УДК 303.436:311.2/.3

В. Г. Саріогло,
доктор економічних наук,
завідувач відділу,
Інститут демографії та соціальних досліджень
імені М. В. Птухи НАН України,
E-mail: sarioglo@idss.org.ua

"Великі дані" як джерело інформації та інструментарій для офіційної статистики: потенціал, проблеми, перспективи

Розглянуто питання, пов'язані з потенційною можливістю використання в офіційній (державній) статистиці так званих "великих даних". Висвітлено їх переваги, серед яких своєчасність, широке охоплення певних частин цільових сукупностей, скорочення витрат на їх отримання. Окреслено проблеми, які необхідно вирішувати при використанні "великих даних". Наведено аргументи щодо наявності у прикладній та в офіційній статистиці прототипів інструментів, які за належного їх розвитку та адаптації дадуть можливість розв'язати основні з указаних проблем.

Ключові слова: *"великі дані", джерела інформації, статистичний інструментарій, офіційна статистика, інформаційні технології.*

Проблемам використання так званих "великих даних" як джерела інформації при вимірюванні важливих соціально-економічних показників в офіційній статистиці протягом останніх декількох років приділяється значна увага. На підтвердження достатньо зазначити, що у 2015 та 2016 роках будь-яка відносно велика конференція зі статистики, що проводилась у США або у країнах ЄС, майже завжди включала доповіді на цю тему. А регулярна конференція Європейської Комісії "Нові методики та технології у статистиці 2015 року" (New Techniques and Technologies for Statistics – NTTS 2015) фактично була присвячена зазначеним питанням. У 2017 році NTTS 2017 проходитиме паралельно, фактично на конкуруючих засадах, із заходом European Big Data Hackathon, на якому групи розробників із європейських країн представлятимуть свої інструменти для об'єднання офіційної статистики та "великих даних" з метою покращання інформаційного забезпечення політики, управління та досліджень.

Таке, для багатьох несподіване, зростання інтересу до питань імплементації "великих даних" у статистику пов'язане, насамперед, зі значним комерційним успіхом цього підходу у США. Це проілюстровано у багатьох публікаціях і навіть у науково-популярних виданнях [1; 2]. Однак щодо дійсної корисності "великих даних" для офіційної статистики та оптимальних шляхів їх упровадження ще не все ясно і тривають наукові дискусії. Це пояснюється тим, що ідея використання "великих даних", зокрема й у статистиці, виникла лише приблизно десять років тому у сфері інформаційних технологій. Протягом останнього десятиліття з'явився розвинутий інструментарій для її реалізації у різних сферах суспільного життя. Для сучасної офіційної статистики такий темп інноваційного розвитку є занадто високим, а помилки на цьому шляху можуть коштувати достатньо дорого.

В Україні "великим даним" також почали приділяти певну увагу – у ВНЗ на кафедрах прикладної математики, статистики, програмування плануються (а може вже десь і викладаються) відповідні курси. Але фахівцями державної і прикладної статистики ця тема лише почала обговорюватися.

Метою цієї статті є висвітлення сутності підходу на основі "великих даних", потенційних напрямів та проблем його застосування в офіційній статистиці. Статтю підготовлено за матеріалами конференції NTTS 2015, іншими публікаціями та результатами дискусій з цих питань.

Насамперед зазначимо, що наразі для поняття "великі дані", як і слід було очікувати, не існує загальноприйнятого трактування. Власне, і сам термін ще може змінитися. Частіше за все під "великими" розуміють дані, які через значні обсяги не можуть бути оброблені стандартними інструментами, а потребують спеціальних програмних і технічних засобів. Дещо більш зрозумілим це визначення стане, якщо вказати джерела "великих даних": інформація з соціальних мереж, камер відеоспостереження, камер відстеження дорожнього руху та фіксації порушень його правил, відеореєстраторів, мобільних пристроїв, зокрема телефонів, інформація з касових апаратів щодо покупок у великих торгових мережах, оцифровані програми телебачення, звукові записи тощо. З наведеного переліку зрозуміло, що найбільшими обсягами "великих даних" володіють великі Інтернет провайдери, провайдери соціальних мереж,

достатньо великі мобільні оператори, телевізійні та кінокомпанії.

