Wiadomości Statystyczne. The Polish Statistician, 2024, *69*(7), 42–58 https://doi.org/10.59139/ws.2024.07.3
Spisy powszechne – problemy i wyzwania / Issues and challenges in census taking
Zgłoszony/submitted: 22.11.2023, poprawiony/revised: 26.03.2024, zaakceptowany/accepted: 23.06.2024

Udostępnianie danych spisowych w przekroju siatki kilometrowej – wyzwania związane z ochroną tajemnicy statystycznej[1]

Tomasz Klimanek[a], Tomasz Józefowski [b], Andrzej Młodak[c], Amelia Wardzińska-Sharif[d]

Streszczenie. Udostępnianie podstawowych danych pochodzących z Narodowego Spisu Powszechnego Ludności i Mieszkań (NSP) w przekroju siatki kilometrowej należy uznać za jeden z najważniejszych kierunków upowszechniania wyników spisowych, a zarazem spełnienie krajowych i międzynarodowych wymogów w tym zakresie. Z uwagi na to, że komórki siatki są relatywnie małe (kwadraty 1 km×1 km), a ryzyko identyfikacji konkretnej osoby i ujawnienia jej danych wrażliwych jest znaczne, niezbędne staje się podjęcie stosownych działań ochronnych. Celem artykułu jest omówienie najważniejszych działań z zakresu kontroli ujawniania danych na przykładzie danych zebranych podczas NSP 2021 oraz zaproponowanie możliwych do zastosowania w tym przypadku metod i narzędzi kontroli. W szczególności dotyczy to podejść niezakłóceniowych, czyli takich, które prowadzą do ukrywania danych wrażliwych. Ponadto w artykule wskazano najistotniejsze problemy i wyzwania związane z podjęciem tego rodzaju działań ochronnych w zależności od zakresu udostępnianych informacji, ponieważ liczba i rodzaj zmiennych mają kluczowe znaczenie dla ryzyka identyfikacji jednostki i doboru narzędzi ochronnych. Przedstawiono także propozycje rozwiązań metodologicznych i technicznych. Analizy wykazały, że ochrona danych stanowi ważne wyzwanie w rozpatrywanym przypadku, zwłaszcza gdy należy kontrolować ujawnianie danych z kilku powiązanych ze sobą baz danych – wówczas konieczne jest uwzględnienie logicznych i matematycznych powiązań między zbiorami. Dodatkowym czynnikiem ryzyka może się okazać gęstość lub hierarchiczność siatki.

[1] Artykuł został opracowany na podstawie referatu wygłoszonego na Konferencji Naukowej „Statystyka w służbie polityki społeczno-gospodarczej: nowe badania, metody i narzędzia", która odbyła się w dniach 23–24 marca 2023 r. w Warszawie / The article is based on a paper delivered at the Scientific Conference 'Statistics in the service of socio-economic policy: new surveys, methods and tools', held in Warsaw on 23–24 March 2023.

[a] Uniwersytet Ekonomiczny w Poznaniu, Instytut Informatyki i Ekonomii Ilościowej; Urząd Statystyczny w Poznaniu, Polska / Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics; Statistical Office in Poznań, Poland. ORCID: https://orcid.org/0000-0001-6411-8145. E-mail: t.klimanek@stat.gov.pl.

[b] Uniwersytet Ekonomiczny w Poznaniu, Instytut Informatyki i Ekonomii Ilościowej; Urząd Statystyczny w Poznaniu, Polska / Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics; Statistical Office in Poznań, Poland. ORCID: https://orcid.org/0000-0001-9485-1946. E-mail: t.jozefowski@stat.gov.pl.

[c] Uniwersytet Kaliski im. Prezydenta Stanisława Wojciechowskiego, Międzywydziałowy Zakład Matematyki i Statystyki; Urząd Statystyczny w Poznaniu, Polska / University of Kalisz, Inter-faculty Department of Mathematics and Statistics; Statistical Office in Poznań, Poland. ORCID: https://orcid.org/0000-0002-6853-9163. Autor korespondencyjny / Corresponding author, e-mail: a.mlodak@stat.gov.pl.

