УДК 001.103-022.56:[31:314]:316.485.26
Jel Classiffication: C82, H56, J11
doi: 10.31767/nasoa.1-2-2023.01

О. О. ГОРОБЕЦЬ,
кандидат економічних наук, доцент кафедри,
e-mail: babutska@ukr.net
ORCID: https://orcid.org/0000-0001-5433-6448;
С. П. ЧЕРВОНА,
кандидат економічних наук, доцент,
доцент кафедри,
e-mail: SPChervona@nasoa.edu.ua
ORCID: https://orcid.org/0000-0003-1581-4448;
кафедра статистики, інформаційних технологій
та математичних методів в економіці,
Національна академія статистики, обліку та аудиту

Можливості використання великих даних у рамках статистичного вивчення населення в умовах війни

У статті акцентується увага на неможливості повного і достовірного вивчення населення та умов його життя в умовах війни. Автори звертають увагу на заповнення прогалин, які виникли внаслідок війни, у статистичних даних з допомогою альтернативних джерел у рамках статистичного вивчення населення. Досліджуючи питання імплементації великих даних у соціальну та демографічну статистику, було зосереджено увагу на тих підходах, які наразі можуть використовуватися на практиці: аналізі результатів опитування; аналізі соціальних медіа; аналізі даних про здоров'я; геопросторовому аналізі; аналізі населення. У рамках дослідження розкрито також питання використання офіційною статистикою мікроданих, інтелектуального аналізу тексту, машинного навчання. Обґрунтовано, що імплементація великих даних у процеси статистичного вивчення населення – це лише питання часу, з огляду на стрімкий розвиток цифровізаційних процесів в Україні.

Ключові слова: *офіційна статистика, демографічна статистика, соціальна статистика, аналіз результатів опитування, аналіз соціальних медіа, аналіз даних про здоров'я, геопросторовий аналіз, аналіз населення.*

O. HOROBETS,
PhD in Economics, Associate Professor of
Department of Statistics, Information Technologies
and Mathematical Methods in Economics;
S. CHERVONA,
PhD in Economics,
Associate Professor, Associate Professor of
Department of Statistics, Information Technologies
and Mathematical Methods in Economics;
National Academy of Statistics, Accounting and Audit

The Usability of Big Data in Statistical Studies of the Population in the Conditions of War

The war in Ukraine aggravated economic, social, geopolitical, and environmental problems, caused a devastating blow on physical and mental health of the Ukrainian population, worsened its quantitative and structural parameters. Apart from violations of human rights and the occurrence of mental traumas among adults and children alike, terrorist methods of warfare used by the Russian Federation led to heavy losses of the civil population and forced replacements in search for safety.

The conventional problems of the Ukrainian official statistics were added by another one posed by the war: the impossibility to study the population and their living standards in the conditions of war. The authors drew attention to the need to fill the gaps in statistical data by use of alternative war-specific sources in course of the statistical studies of the population. When investigating the issue of implementing big data in the social and demographic statistics, emphasis was made on the methods that could be fit for practical applications now: analysis of surveys, analysis of social media, analysis of health data, geospatial analysis and population analysis. The study allowed for highlighting issues of using microdata, intellectual analysis of text or machine training by the official statistics. It was revealed that the issue of implementing data in the official statistics, for population studies in particular, was but a matter of time, considering the rapid development of digitalizing processes in Ukraine.

Keywords: *official statistics, demographic statistics, social statistics, analysis of survey, analysis of social media, analysis of health data, geospatial analysis, analysis of population.*

Постановка проблеми. Війна в Україні вкрай загострила економічні, соціальні, геополітичні, екологічні проблеми, завдала руйнівного удару по фізичному та психічному здоров'ю населення, а також по чисельності та структурі населення нашої країни. Окрім порушень прав людей та психологічного травмування як дорослих, так і дітей, терористичні методи ведення Росією активної фази війни призвели й до значних людських втрат серед мирного населення та зростання потоку вимушених переміщень у пошуку порятунку.

Перед офіційною статистикою України з-посеред інших проблем війною було спричинено й таку: неможливість повного і достовірного вивчення населення та умов його життя.

Проблема неповноти даних, яка виникла у статистиці внаслідок війни, зокрема і в зміні чисельності та структури населення, умовах його життя, ще недостатньо висвітлена у наукових працях, однак існує багато робіт, присвячених питанням імплементації великих даних в офіційну статистику як альтернативних, а також тих, які дозволяють отримати ширше уявлення про структуру населення, його особливості та переміщення. Отже, великі дані можуть бути цінним ресурсом для статистичних досліджень населення, оскільки вони дозволяють збирати й аналізувати величезні обсяги даних із різноманітних джерел.

З огляду на зазначене, пошуки альтернативних методів збирання даних для офіційної статистики набули пріоритетного значення. Це й обумовило проведення дослідження з метою виявлення та представлення альтернативних методів статистичного вивчення населення, які б дозволили фахівцям використовувати їх в умовах війни.

Аналіз останніх досліджень і публікацій. У 2019 р. під час Конференції європейських статистиків, присвяченої питанням дослідження населення ЄС, наголошувалося, що статистичне вивчення населення має першочергове значення для будь-якої статистичної системи. У матеріалах Конференції зазначено, що дослідження населення, є мабуть, найстарішою статистикою, разом з тим у статистичній спільноті відбулося небагато дискусій щодо методів вивчення населення. На сьогодні майже відсутні рекомендації міжнародної статистичної спільноти, присвячені цій темі [1]. Вважаємо, що виявлені проблеми, пов'язані з дослідженням населення у ЄС, також актуальні для України, де вони додатково загострилися через війну.

