечення належного рівня технічної захищеності окремих Інтернет-сайтів.

Фінансовий аспект може стати суттєвим гальмом отримання великих даних для потреб офіційної статистики, передусім коли йдеться про дані, утримувачем яких є приватний сектор. Якщо умови отримання таких великих даних для статистичних цілей не мають відповідного юридичного підґрунтя, це може значно підвищити вартість їх використання. Статистичні служби мають знайти прийнятний консенсус між витратністю і бажаною якістю, передусім своєчасністю отримання готового статистичного продукту та зменшенням звітного навантаження на респондентів за рахунок залучення великих даних. Тобто навіть за умови високої фінансової витратності потенційні значні вигоди використання великих даних для суспільства можуть переважити вартісний фактор. Наприклад, позитивний ефект від своєчасно наданої на базі великих даних статистичної інформації з критичних питань охорони здоров'я може стати вирішальним у прийнятті рішення щодо доцільності їх використання [5].

Виклики у питаннях управління великими даними в офіційній статистиці (їх менеджменту) пов'язані, по-перше, зі значним зростанням фізичних обсягів первинних даних, які надходитимуть до статистичної системи, по-друге – зі зміною природи первинних даних і по-третє – із людським фактором. Це потребує розробки принципово нової політики з управління інформацією, її захисту, а також процесом академічної і професійної підготовки статистичного персоналу та стосунками із респондентами.

З методологічного погляду фундаментальною проблемою використання великих даних в офіційній статистиці є забезпечення їх репрезентативності. За відсутності бази вибірки для такого типу даних складнощі виникають у визначенні як цільової сукупності загалом, так і вибіркової сукупності. Традиційні статистичні спостереження базуються на переписах та реєстрах, на основі яких можна визначити усі типи сукупностей, необхідні для реалізації статистичного спостереження. Великі же дані формуються спонтанно й поза межами реєстрів, що значно ускладнює їх прив'язку до такої центральної концепції офіційної статистики, як статистична одиниця.

Інша методологічна проблема полягає в тому, що існуючі статистичні методи розраховані на послідовний, глибокий та тривалий аналіз даних невеликих за розміром вибірок і це значно гальмує статистичний виробничий процес. У зв'язку з цим виникає потреба у таких інноваційних методах, які б: а) дозволяли швидко аналізувати дуже великі масиви даних, наприклад методи візуалізації інформації, обробки текстової інформації та представлення даних як безперервного, але водночас регулярного потоку інформації, що вбачається можливим з урахуванням постійно зростаючої потужності комп'ютерної техніки; б) були придатні для аналізу й підключення інформації, не охопленої безпосередньо статистичним процесом, наприклад методи встановлення великомасштабних інформаційних зв'язків та спеціальні статистичні методи для роботи з великими масивами інформації.

Головною вимогою до таких методів має бути їх придатність для швидкого ув'язування між собою, аналізу й обробки великих обсягів інформації.

Ще однією методологічною проблемою використання великих даних у статистичних цілях є забезпечення відповідності технічного арсеналу статистичних досліджень у частині:

1) вимірювання якості даних, що виникають поза статистичною системою, оскільки залежність від зовнішніх джерел інформації значно обмежує можливості застосування статистичних технік порівняно з даними цільових обстежень;

2) загальної обмеженості сфери застосування даних із зовнішніх джерел;

3) складності інтегрування інформації різного зовнішнього походження у статистичні бази даних для отримання якісного кінцевого продукту;

4) складності формування цінової пропозиції за умови відсутності близьких за типом продуктів на інформаційному ринку.