[d] Główny Urząd Statystyczny, Departament Systemów Teleinformatycznych, Geostatystyki i Spisów, Polska / Statistics Poland, ICT Systems, Geostatistics and Census Department, Poland. ORCID: https://orcid.org/0009-0001-8231-8113. E-mail: a.wardzinska@stat.gov.pl.

Słowa kluczowe: siatka kilometrowa, kontrola ujawniania danych, metody niezakłóceniowe, spis powszechny
JEL: C54, C63, C83

Data dissemination in the kilometre grid: challenges connected to the protection of statistical confidentiality

Abstract. Providing basic data from the National Population and Housing Census in a kilometre grid is one of the most important ways of disseminating census results, which at the same time meets the applicable national and international requirements. Due to the fact that the tiles of the kilometre grid are relatively small (squares with one-kilometre-long sides) and thus the risk of identifying a concrete person and disclosing sensitive information about him or her is significant, it is necessary to employ data-protection procedures. The aim of the paper is to discuss the most important directions in the statistical disclosure control on the example of data collected during the National Population and Housing Census 2021, and to propose methods and tools from the aforementioned realm that would be applicable. These will be mainly non-perturbative approaches, i.e. ones that cause suppression of sensitive information. The paper also brings to light the most important issues and challenges dependent on the scope of information disclosed and related to this type of data-protection procedures, as the number and type of variables determine the risk of the identification of individuals and influence the selection of suitable protection tools. The article sets forth proposals for methodological and technical solutions in the field. The analyses demonstrate that data protection poses a significant challenge in the studied case, especially if several mutually-connected databases are to be protected. In such a situation, it is necessary to take into account the logical and mathematical connections between the data sets. An additional risk factor can also be the density or hierarchical character of the grid.
Keywords: kilometre grid, statistical disclosure control, non-perturbative methods, census

1. Wprowadzenie

Współcześnie można zaobserwować rosnące zapotrzebowanie użytkowników informacji statystycznych na coraz bardziej szczegółowe dane. Owa szczegółowość dotyczy nie tylko większej liczby zmiennych opisujących różne aspekty określonych zjawisk, lecz także coraz niższych poziomów agregacji przestrzennej, na których prezentowane są dane. To ważne zwłaszcza w przypadku spisów powszechnych, które z uwagi na zakres uzyskiwanych danych i liczbę badanych respondentów (a w ostatnich latach także skalę wykorzystania źródeł administracyjnych) dostarczają szczególnie użytecznej wiedzy.

W związku z tym – zgodnie z rekomendacjami międzynarodowymi (United Nations, 2017) – najważniejsze dane spisowe są udostępniane w przekroju określonych siatek przestrzennych. W Polsce stosuje się na ogół siatkę kilometrową (ang. *kilometre grid*) o komórkach w rozmiarze 1 km×1 km (powierzchnia 1 km^2). W obszarze wyznaczonym przez daną komórkę nalicza się agregaty określonych zmiennych kategorialnych dla rekordów odpowiadających jednostkom zlokalizowanym w tym

obrębie. Ułatwia to użytkownikom dokonywanie efektywnych analiz zjawisk społeczno-gospodarczych, w tym zróżnicowania ich natężenia przestrzennego[2].

Przed udostępnieniem danych na tak dużym poziomie szczegółowości należy szczególnie zadbać o właściwą ochronę tajemnicy statystycznej. Zgodnie z podstawową zasadą kontroli ujawniania danych (ang. *statistical disclosure control* – SDC) taka ochrona powinna z jednej strony zapewnić minimalizację ryzyka identyfikacji jednostki, a z drugiej – minimalizację straty informacji będącej wynikiem zastosowania odpowiednich mechanizmów ochrony: niezakłóceniowych, czyli prowadzących do ukrycia niektórych danych, lub zakłóceniowych, czyli zniekształcających w przyjęty sposób pewne informacje w celu ochrony poufności (Hundepool i in., 2012). Spośród wielu tego typu narzędzi (np. Hundepool i in., 2012; Młodak i in., 2023) należy wybrać te, które w danej sytuacji okażą się najefektywniejsze.