У своїй статті О. Осауленко акцентує увагу на тому, що сьогодні офіційна статистика України практично позбавлена всіх основних джерел статистичних даних, а саме: первинних даних, отримуваних з допомогою статистичних спостережень та обстежень (від суб'єктів комерційної діяльності, фізичних осіб, домашніх господарств, а також державних структур, суспільних об'єктів тощо), вторинних даних з адміністративних джерел, а також інформації, що збиралася спеціалізованими урядовими агенціями [2].

Зважаючи на виклики, які постали перед вітчизняними статистиками у зв'язку з війною, доцільно зазначити, що після Другої світової війни для вивчення населення почали широко використовувати вибіркові обстеження. Демографи з Бюро перепису населення у співпраці з прикладними статистиками розробляли вибіркові методи для задоволення вимог щодо визначення своєчасних заходів, спрямованих на зниження

рівня безробіття [3]. До кінця 1950-х років у США вибіркові обстеження були основними у дослідженнях сфери соціальних наук [4].

Аналізуючи роль та значення великих даних в інформаційному забезпеченні соціальної та демографічної статистики, I. Bernal зазначає, що великі джерела даних дають суттєві переваги для статистичного виробництва у цих галузях, наприклад збільшення охоплення, підвищення рівня деталізації, частоти, а також забезпечують своєчасність і скорочення витрат [5].

На думку E. Magpantay, статистичне співтовариство зобов'язане досліджувати використання великих даних з метою задоволення очікувань суспільства щодо покращення продуктів та реалізації ефективніших способів роботи статистичних служб, зокрема в частині моніторингу цілей сталого розвитку, шляхом покращення своєчасності, деталізації та актуальності показників без шкоди для їх неупередженості та методологічної надійності [6]. Цікавою є думка фахівців з охорони здоров'я із міжнародної довідкової організації щодо населення (Population Reference Bureau, PRB), які зазначають що використання великих даних забезпечує ефективний розподіл ресурсів за унеможливлення проведення підрахунків населення протягом тривалого часу внаслідок конфліктів або воєн, зокрема щодо визначення найпривабливіших, з погляду охоплення населення, місць для розташування клінік, а також дає змогу територіально досліджувати насиченість надання тих чи інших медичних послуг у регіоні [7].

Натомість S. Ruggles зосереджує увагу дослідників на питанні використання мікроданих у рамках вивчення населення. Він зазначає, що ця величезна нова скарбниця – мікродані у поєднанні з новими технологіями – має потенціал трансформувати просторово-часовий аналіз демографічної поведінки та економічної діяльності. Зокрема послідовні великомасштабні мікродані, які охоплюють багато десятиліть і враховують національні кордони з дрібними географічними деталями, створюють унікальну лабораторію для вивчення демографічних процесів і тестування соціальних і економічних моделей. Ці дані дозволять проводити нові види досліджень, які враховують, зокрема:

- ✓ сегрегацію проживання;
- ✓ міграції та моделі поселення мігрантів;
- ✓ розростання міст;
- ✓ економічний та ідейний контекст зниження народжуваності, зростання смертності та співжиття між поколіннями;
- ✓ депопуляцію у сільській місцевості та консолідацію сільського господарства;
- ✓ виявлення концентрованої бідності;
- ✓ причини та рівні змін в екосистемах як функцію взаємодії людини й довкілля [8].

З огляду на зазначене, варто додати, що мікродані характеризуються якісною структурою та надають узгоджену інформацію від індивіда до цілої нації (індивід → домогосподарство → мікрорайон → регіон → нація).

S. Keller та ін. пропонують групувати великі дані на органічні (характеризуються органічною появою, а саме: дані про місцезнаходження, реєстрацію виборців, транзакції, медичні записи, метеорологічні та сейсмічні записи та ін.) та вироблені дані (вторинні, що характеризуються штучною появою, тобто є результатом попередньої дії, наприклад адміністративні дані, зібрані в рамках перепису населення). Науковці зазначають, що використовувати органічні великі дані із соціальних мереж і комерційних транзакцій для кращого розуміння суспільства дуже цікаво [9].

Виклад основного матеріалу. Досліджуючи питання імплементації великих даних у соціальну та демографічну статистику в рамках заповнення прогалин, які виникли внаслідок війни, доцільно зосередити увагу на тих способах, які наразі можна використовувати на практиці, а саме:

1) аналіз результатів опитування;
2) аналіз соціальних медіа;
3) аналіз даних про здоров'я;
4) геопросторовий аналіз;
5) аналіз населення.

У рамках аналізу результатів опитування великі дані можна, зокрема, використовувати для аналізу відповідей на запитання у ході інтерв'ювання/анкетування великих

вибірок населення. Це, своєю чергою дозволить дослідникам визначати тенденції та закономірності даних.