Але варто взяти до уваги, що великі Інтернет-компанії та мобільні оператори існували і 15 років тому, особливо у розвинених країнах, а "великі дані" з'явилися помітно пізніше. Це пояснюється тим, що наявні в таких компаніях дані й інформація практично не використовувалися для аналізу соціальних та економічних явищ і процесів. Дані отримувалися (або збиралися), передавалися і зберігалися. Не існувало ні систематичного попиту на таку інформацію, ні спеціалізованого інструментарію для її обробки та аналізу. Отже, слід погодитися з тими, хто під "великими даними" розуміє не просто великі та постійно зростаючі обсяги даних, а також методологію, операції, алгоритми, що дають можливість оперувати даними дуже великих масштабів та отримувати з них необхідну інформацію. Таку, наприклад, як уподобання покупців, основні характеристики заміщення товарів і послуг через зміну попиту та пропозиції, динаміка цін за групами товарів, динаміка поширення епідемій та багато чого іншого.

На наш погляд, для більшої ясності тут доцільно відмітити, що у загальноприйнятому сенсі під даними розуміють достатньо прості об'єкти – окремі числа або їх набори (масиви), за якими людині неможливо або дуже важко зрозуміти що вони характеризують. У найбільш поширеному і найпростішому випадку – це числа "0" та "1", які характеризують наявність або відсутність певної ознаки, об'єкта, явища і т. ін. Під інформацією частіше за все мають на увазі зрозуміле людині повідомлення, отримане на основі даних. Ці визначення належать до сфери інформаційних технологій, де вся інформація є оцифрованою і представляється у вигляді послідовності бітів (байтів і т. д.).

Як важливий штрих до визначення "великих даних" слід додати, що особливістю цього підходу є не тільки і не стільки великі масштаби даних, а намагання та можливість використати всі наявні дані щодо аналізованих явища або процесу. Так, для отримання інформації щодо договірних поєдинків у боротьбі сумо та їх учасників довелося проаналізувати результати всіх 64 тисяч поєдинків, що відбулися протягом 10 років у вищій лізі в Японії [1]. Такий масштаб оброблених та проаналізованих даних не є занадто великим, але ключовою ознакою тут є те, що були використані всі дані.

Великі масштаби даних диктують також зміну відношення до проблем їх точності. Як відомо, протягом останніх 20–30 років офіційна статистика розвинених країн приділяла і приділяє дуже значну увагу забезпеченню прийнятних рівнів точності отриманих даних [3]. Однією з причин цього є широке застосування вибіркових методів спостереження. Значні витрати ресурсів на проведення таких спостережень, високе навантаження на респондентів вимагають максимального скорочення обсягів вибірки і, відповідно, даних, що збираються, за умови забезпечення належного рівня якості результатів спостереження. Розвиваються методи оптимізації дизайну вибірки, підвищення ефективності оцінювання показників шляхом об'єднання даних різних обстежень, використання зовнішньої інформації тощо. До речи, деякі з цих методів також спочатку були запропоновані у сфері ІТ-технологій. А відтак, легко уявити, що при використанні даних, які характеризують малу частину цільової сукупності для отримання інформації щодо всієї сукупності, рівень точності даних відіграє визначальну роль у загальному рівні якості інформації, а при використанні всіх, або майже всіх даних щодо цільової сукупності рівень точності даних може бути значно нижчим за тих самих вимог щодо загального рівня якості отриманої інформації.

Зазначимо ще таку важливу особливість "великих даних". На їх основі для отримання інформації та формування суджень можуть використовуватися достатньо прості статистичні процедури. Це пояснюється тим, що необхідна інформація є щодо всіх одиниць сукупності. Аналогічна ситуація зі спрощенням статистичних процедур спостерігається, наприклад, при переході від аналізу агрегованих статистичних даних до аналізу даних макрорівня. При побудові аналітичних моделей за агрегованими даними залежно від цілей аналізу доводиться враховувати різні характеристики груп одиниць цільової сукупності та формувати доволі складні рівняння або системи рівнянь з використанням найбільш адекватних статистичних методів. При використанні даних мікрорівня для кожної одиниці спостереження можуть бути побудовані достатньо прості моделі зміни її характеристик або поведінки, які в сукупності відображають різноманітність характеристик різних груп населення.

До безумовних переваг "великих даних" слід віднести такі:

1) своєчасність – дані можуть отримуватись у режимі реального часу;

2) дуже широке охоплення – дані отримуються теоретично по всіх одиницях сукупностей, які здійснюють відповідні дії (покупки за кредитними картками, користування соціальними мережами, пошук товарів в Інтернеті та ін.) або володіють певними пристроями (мобільними телефонами, планшетами, комп'ютерами тощо);

3) для отримання таких даних не потрібно розробляти запитальники та проводити обстеження, навчати й оплачувати інтерв'юерів.