Celem artykułu jest omówienie najważniejszych działań z zakresu kontroli ujawniania danych na przykładzie danych zebranych podczas Narodowego Spisu Powszechnego Ludności i Mieszkań (NSP) 2021 oraz zaproponowanie możliwych do zastosowania w tym przypadku metod i narzędzi kontroli.

2. Znaczenie danych w przekroju siatki kilometrowej

Rolą statystyki publicznej jest zbieranie, gromadzenie i analizowanie danych statystycznych, a także opracowywanie wyników badań statystycznych i udostępnianie ich w formie wynikowych informacji statystycznych. Z uwagi na rosnące zapotrzebowanie na dane przestrzenne, przed rundą spisów 2010–2011 Główny Urząd Statystyczny (GUS) wdrożył technologię GIS[3] (Geographical Information System), w ramach której utworzono Przestrzenną Bazę Adresową (PBA). W skład bazy wchodzą następujące warstwy:

- punkty adresowe dla budynków, w których znajduje się co najmniej jedno mieszkanie, wraz z lokalizacją (współrzędnymi);
- wektorowe granice podziału statystycznego, tj. rejonów statystycznych i obwodów spisowych.

Warstwy te są na bieżąco aktualizowane przez pracowników statystyki publicznej. Dzięki wprowadzeniu do danych statystycznych punktów adresowych zmieniono

[2] Warto zauważyć, że siatka kilometrowa jest szczególnie przydatna do przeprowadzania porównań w skali regionalnej, kontynentalnej i globalnej, np. analiz zmian klimatycznych (tereny zalewowe a gęstość zaludnienia) lub odległości od domu do szkoły lub miejsca pracy. Nie będzie natomiast użyteczna w przypadku analiz terytorialnych prowadzonych przez samorządy z uwagi na odrębność tego układu od podziału administracyjnego lub kształtu funkcjonalnych obszarów miejskich.

[3] GIS to system pozyskiwania, gromadzenia, weryfikowania, integrowania, analizowania, transferowania i udostępniania danych przestrzennych. W szerokim rozumieniu obejmuje on metody, środki techniczne (w tym sprzęt i oprogramowanie), bazę danych przestrzennych, organizację, zasoby finansowe i ludzi zainteresowanych jego funkcjonowaniem (System informacji przestrzennej, n.d.).

dotychczasowy system identyfikacji przestrzennej z przyporządkowania obszarowego (obwody spisowe) na przyporządkowanie punktowe. Miało to zasadnicze znaczenie dla zastosowania geomatyki (geoinformatyki) w statystyce publicznej. Zmiana przyporządkowania umożliwiła bardziej elastyczne grupowanie danych statystycznych dla dowolnie małych obszarów. Pozwoliło to także na utworzenie bazy mikrodanych o charakterze przestrzennym, dającej możliwość przeprowadzania analiz geostatystycznych różnych zjawisk dotyczących m.in. demografii (np. średnia odległość od domu do pracy, szkoły lub szpitala).

Odniesienie danych statystycznych do punktu na mapie pozwoliło również na uniezależnienie się w prowadzonych badaniach od zmian w podziale terytorialnym kraju, skutkujących zwykle zmianami obwodów spisowych i wynikającymi stąd pracochłonnymi przeliczeniami. Ułatwiło także analizę porównawczą szeregów czasowych niezależnie od zmian zachodzących w tym podziale (Dygaszewicz, 2012).

Dzięki użyciu technologii GIS w statystyce publicznej GUS może od wielu lat brać udział w międzynarodowych projektach dotyczących wykorzystania informacji geograficznych w statystyce inicjowanych w ramach Europejskiego Systemu Statystycznego (ESS). Podczas pracy przy tych projektach stworzono m.in. metodę przedstawiania danych demograficznych w innym podziale niż dotychczas stosowany podział administracyjny. Wypracowano sposoby prezentacji danych o rozmieszczeniu ludności w podziale ewidencyjnym, statystycznym, a także za pomocą siatki kilometrowej i przeanalizowano te rozwiązania pod kątem jak największej przydatności dla potencjalnego użytkownika. Dzięki temu już dane z NSP 2011 zostały zaprezentowane w Portalu Geostatystycznym GUS[4] w przekroju siatki kilometrowej.