Серед найпоширеніших методів аналізу відповідей на опитування доцільно виокремити такі:

- попередня обробка даних (очищення, стандартизація та підготовка даних до подальшого аналізу);
- описовий аналіз (узагальнення та візуалізація даних для визначення закономірностей і тенденцій);
- прогнозне моделювання (статистичні методи для прогнозування майбутніх результатів на основі даних опитування, зокрема регресійний аналіз, дерева рішень та ін.);
- машинне навчання (навчання алгоритмів для виявлення шаблонів у даних задля виявлення закономірностей, тенденцій та зв'язків, які не можуть бути виявлені в рамках традиційного статистичного аналізу) і створення прогнозів на основі цих шаблонів);
- інтелектуальний аналіз тексту. У рамках цієї групи використовуються, зокрема, методи обробки природної мови (Natural language processing, NLP) для аналізу текстових відповідей на запитання відкритого опитування з метою виявлення загальних тем, ураховуючи:
  а) аналіз настроїв (Sentiment analysis), тобто розуміння емоцій з допомогою програмного забезпечення для визначення загального настрою відповідей [10];
  б) кластерний аналіз (групування у кластери респондентів на основі їх відповідей із формуванням підгруп респондентів зі схожими характеристиками чи ставленням до певних явищ і процесів).

На сьогодні одним із потужних механізмів формування екосистеми великих даних є соціальні медіа. Аналіз соціальних медіа – це здатність збирати дані із соціальних каналів, знаходити у них сенс, наприклад щодо підтримки бізнес-рішень, а також вимірювати ефективність дій на основі цих рішень [11]. Аналіз соціальних мереж використовує спеціально розроблені програмні платформи, які працюють подібно до інструментів вебпошуку. Дані про ключові слова чи теми витягуються з допомогою пошукових запитів або вебсканерів, які охоплюють канали. Фрагменти тексту повертаються, завантажуються в базу даних, класифікуються й аналізуються для отримання значущих ідей. Аналіз соціальних медіа ґрунтується на концепції свого роду соціального прослуховування.

A. Olteanu зі співавторами зосереджують увагу на тому, що соціальні дані сформували абсолютно нові галузі досліджень на перетині інформатики та соціальних наук, таких як обчислювальна соціальна наука та соціальне обчислення, галузі, які також розгалужуються на багато суміжних сфер застосування, включаючи кризову інформатику, обчислювальну журналістику та ін. [12].

Дані соціальних медіа (або соціальні дані) відкривають безпрецедентні можливості для відповідей на важливі питання, пов'язані з політикою, зокрема суспільним та політичним устроями. Так, аналіз соціальних медіа активно використовують в галузі охорони здоров'я [13; 14].

У рамках дослідження, проведеного D. Valdez із колегами по Indiana University, проаналізовано 86,6 тис. загальнодоступних твітів англійською мовою, вивчено еволюцію хештегів у часі з допомогою тематичного моделювання (метод прихованого розподілу Діріхле), досліджено колективні зміни в суспільному настрої щодо еволюції циклів новин про пандемію, розглядаючи середньодобовий настрій усіх твітів хронології із використанням інструменту Valence Aware Dictionary and Sentiment Reasoner (VADER). Вибір на користь соціальних медіа науковці обґрунтували тим, що традиційні методи опитування є трудомісткими та дорогими, а для дослідження потрібні своєчасні та проактивні джерела даних, щоб реагувати на швидкозмінний вплив політики охорони здоров'я на психічне здоров'я населення. Зараз багато людей у США використовують соціальні медіа-платформи, зокрема такі як Twitter, щоб висловити найдрібніші подробиці свого повсякденного життя та соціальних стосунків [15].

Досліджуючи питання забезпечення великими даними прогалин у статистичних

даних, варто звернути особливу увагу на збирання та аналіз даних про здоров'я. Адже у цьому випадку великі дані формуватимуться із масиву медичних заяв, записів, цифрової фіксації анамнезу пацієнтів та рекомендацій щодо лікування. Ці дані можуть сформувати уявлення про здоров'я та благополуччя населення, допомогти визначити фактори, які впливають на здоров'я. Однак своєрідним підводним каменем зазначеного є те, що медична галузь є найменш цифровізованою, а отже, у цьому середовищі слабкіше розвинена екосистема великих даних, яка насамперед характеризується низькою якістю даних та їх погано організованими наборами [16].

Зазвичай офіційна статистика оприлюднює дані та формує звіти на основі реєстрів хронічних захворювань, які є досить обмеженими, а тому потребують значних удосконалень і аж ніяк не характеризують населення та стан його здоров'я. Разом з тим існують методи, з допомогою яких науковці усе ж таки проводять деталізоване вивчення та дослідження пацієнтів. Використання цифрових версій медичних записів пацієнтів, включаючи інформацію про їх діагнози, лікування та результати, дає змогу сформувати повнішу інформацію щодо захворюваності населення.

Серед найпоширеніших методів аналізу великих даних про здоров'я зазначимо насамперед аналіз електронних медичних карток (або записів), прогностичну аналітику, аналіз клінічних випробувань, дані спостережень за станом здоров'я, аналіз інформації, отримуваної із носимих пристроїв (wearables devices). Наведені методи мають і позитивні, і негативні сторони як з боку методології (статистики), так і з боку практичного втілення (медицини).

У контексті аналізу електронних медичних карток варто навести слова S. Upadhyay та H. Hu, які зазначали: незважаючи на те, що переваги аналізу медичних карток добре сприйняті, попередні дослідження показують неоднозначні результати їх упровадження [17]. Частково підтверджують наведене твердження M. A Gianfrancesco та N. D. Goldstein. На їх думку, електронні записи про стан здоров'я широко використовуються в епідеміологічних дослідженнях, але достовірність результатів залежить від припущень, зроблених щодо системи охорони здоров'я, пацієнта та постачальника послуг [18].