Обсяги даних безперервно зростають – лише користуйся. Насправді, здається що в офіційної статистики немає альтернативи все більш широ-

кому використанню таких даних. І все ж на нинішньому етапі розвитку статистики можливість використання "великих даних" викликає багато запитань. Так, професор Д. Пфефферманн (Danny Pfeffermann), чинний президент Міжнародної асоціації статистиків з обстежень, зазначає, що проблема збирання та використання "великих даних" для офіційної статистики є однією з найбільш складних із тих, які доведеться вирішувати у найближче десятиріччя [4]. Ця проблема суттєво ускладнюється необхідністю інтеграції комп'ютерних наук у статистику для забезпечення можливості роботи з "великими даними". Основними питаннями, що доведеться вирішувати, на його думку є: охоплення цільових сукупностей та зміщення оцінок показників; доступність "великих даних", зокрема зміна законодавства з метою забезпечення статистичним службам доступу до них (з огляду на те, що ці дані у більшості випадків збираються приватними компаніями та належать їм); захист даних, отриманих від приватних компаній; збереження великих обсягів даних, їх обробка та аналіз; статистичне об'єднання значної кількості різних масивів даних, оскільки масиви приватних компаній є вузько спеціалізованими і, як правило, не містять одночасно всіх необхідних для офіційної статистики характеристик одиниць цільової сукупності; запобігання ризикам, пов'язаним з можливістю маніпуляцій з даними. Ці питання Д. Пфефферманн виразив лаконічною формулою:

"Великі дані" ⟶ Великі проблеми ⟶ ⟶Великий головний біль.

Для розв'язання зазначених питань у статистиці вже існують певні інструменти, хоча вони наразі й не дуже розвинуті та релевантні проблемам. Це лише підтверджує відоме положення, що в науці дуже часто питання виникають майже одночасно з можливостями для їх розв'язання. Так, для запобігання зміщенню охоплення цільової сукупності та оцінювання показників "великі дані" можуть використовуватися разом з даними стандартних статистичних обстежень. Такий підхід, зокрема, уможливлює суттєве поглиблення аналізу ринку праці на основі даних вибіркових обстежень робочої сили та заробітної плати й даних рекрутингових агенцій щодо вакансій і пропонованих рівнів оплати праці, оцінку індексів споживчих цін на основі стандартних обстежень цін і даних щодо продажів у супермаркетах та інтернет-магазинах тощо. Залишається "лише" розробити відповідні статистичні моделі та алгоритми.

Проблема збереження та обробки "великих даних" органами офіційної статистики може потенційно вирішуватися на основі "хмарних" технологій і така можливість опрацьовується статистиками у багатьох країнах [5]. Якщо "хмари" для цілей офіційної статистики будуть створені організаціями з солідною репутацією та можливостями щодо захисту інформації, зокрема Євростатом, то проблема збереження конфіденційних даних безумовно може бути суттєво пом'якшена.

Представляється доцільним приділити увагу ілюстрації можливих підходів до використання "великих даних" при оцінюванні статистичних показників. Насамперед слід зазначити, що, на нашу думку, роль вибіркового методу спостереження за цих умов не лише не знизиться, а навіть зросте. Навіть якщо в розпорядженні дослідника є масив інформації, який характеризує всю сукупність одиниць спостереження, часто зручніше працювати з певною вибіркою даних із такого масиву або вважати, що дані для всієї сукупності одиниць є вибіркою з певної "суперсукупності" з певним розподілом досліджуваних характеристик [6]. Такий підхід дає можливість використовувати при оцінці показників і аналізі явищ дуже потужний статистичний інструментарій, розроблений для вимірювання імовірнісних явищ і процесів.