Zgodnie z definicją zawartą w Rozporządzeniu Rady Ministrów z dnia 15 października 2012 r. w sprawie państwowego systemu odniesień przestrzennych siatka kilometrowa to siatka odniesienia dla współrzędnych płaskich prostokątnych, przy czym:

- współrzędne geodezyjne narożników arkuszy map i linie siatki kartograficznej opisuje się w pełnych stopniach, minutach lub sekundach;
- linie siatki kilometrowej opisuje się w metrach lub kilometrach;
- dopuszcza się podawanie tylko punktów przecięcia siatek odniesienia, o których mowa powyżej.

Siatka kilometrowa otrzymuje nazwę od układu współrzędnych, dla którego została obliczona, przy czym:

- początek siatki pokrywa się z początkiem układu współrzędnych płaskich prostokątnych;

[4] https://geo.stat.gov.pl/.

- linie siatki biegną z południa na północ i z zachodu na wschód;
- punktem odniesienia komórki siatki jest lewy dolny narożnik komórki siatki.

Układ odniesienia siatki kilometrowej, w której GUS publikuje obecnie dane z NSP 2011 i 2021, to Europejski Ziemski System Odniesienia 1989 (ETRS89-LAEA). W skład kodu komórki siatki (np. PL_CRS3035RES1000mN3057000E5131000) wchodzą:

- CRS3035 – niepowtarzalny identyfikator przyjętego układu odniesienia (ETRS89-LAEA);
- RES1000 m – rozdzielczość komórki siatki (1000 m);
- N[y] – koduje wartość północną [N], gdzie y to współrzędna w metrach lewego dolnego rogu komórki siatki;
- E[x] – koduje wartość wschodnią [E], gdzie x to współrzędna w metrach lewego dolnego rogu komórki siatki.

Ponadto w Rozporządzeniu (WE) nr 1059/2003 Parlamentu Europejskiego i Rady z dnia 26 maja 2003 r. w sprawie ustalenia wspólnej klasyfikacji Jednostek Terytorialnych do Celów Statystycznych (NUTS) ustanowiono metodykę dla wszystkich członków Unii Europejskiej (UE) dotyczącą określania typologii terytorialnej na podstawie rozmieszczenia ludności w komórkach siatki kilometrowej. Z kolei Rozporządzenie wykonawcze Komisji (UE) 2018/1799 z dnia 21 listopada 2018 r. w sprawie ustanowienia tymczasowego bezpośredniego działania w dziedzinie statystyki dotyczącego rozpowszechniania wybranych tematów spisu ludności i mieszkań w 2021 r. i geokodowanych w siatce kilometrowej (dalej: rozporządzenie 2018/1799) jako tymczasowe bezpośrednie działanie w dziedzinie statystyki towarzyszące spisom powszechnym ludności i mieszkań w 2021 r. określa obowiązek opracowania głównych rezultatów ostatnich spisów powszechnych na podstawie siatki kilometrowej obejmującej całą Europę. Należy również wspomnieć, że Komisja Europejska planuje nałożenie na kraje członkowskie UE obowiązku corocznej publikacji danych demograficznych w przekroju siatek kilometrowych na podstawie danych z rejestrów administracyjnych. Według wstępnych założeń od 2025 r. corocznie mają być publikowane dane dotyczące ludności, a od 2031 r. – bardziej szczegółowe dane dotyczące mieszkań.

GUS opublikował dane demograficzne z NSP 2011 i 2021 w przekroju siatki kilometrowej (m.in. ogólną liczbę ludności, ludność według płci i biologicznych grup wieku oraz współczynnik feminizacji). Można je przeglądać w Portalu Geostatystycznym i pobierać z niego oraz za pośrednictwem usługi INSPIRE. Dzięki wykorzystaniu układu współrzędnych ETRS89-LAEA polskie dane mogą być łączone i porównywane z danymi europejskimi.

Należy również zaznaczyć, że w rozporządzeniu 2018/1799 określono zakres danych w przekroju siatki kilometrowej dostarczanych do Eurostatu przez kraje

członkowskie UE. Spośród nich w niniejszym artykule wzięto pod uwagę następujące zmienne i kategorie:

- ludność ogółem;
- płeć:
  – mężczyźni,
  – kobiety;
- grupy wieku:
  – poniżej 15 lat,
  – od 15 do 64 lat,
  – 65 lat i więcej.