Разом з тим у дослідженні R. Verheij та ін. було виявлено 13 потенційних джерел упередженості даних (так званого забруднення даних) на основі аналізу медичних карток, що охоплюють майже усі аспекти надання медичної допомоги – від вибіркового входу в систему охорони здоров'я, варіацій у догляді та практиках документування до ідентифікації і вилучення потрібних даних для аналізу [19].

У рамках аналізу медичних карток варто виокремити низку ризикових моментів, з якими можуть зіштовхнутися науковці у ході дослідження:

- ✓ помилки в репрезентативності даних;
- ✓ обмежена доступність даних та можливість їх інтерпретації (у тому числі узгодженість);
- ✓ неструктурованість даних;
- ✓ відсутність можливості вимірювання даних;
- ✓ відсутність інформації про пацієнта.

На сьогодні найпопулярнішими базами медичних даних є The Medical Information Mart for Intensive Care [20], PCORnet [21], The National Health Services (NHS England) [22], eICU [23], Veterans Health Information Systems and Technology Architecture (VistA) [24], The National Surgical Quality Improvement Project [25].

У тій чи іншій мірі наступні методи аналізу великих даних про здоров'я ґрунтуються на цифровому середовищі медичних карток або ж записів. Так, застосування прогностичної аналітики, допомагало б виявити шляхом ідентифікації факторів ризику пацієнтів із високим ризиком розвитку певних станів здоров'я (наприклад, передбачити появу серцево-судинних захворювань). Однак для належного використання цього методу необхідно мати якісну систему медичних записів. Ідентична ситуація виникає і з аналізом клінічних випробувань, в рамках якого з допомогою великих даних з'явилася б можливість визначати найбільш ефективні методи лікування для різних груп пацієнтів, тим самим сприяючи розробці ефективних випробувань та розвитку персоналізованих методів лікування.

Також великі дані сприяли б удосконаленню спостережень за станом здоров'я населення у частині моніторингу та відстеження поширення інфекційних захворювань

(наприклад, Covid-19) шляхом аналізу даних із лікарських записів, соціальних мереж пацієнтів та пошукових запитів в інтернеті. Це дало б змогу працівникам сфери охорони здоров'я своєчасно виявляти спалахи пандемій та вживати заходів для запобігання поширенню хвороби.

Цікавим з погляду вивчення населення та недостатньо опрацьованим для медичної статистики є аналіз носимих пристроїв (фітнес-трекерів, розумних годинників, окулярів доповненої реальності, розумних капелюхів, діагностів сну та ін.). У контексті великих даних варто зазначити, що у 2022 р. світовий ринок цих пристроїв оцінювався в 61,3 млрд дол. США; очікується, що у 2023–2030 рр. він у середньому зростатиме на 14,6% щорічно [26]. З допомогою таких пристроїв можна збирати велику кількість даних про здоров'я (наприклад, дані щодо частоти серцевих скорочень, режиму сну, рівня фізичної активності, рівня холестерину та ін.). Аналітику великих даних можна використовувати для виявлення закономірностей і тенденцій у різних сферах життя населення, а також для надання персоналізованих рекомендацій щодо здоров'я окремим особамТак, у рамках вивчення певної групи населення S. Huhn зі співавторами обґрунтували, що носимі пристрої можуть генерувати цінні дані для глобальних досліджень з охорони здоров'я [27]. Підтверджуючи цю думку, J. Dunn акцентував увагу на тому, що переваги споживчих пристроїв порівняно з дослідницькими й анкетуванням полягають у низькій вартості, зручності та непомітності, а також у можливості збирати дані в природному середовищі учасників дослідження. Також науковці акцентують увагу на тому, що носимі пристрої вже революціонізували біомедицину завдяки мобільному та цифровому здоров'ю, дозволяючи безперервний лонгітьюдний моніторинг стану здоров'я за межами клініки. Також носимі пристрої полегшують розробку алгоритмів для автоматизованого прогнозування, профілактики та втручання [28].

Варто зазначити, що носимі пристрої уже були використані для прогнозування спалахів інфекційних захворювань. Так, J. Radin зі співавторами використали деідентифіковані дані датчиків від 200 000 осіб, які застосовували носимий пристрій Fitbit з 01.03.2016 р. по 01.03.2018 р. у США [29]. Автори зазначають, що трекери активності та фізіологічні трекери усе частіше використовуються як у США, так і в усьому світі для моніторингу індивідуального здоров'я. Отримавши доступ до цих даних, можна покращити географічно уточнений епіднагляд за грипом у реальному часі. Така інформація є життєво необхідною для своєчасного реагування на спалахи епідемій. Варто також зазначити, що з допомогою носимих пристроїв цілком реально поліпшити медичне обслуговування в різних умовах, від лікарняного та клінічного догляду до амбулаторного догляду вдома та у географічно віддалених місцях, включаючи сільські райони та середовища з низьким рівнем ресурсів.

Разом з перевагами носимих пристроїв науковці, водночас, наголошують і на їхніх недоліках, серед яких складність вилучення необхідних даних із усього середовища великих даних, яке формують ці пристрої.

Серед потенційних методів використання великих даних для статистичного вивчення населення та умов його життя доцільно вказати геопросторовий аналіз, який уможливлює збирання, аналіз, моделювання та візуалізацію даних про місцезнаходження, а також сприяє формуванню уявлення про тенденції і поведінку населення у певних географічних регіонах. Окрім цього, цей підхід дає змогу визначати густоту рослинного покриву, ступінь вологості та температуру ґрунту, стан посівів, виявляти вирубку лісів, прогнозувати пожежі за критичними температурами, ідентифікувати розливи нафти, створювати найшвидші та найбезпечніші маршрути, бачити найвигідніше місце розташування магазину, розподіляти допомогу в райони, найбільш постраждалі від стихійних лих або війни, розподіляти освітні та медичні заклади для найповнішого задоволення потреб населення досліджуваного регіону.