Технологічна проблематика переходу до широкого використання великих даних в офіційній статистиці пов'язана, передусім, з питанням їх високих швидкісних характеристик. Значне підвищення швидкості надходження даних, доступу до них і їх обробки викликає необхідність інтенсивного використання спеціальних стандартних програмних продуктів, наприклад, як Application Programme Interfaces (API) або в окремих випадках потокового його варіанту – Streaming API. Завдяки цим продуктам, які вже досить активно використовуються статистичними службами розвинених країн, стає можливим отримання безпосереднього, у реальному часі, доступу до великих адміністративних даних. Це, своєю чергою, дозволяє інтегрувати суто адміністративні дані з великими даними з інших джерел, таких як, наприклад, комерційні дані, дані мобільних телефонів, соціальні мережі, медіаресурси тощо.

Інтегрування й комбінація різного роду статистичної і не статистичної за своєю природою інформації вимагає від офіційної статистики активного розвитку методів статистичного моделювання. Разом з тим, якщо обробка й аналіз великих, у тому числі індивідуальних даних у найбільш статистично розвинених країнах поступово стають ефективним засобом отримання статистичної інформації, то питання їх збереження, захисту й архівації усе ще залишаються слабким місцем технологічного процесу.

Важливим питанням є вибір тематики великих даних, яка може бути корисною для офіційної статистики. Різноманітність цієї тематики і її постійна самоактуалізація є одночасно перевагою і недоліком великих даних. Перевагою – оскільки у статистиків є широкий набір актуальної для них інформаційної тематики, а недоліком – тому що таке розмаїття ускладнює завдання її вибору й аналізу. Відповідно до досвіду провідних статистичних служб світу і проектів Євростату, основними перспективними статистичними галузями застосування великих даних є статистики транспорту, цін (побудова індексу споживчих цін), туризму, використання інформаційних та комунікаційних технологій, а також соціальна медіастатистика.

Перехідний період з упровадження у статистичну практику великих даних передбачає комбінування різного роду таких даних, поступове заміщення окремих традиційних статистичних спостережень аналогічними за тематикою великими даними, а також пошук і використання принципово нових таких даних з метою задоволення зростаючих потреб користувачів.

Список використаних джерел

1. UNECE (2013). What does Big Data mean for official statistics? Conference of European Statisticians. 10 March. Retrieved from https://statswiki.unece.org/pages/viewpage.action?pageId=77170614

2. CROS. (2014). Collaboration in Research and Methodology for Official Statistics. ESP Rome 2014. Retrieved from https://ec.europa.eu/eurostat/cros/content/esp-rome-2014_en

3. Васєчко О. О. Сучасні виклики статистичних вищої освіти і науки // Статистика України. 2014. № 4. С. 4–16.

4. Big Data. Gartner Glossary. Information Technology. Retrieved from https:// www.gartner.com/it-glossary/big-data/

5. Осауленко О. Г. Офіційна статистика в системі національної інформаційної безпеки: монографія. Київ: ТОВ «Август Трейд», 2017. 367 с.

ОСОБЛИВОСТІ МОНІТОРИНГУ СОЦІАЛЬНИХ ЯВИЩ В УМОВАХ ДИДЖИТАЛІЗАЦІЇ: МЕТОДИЧНІ ТА ПРАКТИЧНІ АСПЕКТИ

Пальян Зінаїда Оганесівна,
кандидат економічних наук, доцент,
доцент кафедри статистики та демографії;
Григор'єва Катерина Олександрівна,
студентка спеціальності
«Економічна аналітика та статистика»;
Київський національний університет імені Тараса Шевченка

В епоху цифрових технологій надзвичайно важливим є питання пошуку сучасних способів обробки значних за обсягом наборів даних, швидкість нагромадження та ускладнення яких з плином часу продовжує зростати. Створення нових методів аналізу масової інформації дає можливість досягти більшої гнучкості й адаптивності у сфері опрацювання зібраних даних. Водночас виникає проблема інтерпретації отриманих результатів, а також перспективи використання новостворених способів статистичної обробки, але вже на інших масивах даних. Не менш актуальним є питання збалансованості між економією витрат ресурсів та якістю і обсягом зібраних даних. Тобто важливо за мінімальних витрат часу, матеріальних і людських ресурсів забезпечити належну точність і структурованість статистичної інформації без втрати необхідного її обсягу.