W ramach dalszych prac autorzy planują analizę kolejnych kategorii i zmiennych[5].

Trzeba dodać, że informacji o liczbie osób pracujących dostarcza się, jeśli są dostępne na tym poziomie w kraju członkowskim. Zbiór przepisów dotyczących regulacji, które odnoszą się do rundy spisów z 2021 r. – ze szczególnym uwzględnieniem poziomu (rozmiaru komórek) siatki kilometrowej – zawiera publikacja Eurostatu (2019).

3. Podstawy formalne i rekomendacje międzynarodowe

Spośród rekomendacji Eurostatu w zakresie metod kontroli ujawniania danych zalecanych do stosowania w spisach powszechnych jedną z najważniejszych jest celowana wymiana rekordów (ang. *targeted record swapping* – TRS). To przykład podejścia pretablicowego, czyli metody ochrony stosowanej w odniesieniu do mikrodanych, zwłaszcza takich, na podstawie których powstają tablice wynikowe. Ten rodzaj wymiany rekordów zamienia nierówne wartości zmiennych – najczęściej geograficznych, choć nie tylko – pomiędzy odpowiednio sparowanymi jednostkami. Polega to na tym, że znajduje się rekordy z wysokim ryzykiem ujawnienia i wymienia się w nich określone dane na informacje pochodzące z innego (podobnego) rekordu. Najistotniejsze założenia TRS są następujące:

- wymiana dokonywana jest jedynie między gospodarstwami domowymi (lub innymi podobnie predefiniowanymi grupami jednostek);
- wyznaczane są jednostki o wysokim poziomie ryzyka ujawnienia w odniesieniu do wartości jednej lub większej liczby rozpatrywanych zmiennych;

[5] Będą to: (1) bieżąca aktywność ekonomiczna (osoby pracujące); (2) kraj/miejsce urodzenia (miejsce urodzenia w innym kraju członkowskim UE; miejsce urodzenia gdzie indziej); (3) miejsce zamieszkania na rok przed spisem (bez zmiany miejsca zamieszkania; przemieszczenie na terytorium kraju zgłaszającego; przemieszczenie poza terytorium kraju zgłaszającego).

- za pomocą schematu losowania probabilistycznego z wysokim prawdopodobieństwem wyłania się próbkę gospodarstw domowych (lub innych predefiniowanych grup jednostek) o wysokim stopniu ryzyka ujawnienia;
- każde gospodarstwo domowe (grupa) z tej próbki jest parowane z innym, podobnym w zakresie określonych zmiennych;
- dokonuje się wymiany danych z zakresu zmiennych docelowych (np. geograficznych) pomiędzy sparowanymi gospodarstwami domowymi (grupami).

Jak wspomniano, z uwagi na pierwotne zastosowanie to podejście odnosi się do najpowszechniej stosowanych w spisach grup jednostek, a więc do określonego poziomu podziału terytorialnego (np. podregionów lub powiatów), którego składowe podlegają wymianie w rekordach. Wymiana jest dokonywana w obrębie gospodarstw domowych. W innego rodzaju badaniach to odniesienie może być odmienne (np. w przypadku badań gospodarczych wymiana następuje między grupami Polskiej Klasyfikacji Działalności i klastrami przedsiębiorstw). Celowaną wymianę rekordów można przeprowadzić, używając pakietu `recordSwapping` środowiska R dostępnego na platformie GitHub[6], ostatnio włączonego także jako procedura `recordSwap` do pakietu `sdcMicro` (Gussenbauer, 2023; w kodzie źródłowym wykorzystano skrypty napisane w języku C++). Ocena ryzyka ujawnienia informacji wrażliwych oparta jest w tym przypadku głównie na regule k-anonimowości.