За наявності масиву великих даних Інтернет-опитувань, серйозною проблемою є статистична оцінка сумарних показників за певною цільовою сукупністю, оскільки респондентами є користувачі Інтернету, які добровільно вирішили взяти участь в обстеженні, нехай і за певну нагороду (у формі, наприклад, можливості отримання специфічної інформації). Припустимо, що дослідників цікавить питання, скільки коштів у певній країні сумарно витрачають на кока-колу молоді люди віком 17–25 років. Нехай для збирання такої інформації з усіх можливих респондентів – сукупності осіб підходящого віку, зареєстрованих на певних сайтах – побудовано імовірнісну вибірку R_{Int}. При цьому ймовірності включення одиниць до вибірки і, відповідно, їх статистичні ваги w_i^{Int}, відображають лише пропорції певних груп одиниць спостереження у вибірці й у Інтернеті, не маючи водночас явного і зрозумілого відношення до структури генеральної сукупності цієї країни – всіх молодих людей відповідного віку. Якщо по вибірці R_{Int} отримано інформацію щодо витрат на кока-колу і щодо певних характеристик респондентів (стать, вік, місце проживання, соціальний статус тощо), то є можливість наблизити оцінки відповідних показників по ній до оцінок за генеральною сукупністю. Для цього лише потрібно використати дані обстеження з імовірнісною вибіркою (що може розглядатись як еталонна або контрольна вибірка) R_C, в якій обстежено достатньо велику кількість молодих людей із цільової сукупності та отримано інформацію, що так або інакше пов'язана зі схильністю респондентів до споживання кока-коли. Такою вибіркою може бути вибірка, побудована для обстеження особливостей споживання населенням продуктів харчування, особливостей витрачання

молоддю грошей тощо. При цьому результати обстеження можуть бути і не дуже актуальними. Результати обстеження за вибіркою R_{Int} (актуальні та по значній кількості одиниць цільової сукупності) можуть бути скориговані для оцінки суми коштів, витрачених на кока-колу, по генеральній сукупності $\hat{D}$ на основі статистичних ваг w_i^{Int}:

$$\hat{D} = \sum_{s=1}^{S} \left(\frac{\sum_{k=1}^{K_s} w_k^C}{\sum_{i=1}^{I_s} w_i^{Int}} \cdot \sum_{i=1}^{I_s} w_i^{Int} \cdot d_i^{Int} \right), \qquad (1)$$

де w_k^C – статистичні ваги одиниць спостереження в обстеженні з вибіркою R_C; d_i^{Int} – витрати на кока-колу респондентів в обстеженні R_{Int}; S – кількість страт або кластерів, за якими здійснюється узгодження вибірки R_{Int} з вибіркою R_C; K_s – кількість одиниць, обстежених за вибіркою R_C у страті s; I_s – кількість одиниць, обстежених за вибіркою R_{Int} у страті s.

З формули (1) випливає, що оцінки за вибіркою R_{Int} фактично масштабуються за наявними пропорціями по генеральній сукупності. Для такого масштабування можуть бути використані різні процедури, зокрема і калібрація статистичних ваг w_i^{Int} за наявності оцінок сумарних значень певних показників за вибірками R_{Int} та R_C. Перспективним напрямом покращання оцінок є також моделювання механізмів формування цільових сукупностей в Інтернеті та механізмів участі респондентів в опитуваннях. На жаль серйозною проблемою залишається можливість адекватної оцінки характеристик надійності показників, визначених за результатами збирання інформації в Інтернеті.

Слід зазначити, що окремі характеристики соціально-економічних процесів можна вже сьогодні отримати з Інтернету, причому з використанням стандартних статистичних процедур. Наприклад, з використанням інструменту Google Trends за запитом "Пошук роботи" можна отримати відносну кількість відповідних запитів в Україні або в окремих регіонах (щодо максимальної кількості запитів за досліджуваний період). При цьому відносна кількість запитів надається кумулятивно за 2–3 дні зі своєчасністю, що складає також 2–3 дні (тобто сьогодні доступні дані за позавчора). Якщо порівняти ці дані з даними щодо відносної кількості безробітних за методологією МОП, які розраховуються Держстатом, то видно, що принаймні сезонні коливання цих двох динамічних рядів є цілком узгодженими (рис. 1, для побудови графіку дані з Інтернету [7] агреговані за кварталами). При цьому за 2010 рік дані практично повністю збігаються. Таким чином, дані за Google Trends можуть дати можливість принаймні оцінити характеристики сезонності відповідних динамічних рядів.