Аналітика геопросторових даних (просторова регресія, кластеризація, інтерполяція, а також кригінг та ін.) ґрунтується на даних з усіх видів джерел, де реалізовано різноманітні сучасні технології, – GPS, датчиків розташування, соціальних мереж, мобільних пристроїв, супутникових зображень [30]. У контексті цього питання варто згадати позицію J. Byers та J. Gill, що наразі у прикладній статистиці існує тенденція до збільшення використання двох важливих інструментів: геопросторових моделей і великих даних [31].

Використання геопросторових даних передбачає чітке дотримання ключових етапів, якими є:

1. Збирання релевантних даних.
2. Очищення та попередня обробка даних.
3. Аналіз просторових шаблонів.
4. Візуалізація даних.
5. Використання методів машинного навчання.

З огляду на зазначене варто підкреслити, що загалом геопросторовий аналіз є цінним інструментом для розуміння складних взаємозв'язків між характеристиками населення та просторовим контекстом, у якому вони мають місце. Геопросторова статистика відіграє важливу роль у статистичному вивченні населення, оскільки дозволяє дослідникам аналізувати дані про населення у прив'язці до його географічного розташування.

Досить гострим питанням для статистиків і демографів України є перепис населення, який востаннє було проведено ще у 2001 р. Вітчизняні демографи неодноразово наголошували на потребі в актуальних даних щодо чисельності населення України, рівня його освіти, статево-вікового складу, сімейної та етнічної структури тощо. Так, Саріогло В. та Огай М. зазначають, що ситуація суттєво ускладнилася з 2014 р. та стала однією з найнагальніших для країни в умовах повномасштабної війни з 24 лютого 2022 р., коли значна кількість осіб була вимушена шукати нові місця для проживання в інших регіонах України або за її кордонами, примусово переселена до країни-окупанта або опинилася на тимчасово окупованих територіях. Також слід ураховувати людські втрати через загибель військових та цивільних осіб [32].

З огляду на наведене варто зазначити, що великі дані можна використовувати для доповнення традиційних даних перепису; це сприятиме додатковому розумінню характеристик населення та допомагатиме визначити ключові сфери його потреб. Загалом великі дані можуть підштовхнути дослідників до вибору моделей аналізу тенденцій, неочевидних при застосуванні традиційних статистичних методів. Науковці при цьому глибше розумітимуть фактори, які формують поведінку населення.

У пошуках альтернативних джерел даних для офіційної статистики, призначених для статистичного вивчення населення в умовах війни, особливе зацікавлення фахівців викликають дані мобільного позиціонування як окрема група в екосистемі великих даних. Це джерело даних демонструє потужний потенціал для моніторингу розселення населення та його мобільності, оскільки мобільні телефони широко розповсюджені, а подібні стандартизовані дані можна отримувати по всьому світу.

Завдяки широкому використанню смартфонів цифрові сліди, залишені під час використання цих пристроїв, надають цінну інформацію про діяльність людини в реальному часі. Ці цифрові сліди полегшують вивчення людської поведінки [33]. Наразі дані мобільного позиціонування визнані одним із найперспективніших нових джерел для створення швидкої та економічно ефективної статистики щодо населення та його мобільності [34]. Водночас D. Salgado та ін. зазначають, що включення цього джерела даних у регулярне виробництво офіційної статистики потребує багатьох зусиль, оскільки різноманітність великих даних та низка супутніх проблем, пов'язаних із їх імплементацією (доступ, методологія, ІТ-інструменти, якість, навички працівників), повинні бути вирішені заздалегідь. Разом з тим науковці акцентують увагу на тому, що отримані дані про населення у традиційних статистичних дослідженнях пропонують інформацію про суспільство з конкретної демографічної позиції (тобто характеризують постійне населення) із заданим ступенем просторового та часового розподілу, а дані мобільної мережі надають можливість досягти безпрецедентних просторових і часових масштабів, а також формують додаткове уявлення про населення (тобто характеризують наявне населення) [35].

Вивчаючи питання імплементації даних мобільного позиціонування в офіційну статистику, варто акцентувати увагу на проблемах, які пов'язані із цим досить чутливим видом даних та потребують негайного вирішення, а саме: надання доступу до даних, формування методологічних рамок та рамок якості, створення програмного забезпечення (або інфраструктури), визначення відповідних показників для різноманітних статистичних сфер.

Дані мобільного позиціонування науковці пропонують використовувати в таких сферах статистики:

- ✓ статистика туризму;
- ✓ статистика міграції;
- ✓ статистика населення;
- ✓ статистика транспорту та щоденних поїздок;
- ✓ статистика інформаційного суспільства;
- ✓ переміщення населення внаслідок впливів стихійних лих [36; 6].

Варто зазначити, що у 2021 р. робоча група з мобільних даних на чолі з Міжнародним союзом телекомунікацій підготувала п'ять посібників з рекомендаціями щодо використання мобільних телефонних даних для статистики переміщення під час катастроф, динамічного картографування населення, статистики інформаційного суспільства, статистики міграції та статистики туризму відповідно. Наразі ці посібники пройшли експертне рецензування та перебувають на стадії перехресної перевірки на послідовність та узгодженість [6].