Innym podejściem rekomendowanym przez Eurostat jest metoda kluczy komórkowych (ang. *cell-key method* – CKM). Zalicza się ona do podejść posttablicowych, czyli stosowanych do ochrony już naliczonych tablic wynikowych, a polega na zakłócaniu wartości w komórkach tablicy z wykorzystaniem pewnych liczb losowych (kluczy rekordów – ang. *record keys*) przyporządkowywanych do rekordów, na podstawie których tworzona jest tablica, oraz specjalnej tabeli zakłóceń (p*-table* – p-tabeli). CKM składa się z następujących kroków:

- każdemu rekordowi przyporządkowywana jest losowa liczba naturalna (klucz rekordu) z zakresu od 1 do n, gdzie n jest predefiniowaną maksymalną możliwą wartością klucza;
- tworzona jest tablica docelowa, po czym dla każdej jej komórki sumowane są klucze tworzących ją rekordów modulo n;
- dzięki użyciu predefiniowanej (dobranej optymalnie dla danego przypadku) p-tablicy (zawierającej liczby 1 i –1) otrzymuje się wartość zakłócenia;
- do każdej komórki docelowej tablicy dodawane jest odpowiadające jej zakłócenie z p-tabeli.

[6] https://github.com/sdcTools/recordSwapping.

Metoda kluczy komórkowych może być stosowana z wykorzystaniem pakietu `cellKey` środowiska R. Warto jednak zauważyć, że ten pakiet wciąż jest testowany, dlatego znajduje się na platformie GitHub[7], choć 17 marca 2023 r. umieszczono go również w środowisku R[8]. W Polsce CKM testowano na danych z NSP 2011.

4. Ludność rezydująca według płci i grup wieku w przekroju siatki kilometrowej

Ochrona tajemnicy statystycznej zastosowana w odniesieniu do ludności rezydującej i skutkująca opublikowaniem danych na Portalu Geostatystycznym[9] opierała się na następujących założeniach:

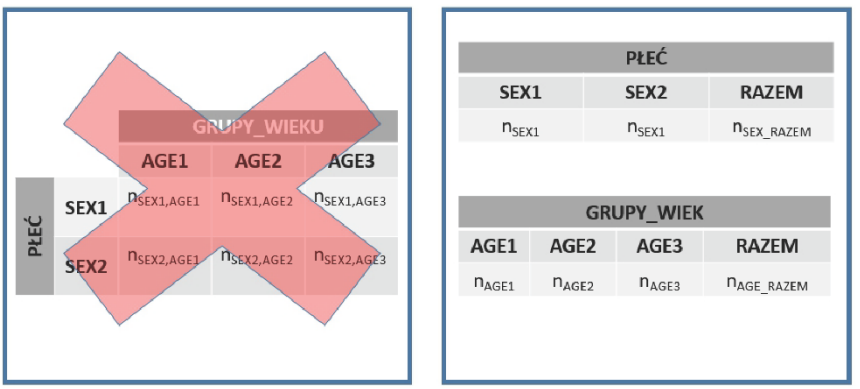
1. Nie należy uwzględniać niezamieszkanych komórek siatki kilometrowej.
2. Informacja na temat ogólnej liczby ludności rezydującej w komórce siatki kilometrowej nie jest objęta tajemnicą statystyczną, a więc jest znana.
3. Zostały udostępnione trzy zmienne na poziomie osób tworzących populację ludności rezydującej:
   - *identyfikator komórki siatki kilometrowej*;
   - *płeć*, przyjmująca warianty:
   – 1 dla mężczyzn (SEX1),
   – 2 dla kobiet (SEX2);
   - *grupa wieku*, przyjmująca warianty:
   – 1 dla osób w wieku poniżej 15 lat (AGE1),
   – 2 dla osób w wieku 15–64 lat (AGE2),
   – 3 dla osób w wieku 65 lat i więcej (AGE3).
4. Ochrona tajemnicy statystycznej za pomocą pierwotnego i wtórnego ukrywania komórek (ang. *primary/secondary supression*) realizowana jest z zastosowaniem zasad:
   - k-anonimowości, gdzie przyjęto $k = 3$;
   - l-różnorodności, gdzie przyjęto $l = 2$.
5. Publikowane są jedynie rozkłady brzegowe zmiennych płeć i grupa wieku w komórkach siatki kilometrowej (schemat).

[7] https://github.com/sdcTools/cellKey.

[8] https://cran.r-project.org/web/packages/cellKey/index.html.