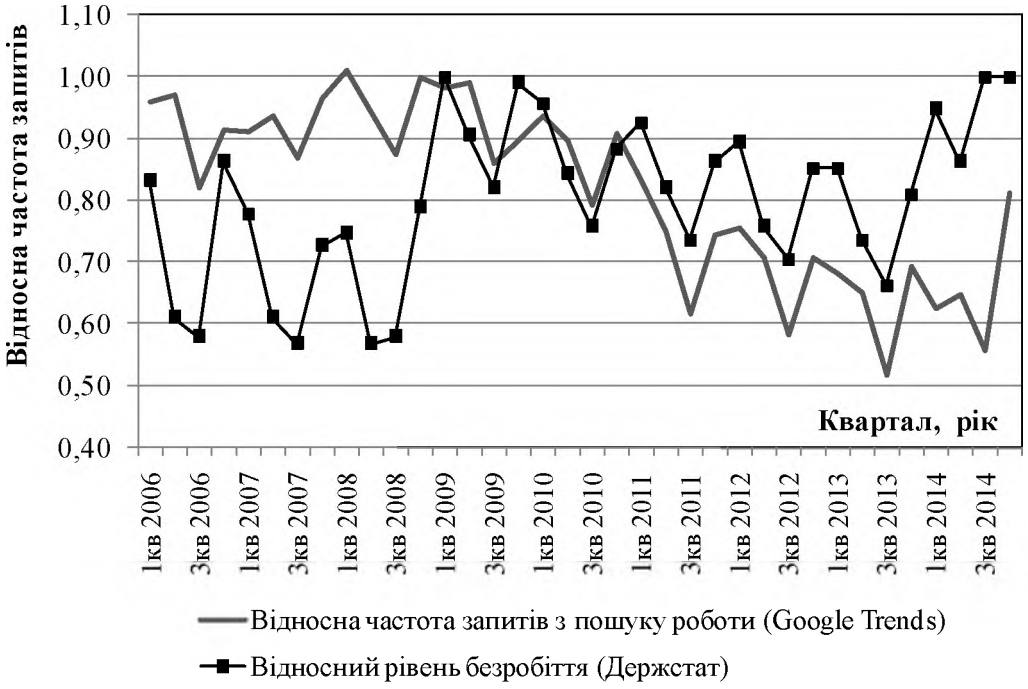

Рис. 1. Порівняння відносного рівня безробіття за методологією МОП та відносної кількості запитів щодо пошуку роботи за даними Google Trends по Україні в цілому за 2006–2014 роки

Слід ураховувати, що результати пошуку інформації в Інтернеті суттєво залежать від формулювання запитів. Для прикладу на рис. 2 наведені дані щодо пошуку житла в Україні залежно від формулювання запиту та його мови.

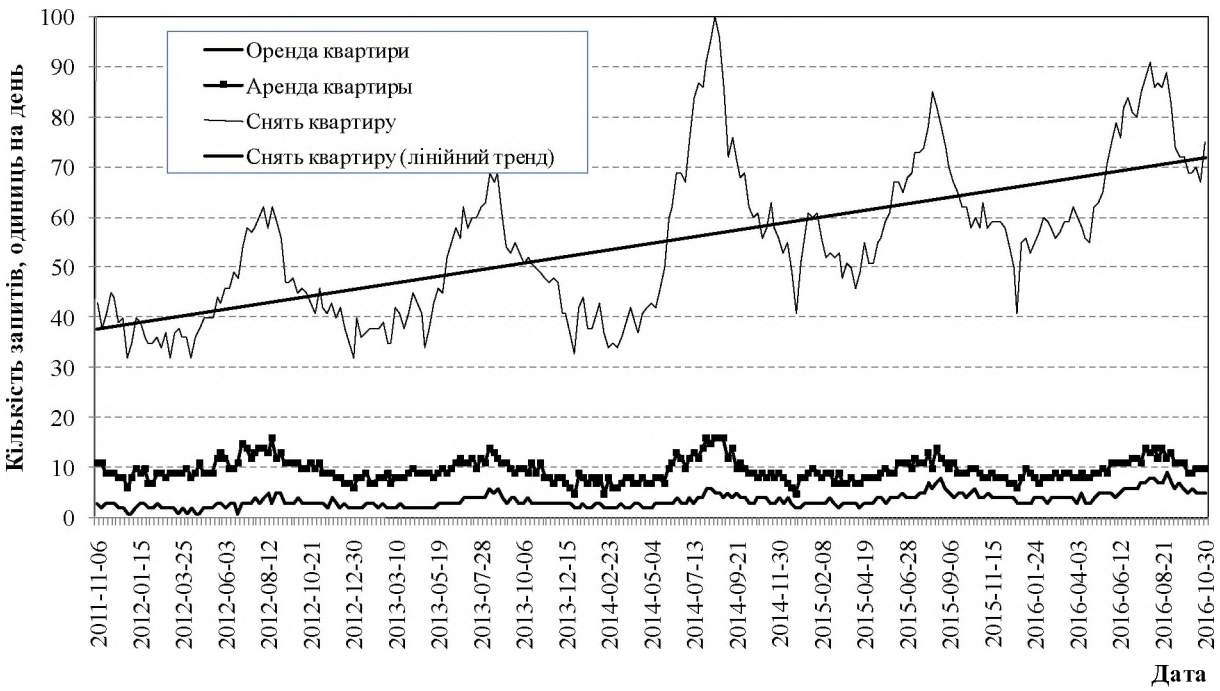

Рис. 2. Порівняння кількості запитів щодо пошуку житла в Україні залежно від формулювання та мови запиту