До особливостей даних мобільного позиціонування слід віднести такі: місцезнаходження мобільних пристроїв можна отримати у реальному часі або історично; власники пристроїв можуть бути відомими або невідомими; дані мобільного позиціонування можуть бути отримані з використанням активних і пасивних методів збирання (так зване активне і пасивне мобільне позиціонування відповідно) [37].

Коротко зупинимося на розгляді зазначених активних і пасивних методів збирання даних. Різниця між ними полягає в тому, що для активного позиціонування формується конкретний цільовий запит для визначення місцезнаходження мобільного пристрою (створюється запит на місцезнаходження та повертається відповідь про нього), тоді як для пасивного позиціонування збираються історичні дані. Як зазначає M. Tiru, існує два основні методи визначення місцезнаходження телефону з допомогою активного позиціонування:

1) використання мережевої інфраструктури операторів мобільного зв'язку (Mobile Network Operators – MNO) та системи мобільного позиціонування (Mobile Positioning System – MPS);
2) використання даних про місцезнаходження в реальному часі з додатків мобільних пристроїв [37].

Україна вже має досвід використання даних мобільного позиціонування в рамках вивчення населення, коли у 2019 р. було оцінено чисельність наявного населення країни. Тоді це оцінювання було проведене трьома методами [38]:

1. Комбінований: дані мобільних операторів (отримана кількість номерів телефонів, якими користувалися клієнти протягом лютого – березня 2019 року) плюс дані статистичного обстеження домогосподарств (отримані дані щодо кількості карток на одного користувача мобільного зв'язку, а також оцінена частку користувачів мобільного зв'язку серед населення для кожної вікової групи та регіонів) плюс дані реєстрів (особи віком 60^+, що одержують пенсію, та діти до 14 років).

2. Комбінований: дані щодо статево-вікової структури населення плюс дані реєстрів (обчислені частки населення за статево-віковою структурою за даними Держстату та визначена чисельність осіб віком 60^+ з урахуванням пенсіонерів із ОРДЛО, які приїжджають за пенсіями. Після цього дані щодо чисельності осіб віком 60^+ екстраполювалися на дані статево-вікової структури).

3. Реєстровий: дані Державного реєстру фізичних осіб (визначалася загальна кількість зареєстрованих громадян у Державному реєстрі фізичних осіб – платників податків з урахуванням тих, хто відмовився від реєстраційного номера облікової картки платника податків (РНОКПП), і тих, хто помер) плюс дані щодо активності (за кожним РНОКПП відбувалася перевірка отримання громадянином будь-яких доходів, пенсій, звернень до інших державних установ для оформлення громадянином субсидій та/або закордонного паспорта) плюс демографічні та міграційні показники (до визначеної раніше кількості осіб додано дані щодо чисельності дітей, які вижили, та міграційне сальдо разом із кількістю іноземних громадян, які перебувають на території України). За оприлюдненими даними тогочасного оцінювання, чисельність наявного населення України складала 37 млн 289 тис. осіб.

Загалом питання вивчення населення з використанням великих даних, окрім теоретичних напрацювань, має й суто практичні аспекти втілення. Зупинимося на розробці екосистеми даних, з допомогою якої цілком можливо повноаспектно характеризувати населення. Так, відповідну низку ініціатив реалізовано у США, де використовують великі дані для вивчення населення, включно зі щорічним опитуванням американської спільноти, в якому збирають дані про демографічні, соціальні й економічні характеристики населення [39]. Уряд Сполученого Королівства запустив кілька ініціатив, які використовують великі дані для вивчення населення, включаючи Data Science Campus [40], який є частиною Офісу національної статистики (ONS). Варто зазначити, що Data Science Campus використовує великі дані та методи науки про дані, щоб допомогти вирішити соціальні та економічні проблеми у Великій Британії. У Сінгапурі впроваджено ініціативу "Розумна нація" [41], метою якої є використання технологій і даних для покращення життя своїх громадян. Китай використовує великі дані для вивчення населення протягом кількох років і вже реалізував низку ініціатив для цього, включаючи створення Національного ресурсного центру інформації про населення [42], який збирає та обробляє дані про населення з низки джерел.

Зазначене дає змогу стверджувати: оскільки технології великих даних продовжують розвиватися, цілком імовірно, що все більше країн почнуть використовувати зазначені методи з метою отримання актуальної та повнішої інформації про своє населення.

Висновки. Ураховуючи цифровий потенціал України, а також інтелектуальний потенціал у галузі ІТ-технологій та швидкість упровадження новітніх технологій на території країни, можна передбачити, що імплементація великих даних в офіційну статистику, зокрема у процеси вивчення населення, – це лише питання часу. Війна, своєю чергою, актуалізувала питання вивчення й оцінювання населення та умов його життя, тим самим змушуючи офіційну статистику шукати альтернативні способи та методи вирішення цього питання. Ураховуючи вітчизняний досвід та досвід інших країн, напрацювання дослідників, у повоєнний період офіційною статистикою, швидше за все, буде проведено модернізований перепис населення із залученням альтернативних джерел даних для отримання всебічної характеристики населення України.

У подальших дослідженнях передбачається узагальнення наукового досвіду з питань імплементації великих даних у статистичні процеси з метою подальшого розроблення методичних підходів до збирання, обробки й аналізу даних, зібраних з допомогою альтернативних джерел.