[9] Zgodnie z harmonogramem podanym na stronie: https://stat.gov.pl/spisy-powszechne/nsp-2021/harmonogram-publikacji-wynikow-nsp-2021/.

Schemat. Rozkłady publikowane w Portalu Geostatystycznym



Źródło: opracowanie własne.

Należy dodać, że w przypadku gdy dwa warianty grup wieku były uznane za bezpieczne, czyli o niskim ryzyku ujawnienia ($n \geq 3$), a trzeci wariant był niebezpieczny ($n < 3$) i jednocześnie warianty bezpieczne były równoliczne, to obok oczywistego ukrycia wariantu niebezpiecznego należało zastosować odpowiedni mechanizm ukrywania jednego z wariantów bezpiecznych. Rozważano podejścia oparte na:

- konsekwentnym ukrywaniu określonego przekroju:

$$P(p_2 = \#|p_1 = u \wedge p_3 = s) = 1, \tag{1}$$

$$P(p_3 = \#|p_2 = u \wedge p_1 = s) = 1, \tag{2}$$

$$P(p_1 = \#|p_3 = u \wedge p_2 = s) = 1; \tag{3}$$

- losowym ukrywaniu określonego przekroju (opartym na rozkładzie równomiernym):

$$P(p_2 = \#|p_1 = u) = P(p_3 = \#|p_1 = u) = 0{,}5, \tag{4}$$

$$P(p_1 = \#|p_2 = u) = P(p_3 = \#|p_2 = u) = 0{,}5, \tag{5}$$

$$P(p_1 = \#|p_3 = u) = P(p_2 = \#|p_3 = u) = 0{,}5; \tag{6}$$

- losowym ukrywaniu określonego przekroju (opartym na rozkładzie grup wieku w populacji):

$$P(p_1 = \#|p_2 = u) = 1 - \frac{n_{p_1}}{n_{p_1} + n_{p_3}}$$
$$P(p_3 = \#|p_2 = u) = 1 - \frac{n_{p_3}}{n_{p_1} + n_{p_3}}, \quad (7)$$

$$P(p_1 = \#|p_3 = u) = 1 - \frac{n_{p_1}}{n_{p_1} + n_{p_2}}$$
$$P(p_2 = \#|p_3 = u) = 1 - \frac{n_{p_2}}{n_{p_1} + n_{p_2}}, \quad (8)$$

$$P(p_2 = \#|p_1 = u) = 1 - \frac{n_{p_2}}{n_{p_2} + n_{p_3}}$$
$$P(p_3 = \#|p_1 = u) = 1 - \frac{n_{p_3}}{n_{p_2} + n_{p_3}}, \quad (9)$$

gdzie:
p_1, p_2, p_3 – grupy wieku odpowiednio: do 15 lat, 15–64 lata oraz 65 lat i więcej,
– przekrój z zastosowaną ochroną tajemnicy statystycznej,
s – przekrój bezpieczny, tj. $n \geq 3$,
u – przekrój niebezpieczny, tj. $n < 3$.

Ostatecznie zastosowano wariant oparty na rozkładzie częstości grup wieku w populacji. Podobne podejście przyjęto w przypadku, gdy liczebność dwóch wariantów grup wieku była równa 0, a trzeci był niebezpieczny ($n < 3$).

Analizę rozkładu komórek siatki w podziale na przekroje potencjalnie bezpieczne ($n \geq 3$) i potencjalnie niebezpieczne ($n < 3$) przedstawiono w tabl. 1.

Tabl. 1. Rozkład liczby rezydentów w komórkach siatki kilometrowej

Liczba rezydentów w komórkach siatki	Komórki siatki		Rezydenci		Kategoria komórek
	liczba	udział w ogólnej liczbie w %	liczba	udział w ogólnej liczbie w %	
1	3 424	1,74	3 424	0,01	potencjalnie niebezpieczne
2	4 213	2,14	8 426	0,02	potencjalnie niebezpieczne
3 i więcej	189 493	96,13	37 007 477	99,97	potencjalnie bezpieczne
Ogółem	197 130	100,00	37 019 327	100,00	.

Źródło: opracowanie własne na podstawie bazy danych z NSP 2021.