З наведених даних видно і тренди, і сезонні особливості, і особливості впливу мови та формулювання запитів. Така інформація є корисною для тих, хто здійснює підприємницьку діяльність на ринку житла. Але з її використанням в офіційній статистиці існує багато проблем. Так, невідомі:

– загальні чисельності груп населення, які шукають житло, зокрема тих, хто не користується для цього Інтернетом;

– невідома кількість повторів у цих даних, оскільки хтось може здійснювати пошук щоденно, а хтось – лише час від часу;

– невідомо здійснюють пошук особи, які шукають житло, чи посередники, забудовники і т. д.

Для розв'язання цих проблем, як зазначалося раніше, необхідно мати статистичні дані, зібрані й оброблені за певними процедурами.

Безумовно, наявність і доступність "великих даних", зокрема даних з Інтернету, вимагає подальшого розвитку статистичної методології у напрямі забезпечення можливості оцінювання статистичних показників з використанням даних із нових джерел і підвищення якості статистичної інформації на цій основі. Ефективним підходом тут є удосконалення системи статистичних показників: вочевидь і офіційній статистиці у багатьох випадках достатньо відображати тенденції та інтенсивність процесів без обов'язкової прив'язки до відповідних генеральних сукупностей. Власно, зараз це неявно здійснюється, зокрема в офіційних обстеженнях підприємств, де все важче отримати інформацію від малих підприємств, мікропідприємств та приватних підприємців і показники оцінюються фактично за не дуже репрезентативною вибіркою підприємств, хоча і поширюються на всю їх сукупність. В обстеженнях домогосподарств у багатьох, насамперед розвинених, країнах рівень участи становить менше 50%, а результати поширюються на всю сукупність, хоча зрозуміло, що оцінки зміщені, як правило – у бік бідніших домогосподарств, домогосподарств сільської місцевості, домогосподарств з особами пенсійного віку. Для подолання цих проблем офіційних обстежень все ширше застосовуються методи моделювання з метою як "виправлення" результатів офіційних обстежень, так і більш адекватного аналізу відповідних явищ і процесів.

Статистичне моделювання, моделювання поведінки, зокрема переваг, настроїв, інформаційного впливу, безумовно є ефективним інструментарієм використання "великих даних", роль якого буде зростати й у подальшому.

Важливу роль у можливості використання "великих даних" в офіційній статистиці відіграє відповідний потенціал працівників органів статистики. Протягом останніх приблизно 20-ти років ефективність роботи статистичних служб майже всіх країн стала суттєво залежати від їх можливостей щодо обробки значних обсягів даних. Мова йде, насамперед, про первинні дані вибіркових та суцільність обстежень, застосування спеціалізованих пакетів статистичних програм, використання моделей при оцінюванні показників, сезонних коригуваннях тощо. Зараз, навіть більш стрімко, для статистиків зростає новий виклик: вони мають володіти достатньо глибокими знаннями й солідни-

ми навичками зі сфери інформаційних технологій. Так, можливість адекватного та ефективного використання даних з сайтів, наприклад для своєчасної (швидкої) оцінки індексів споживчих цін або динаміки зміни та структури сукупності вакансій, вимагає чітких уявлень про особливості побудови та функціонування відповідних сайтів та он-лайн сервісів. Крім того, статистики мають володіти навичками роботи зі спеціальними інструментами (програмами) автоматичного збирання даних з сайтів, які мають загальну назву web scraper або web crawler [8]. Такими програмами є, зокрема, *Nutch/Solr* (https://nutch.apache.org), *JSOUP* (http://jsoup.org), *HTTrack* (http://www.httrack.com/), *import.io* (https://www.import.io/), причому останні дві поширюються на безоплатній основі. Ці інструменти дають можливість у автоматичному режимі збирати різноманітну інформацію з сайтів та формувати таблиці встановлених форм, не вимагаючи від користувачів глибоких знань у сфері web-програмування.