References

1. Towards a single population base in the EU. (2019). Conference of European Statisticians (Geneva, 18–20 September 2019). Working paper 2. Retrieved from https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2019/mtg1/WP2_Eurostat_Lanzieri.pdf
2. 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, 23, 5, 185–199. Retrieved from https://sit.stat.gov.pl/SiT/OnlineFirst/01_Oleksandr%20Osaulenko_Olena%20Horobets%20%20sit%2023%205.pdf
3. Rossi, P. H., Wright, J. D., & Anderson, A. B. (2013). Sample surveys: History, current practice, and future prospects. Handbook of Survey Research. P. H. Rossi, J. D. Wright, A. B. Anderson (Eds.). (pp. 1–20). New York: Academic Press. Retrieved from https://www.jstor.org/stable/522794
4. Billari, F. C., & Zagheni, E. Big Data and Population Processes: A Revolution? (2017). Proceedings of the Conference of the Italian Statistical Society (28–30 June 2017, Florence). Statistics and Data Science: new challenges, new generations. A. Petrucci, R. Verde (Eds.). (pp. 167–178). Retrieved from https://media.fupress.com/files/pdf/24/3407/3407_11724
5. UNESCAP. (2021). Big data for population and social statistics. Stats Brief, 29. Retrieved from https://www.unescap.org/sites/default/d8files/knowledge-products/Stats_Brief_Issue29_Big_data_for_population_and_social_statistics_Apr2021.pdf

6. Magpantay, E. (2022). Information society indicators for the SDGs using mobile phone big data. Mobilizing Big Data and Data Science for the Sustainable Development Goals. Session 1.4. (26 January 2022, Dubai). Retrieved from https://unstats.un.org/bigdata/blog/2021/expo2020/sessions/1-4/presentations/Magpantay_Information%20society_session1.4.pdf
7. Ashford, L. S., Kaneda, T., & Letouzé, E. (2022). Demystifying Big Data for Demography and Global Health. Population Bulletin, 76, 1. Population Reference Bureau. Retrieved from https://www.prb.org/wp-content/uploads/2022/02/population-bulletin-vol-76-no1-demystifying-big-data.pdf
8. Ruggles, S. (2014). Big Microdata for Population Research. Demography, 51 (1), 287–297. Retrieved from https://www.jstor.org/stable/42919999
9. Keller, S. A., Koonin, S. E., & Shipp, S. (2012). Big data and city living – what can it do for us? Significance. Special Issue: Big Data, 9, 4, 4–7. Retrieved from https://rss.onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2012.00583.x
10. Karn, A., Shrestha, A., Pudasaini, A., Mahara B., & Jaiswal, A. (2018). Statistic-Based Sentiment Analysis of Social Media Data. International Research Journal of Innovations in Engineering and Technology (IRJIET), 2, 5, 28–32. Retrieved from https://www.researchgate.net/publication/333943106_Statistic-Based_Sentiment_Analysis_of_Social_Media_Data
11. What is social media analytics? IBM. Retrieved December 20, 2022 from https://www.ibm.com/topics/social-media-analytics
12. Olteanu, A., Castillo, C., Diaz, F. & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Frontiers in Big Data, 2, 1–33. Retrieved from https://www.frontiersin.org/articles/10.3389/fdata.2019.00013/full
13. Gao, Ya., Xie, Z., & Li, D. (2021). Electronic Cigarette Users' Perspective on the COVID-19 Pandemic: Observational Study Using Twitter Data. JMIR Public Health Surveillance, 7 (1):e24859. doi: 10.2196/24859
14. Cho, S. E., Jung, K. & Park, H. W. (2013). Social Media Use during Japan's 2011 Earthquake: How Twitter Transforms the Locus of Crisis Communication. Media International Australia, 149 (1), 28–40. Retrieved from https://journals.sagepub.com/doi/10.1177/1329878X1314900105
15. Valdez, D., ten Thij, M., Bathina, K., Rutter, L. & Bollen, J. (2020). Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data. Journal of Medical Internet Research, 22 (12):e21418. Retrieved from https://www.jmir.org/2020/12/e21418/
16. Marshall, J., Chahin, A., & Rush, B. (2016). Review of Clinical Databases. Secondary Analysis of Electronic Health Records. MIT Critical Data. Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-43742-2_2
17. Upadhyay S., & Hu H. (2022). A Qualitative Analysis of the Impact of Electronic Health Records (EHR) on Healthcare Quality and Safety: Clinicians' Lived Experiences. Health Services Insights, 15, 1–7. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8902175/
18. Gianfrancesco, M. A., & Goldstein, N. D. (2021). A narrative review on the validity of electronic health record-based research in epidemiology. BMC Medical Research Methodology, 21, 234. DOI https://doi.org/10.1186/s12874-021-01416-5
19. Verheij, R., Curcin, V., Delaney, B., & McGilchrist, M. M. (2018). Possible sources of bias in primary care electronic health record data use and reuse. Journal of Medical Internet Research, 20 (5):e185. doi: 10.2196/jmir.9134
20. Medical Information Mart for Intensive Care. mimic.mit.edu. Retrieved December 20, 2022 from https://mimic.mit.edu/
21. The National Patient-Centered Clinical Research Network. PCORnet. Retrieved December 20, 2022 from https://pcornet.org/
22. National Health Services (NHS in England). www.england.nhs.uk. Retrieved December 20, 2022 from https://www.england.nhs.uk/
23. eICU Collaborative Research Database. eicu-crd.mit.edu. Retrieved December 20, 2022 from https://eicu-crd.mit.edu/
24. VA Technical Reference Model v 23.2. U. S. Department of Veterans Affairs. Retrieved from https://www.oit.va.gov/Services/TRM/TRMRedirectPage.aspx?type=W^&tid=8338^