Отже, розвиток сучасної статистичної методології спрямований, зокрема, і на розв'язання проблем, пов'язаних з можливістю використання "великих даних", тобто розробки методів об'єднання даних з різних джерел, методів моделювання при оцінюванні показників, методів формування та використання синтетичних сукупностей одиниць спостереження та ін. Результати оцінки ситуації з потенціалом використання "великих даних" в офіційній статистиці свідчать, що насправді альтернативи цьому ані для статистичних органів, ані для користувачів інформації немає. Користувачі – суспільство й окремі громадяни, бізнес-середовище, експертне середовище, науковці, ЗМІ, органи державної влади, політики – вже використовують і будуть все ширше використовувати інформацію з джерел "великих даних". Статистичні органи мають відіграти свою роль як виробника максимально об'єктивних і якісних даних, просіюючи "великі дані" через сито статистичних процедур і тим самим формуючи для користувачів точні, своєчасні, зрозумілі кількісні та якісні характеристики різних аспектів суспільного життя, функціонування економіки, державних інститутів тощо.

Порівняно з традиційними джерелами даних, "великі дані" характеризуються такими перевагами, як відносна дешевизна, висока своєчасність та новизна, що відкриває можливість оцінки характеристик та процесів, які неможливо було оцінити на основі звичайних джерел даних. Основними проблемами використання таких даних є труднощі у встановленні реального ступеня охоплення цільових сукупностей, можливість суттєвих зміщень оцінок показників на основі "великих даних", доступність останніх, необхідність збереження й обробки дуже великих обсягів даних, ризики, пов'язані з можливістю маніпуляцій з даними та ін.

Можливість використання "великих даних" в офіційній статистиці суттєво залежить від кваліфікації працівників. Сучасний статистик або статистики найближчого майбутнього мають володіти певними специфічними знаннями та навичками з інформаційних технологій, а саме: знати особливості побудови та функціонування web-сайтів, он-лайн сервісів, систем мобільного зв'язку тощо; використовувати інструменти автоматизованого збирання даних з сайтів, обробки значних обсягів інформації; вміти виконувати розрахунки та зберігати дані на основі "хмарних" технологій тощо.

Значну увагу в подальших дослідженнях доцільно приділити вдосконаленню систем статистичних показників на основі більш широкого використання показників, отриманих на основі "великих даних".

Список використаних джерел

1. Майер-Шенбергер В. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / В. Майер-Шенбергер, К. Кукьер. – М. : Манн, Иванов и Фербер, 2014. – 240 с.

2. Френкс Б. Укрощение больших данных: как извлекать знания из массивов информации с помощью глубокой аналитики / Б. Френкс. – М. : Манн, Иванов и Фербер, 2014. – 352 с.

3. Handbook on Data Quality Assessment Methods and Tools [Electronic resource] // European Commission, Eurostat, 2007. – Access mode :
http://ec.europa.eu/eurostat/ramon/statmanuals/files/Handbook_on_data_qual_assess_tools.pdf

4. Pfeffermann D. Methodological Issues and Challenges in the Production of Official Statistics / D. Pfeffermann // Journal of Survey Statistics and Methodology. – 2015. – Vol. 3, № 4. – P. 425-483.

5. High availability in clouds: systematic review and research challenges [Electronic resource] / P. T. Endo, M. Rodrigues, G. E. Goncalves et al. / Journal of Cloud Computing: Advances, Systems and Applications. – 2016. – Access mode :
http://journalofcloudcomputing.springeropen.com/articles/10.1186/s13677-016-0066-8

6. Sarndal C.-E.Model Assisted Survey Sampling / C.-E. Sarndal, B. Swensson, J. Wretman. – New York : Springer, 1992. – 695 p.

7. Google Trends [Electronic resource]. – Access mode :
https://www.google.com/trends/

8. Using Internet as a Data Source for Official Statistics: a Comparative Analysis of Web Scraping Technologies [Electronic resource] // G. Barcaroli, M. Scannapieco, M. Scarno, D. Summa. – Access mode : http://www.academia.edu/20268756/Forecasting_skyrocketing_unemployment_with_big_data

References

1. Maier-Shenberher, V., & Kuker, K. (2014). *Bolshie dannye. Revoliutsiia, kotoraia izmenit to, kak my zhyvem, rabotaem i myslim [Big data: A Revolution That Will Transform How We Live, Work, and Think].* Moscow: Mann, Yvanov i Ferber [in Russian].

2. Frenks, B. (2014). *Ukroshchenie bolshykh dannykh: kak izvlekat znaniia iz massivov informatsii s pomoshchiu hlubokoi analitiki [Taming the Big Data: How to extract knowledge from data arrays using deep analytics].* – Moscow: Mann, Yvanov i Ferber [in Russian].