25. National Surgical Quality Improvement Program. The American College of Surgeons. Retrieved from https://www.facs.org/quality-programs/data-and-registries/acs-nsqip/
26. Wearable Technology Market Size, Share & Trends Analysis Report By Product (Head & Eyewear, Wristwear), By Application (Consumer Electronics, Healthcare), By Region (Asia Pacific, Europe), And Segment Forecasts, 2023–2030. (2023). www.grandviewresearch.com. Retrieved from https://www.grandviewresearch.com/industry-analysis/wearable-technology-market
27. Huhn, S., Matzke, I., Koch, M., Gunga, H. Ch., Maggioni, M. A., & Sié, A. et al. (2022). Using wearable devices to generate real-world, individual-level data in rural, low-resource contexts in Burkina Faso, Africa: A case study. Front Public Health, 10:972177. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9561896/#B2
28. Dunn, J., Runge, R., & Snyder, M. (2018). Wearables and the medical revolution. Personalized Medicine, 15, 5. Retrieved from https://www.futuremedicine.com/doi/10.2217/pme-2018-0044
29. Radin, J. M., Wineinger, N. E., Topol, E. J. & Steinhubl, S. R. (2020). Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. Lancet Digit Health, 2(2):e85-e93. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8048388/
30. Geospatial Analytics Definition. HEAVY.AI. Retrieved December 20, 2022 from https://www.heavy.ai/technical-glossary/geospatial-analytics
31. Byers, J. S., & Gill, J. (2022). Applied Geospatial Bayesian Modeling in the Big Data Era: Challenges and Solutions. Mathematics, 10 (21), 4116. Retrieved from https://www.mdpi.com/2227-7390/10/21/4116
32. Sarioglo, V., & Ogay, M. (2022). Approach to population estimation in Ukraine using mobile operators' data. Statistics in Transition new series and Statistics of Ukraine. Joint Special Issue: A New Role for Statistics, 23, 5, 185–198. Retrieved from https://sit.stat.gov.pl/SiT/OnlineFirst/013%20Sarioglo%20V%20sit%2023%205%20specjalny.pdf
33. Okmi, M., Por, L., Ang, T. F. & Ku, Ch. S. (2023). Mobile Phone Data: A Survey of Techniques, Features, and Applications. Sensors, 23 (2), 908. Retrieved from https://www.mdpi.com/1424-8220/23/2/908
34. Aasa, A., Kamenjuk, P., Saluveer, E, Šimbera, J., & Raun J. (2021). Spatial interpolation of mobile positioning data for population statistics. Journal of Location Based Services, 15 (4), 239–260. Retrieved from https://www.tandfonline.com/doi/full/10.1080/17489725.2021.1917710
35. Salgado, D., Sanguiao, L., Oancea, B., Barragán, S., & Necula, M. (2021). An end-to-end statistical process with mobile network data for official statistics. EPJ Data Science, 10, 20. Retrieved from https://doi.org/10.1140/epjds/s13688-021-00275-w
36. Report of the Committee of Experts on Big Data and Data Science for Official Statistics. (2022). E/CN.3/2022/25. unstats.un.org. Retrieved from https://unstats.un.org/bigdata/documents/reports/UNCEBD%20report%20-%202022-25-BigData-E.pdf
37. Tiru, M. (2014). Overview of the Sources and Challenges of Mobile Positioning Data for Statistics. unstats.un.org. Retrieved from https://unstats.un.org/unsd/trade/events/2014/beijing/Margus%20Tiru%20-%20Mobile%20Positioning%20Data%20Paper.pdf
38. Opryliudneno rezultaty otsinky chyselnosti naiavnoho naselennia Ukrainy [The results of the assessment of the current population of Ukraine have been published]. (2020). Government portal. Retrieved from https://www.kmu.gov.ua/news/oprilyudneno-rezultati-ocinki-chiselnosti-nayavnogo-naselennya-ukrayini [in Ukrainian].
39. American Community Survey. (2021). Vision and Eye Health Surveillance System (VEHSS). Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/visionhealth/vehss/data/national-surveys/american-community-survey.html#print
40. Data Science for public good. Data Science Campus. Retrieved December 20, 2022 from https://datasciencecampus.ons.gov.uk/
41. Smart Nation Singapore. Smart Nation and Digital Government Office. Retrieved December 20, 2022 from https://www.smartnation.gov.sg/
42. China Population Information and Research Center (CPIRC). www.pop.upenn.edu. Retrieved December 20, 2022 from https://www.pop.upenn.edu/resource/china-population-information-and-research-center-cpirc

Посилання на статтю:
Горобець О. О., Червона С. П. Можливості використання великих даних у рамках статистичного вивчення населення в умовах війни. *Науковий вісник Національної академії статистики, обліку та аудиту: зб. наук. праць.* 2023. № 1-2. С. 5–16. doi: 10.31767/nasoa.1-2-2023.01.

Link to the article:
Horobets, O., Chervona, S. (2023). Mozhlyvosti vykorystannia velykykh danykh u ramkakh statystychnoho vyvchennia naselennia v umovakh viiny [Possibilities of using big data in statistical study of the population in the war's conditions]. Naukovyi visnyk *Natsionalnoi akademii statystyky, obliku ta audytu – Scientific Bulletin of the National Academy of Statistics, Accounting and Audit*, 1-2, 5–16. doi: 10.31767/nasoa.1-2-2023.01.