3. Handbook on Data Quality Assessment Methods and Tools (2007). European Commission, Eurostat. *ec.europa.eu.* Retrieved from http://ec.europa.eu/eurostat/ramon/statmanuals/files/Handbook_on_data_qual_assess_tools.pdf [in English].

4. Pfeffermann, D. (2015). Methodological Issues and Challenges in the Production of Official Statistics. *Journal of Survey Statistics and Methodology, Vol. 3, 4 425-483* [in English].

5. Endo P. T., Rodrigues, M., Goncalves, G. E., Kelner, J., Sadok, D. H., Curescu, C. et al. (2016). High availability in clouds: systematic review and research challenges. Journal of Cloud Computing: Advances, Systems and Applications. *journalofcloudcomputing.springeropen.com.* Retrieved from http://journalofcloudcomputing.springeropen.com/articles/10.1186/s13677-016-0066-8 [in English].

6. Sarndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer [in English].

7. Google Trends. www.google.com. Retrieved from https://www.google.com/trends/ [in English].

8. Barcaroli, G., Scannapieco, M., Scarno, M., & Summa, D. (2015). Using Internet as a Data Source for Official Statistics: a Comparative Analysis of Web Scraping Technologies. *www.academia.edu.* Retrieved from http://www.academia.edu/20268756/Forecasting_skyrocketing_unemployment_with_big_data [in English].

В. Г. Сариогло,
доктор экономических наук,
заведующий отделом,
Институт демографии и социальных исследований
имени М. В. Птухи НАН Украины,

“Большие данные” как источник информации и инструментарий для официальной статистики: потенциал, проблемы, перспективы

Рассмотрены вопросы, связанные с потенциальной возможностью использования в официальной (государственной) статистике так называемых “больших данных”. Освещены их преимущества, среди которых своевременность, широкий охват определенных частей целевых совокупностей, сокращение расходов на их получение. Обозначены проблемы, которые необходимо решать при использовании "больших данных". Аргументировано наличие в прикладной и в официальной статистике прототипов инструментов, которые при надлежащем их развитии и адаптации позволят решить основные из указанных проблем.

Ключевые слова: *“большие данные”, источники информации, статистический инструментарий, официальная статистика, информационные технологии.*

V. H. Sarioglo,
DSc in Economics,
Head of Department,
Ptoukha Institute for Demography and Social Studies
of the NAS of Ukraine

“Big Data” as an Information Source and a Toolkit for Official Statistics: Capacities, Problems, Prospects

Issues are discussed, related with potential use by official statistics of the so called “Big Data”, which refers to data extracted from websites, mobile phones, cash machines in retail sales networks, traffic surveillance cameras etc. These data are nicknamed as “big” mainly due to large scopes, not enabling for their processing by standard statistical tools but requiring special software and techniques.

It is argued that "Big Data" have advantages such as timeliness, wide coverage of targeted population segments; their collection does not require special questionnaires or surveys, training or recruiting numerous paid personnel like supervisors or interviewers. When "Big Data" are used, accuracy requirements can be loosened, analysis of phenomena and processes can be made by quite simple procedures. As scopes of these data are increasing incessantly, often second by second, the only thing to do is to process them in a proper way, to analyze and use the output information.

It is emphasized that use of "Big Data" is complicated due to the need to address problems like indeterminacy of the covered data sets; bias of estimates; accessibility of data, because they are mostly collected by private companies or belong to them; protection of private data, storage of large scopes of "Big Data" and their processing; statistical incorporation of numerous large data sets; risks of potential manipulation with data etc.

Arguments are given that applied and official statistics have prototypes of tools capable to solve a major part of the above problems, once properly developed and adapted. They include methods for calibration of survey results, statistical aggregation of data, or model-based assessment of data. As regard "cloud" technologies for data storage and processing, their use can solve the problems of weak capacity of data carriers in statistical offices, and the problems of storage of private and confidential data.

Results of studies conducted by leading statisticians of our days demonstrate that official statistics has no alternatives to use of "Bid Data". The sooner this advanced field of statistics and information technologies comes in focus of the State Statistics Service, universities and research institutions, the easier new information sources and new statistical toolkit can be integrated in the official statistics within the forthcoming ten or fifteen years.

Keywords: *"Big Data", information sources, statistical toolkit, official statistics, information technologies.*

Бібліографічний опис для цитування:

Саріогло В. Г. "Великі дані" як джерело інформації та інструментарій для офіційної статистики: потенціал, проблеми, перспективи / В.Г. Саріогло // Статистика України. – 2016. – № 4. – С. 12–19.