W wyniku prac przeprowadzonych z wykorzystaniem oprogramowania SAS dokonano pierwotnego albo pierwotnego i wtórnego ukrycia przekrojów niebezpiecznych, co skutkowało pewną utratą informacji. Należy jednak podkreślić, że w przypadku zmiennej:

- *płeć* bezwzględna strata informacji obliczona jako różnica między liczbą ludności rezydującej w oryginalnym zbiorze danych a liczbą ludności rezydującej opublikowaną na Portalu Geostatystycznym wyniosła 91 719 osób, tj. 0,25%;
- *grupa wieku* bezwzględna strata informacji obliczona jako różnica między liczbą ludności rezydującej w oryginalnym zbiorze danych a liczbą ludności rezydującej opublikowaną na Portalu Geostatystycznym wyniosła 261 560 osób, tj. 0,71%.

Analiza przestrzennego rozkładu zastosowanych technik ochrony tajemnicy statystycznej pokazuje, że ukrycia komórek siatki kilometrowej występują na terenie całego kraju, choć w nieco większym stopniu w północno-wschodniej Polsce (mapa, s. 54). Na submapie A przedstawiono jednoczesne osobne ukrycia poszczególnych kategorii płci i grup wieku, na submapie B – jednoczesne ukrycia agregacji krzyżowych.

5. Wyzwania związane z udostępnianiem danych o ludności w przekroju siatki kilometrowej

Opisane wyżej wyniki wskazują na złożoność problemu, z jakim przyszło się zmierzyć statystykom. Już w przypadku czynności przedstawionych we wcześniejszej części artykułu można było zauważyć potrzebę zastosowania metod probabilistycznych w ukrywaniu wtórnym. Dotyczyło to zwłaszcza sytuacji, w których typowe metody okazały się niewystarczające. Co ważne, w przypadku większej liczby zmiennych przewidzianych do udostępnienia (zob. cz. 2) kombinacje wartości podlegające ochronie mogą przyjmować coraz bardziej złożoną postać. W takiej sytuacji konieczne staje się zastosowanie bardziej zaawansowanych narzędzi SDC.

Z uwagi na rozmiar zbioru (ponad 197 tys. komórek siatki kilometrowej), charakter zawartych w nim informacji i potencjalne wykorzystywanie udostępnianych danych możliwe są dwa zasadnicze podejścia:

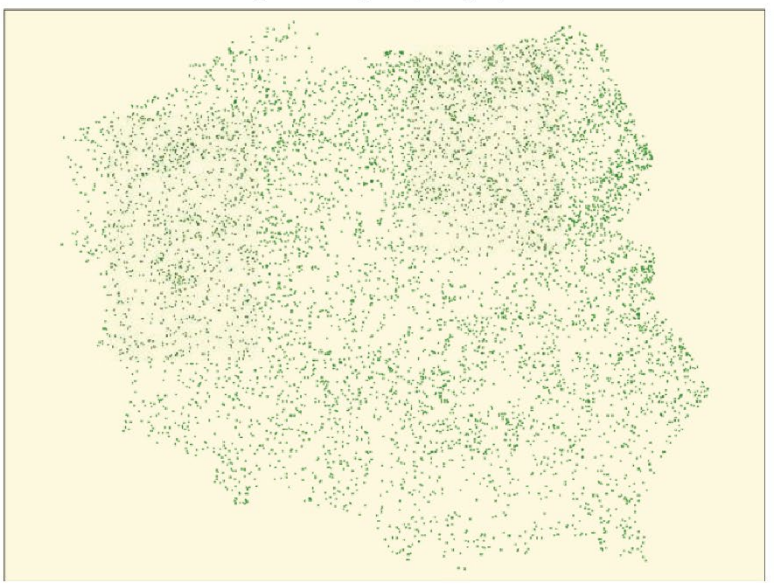
- potraktowanie zbioru jako bazy mikrodanych i zastosowanie w odniesieniu do niego odpowiednich metod SDC (np. celowana wymiana rekordów w połączeniu z mikroagregacją lub zaokrąglaniem[10]);

[10] Jak pokazują doświadczenia z innych badań – opisane m.in. w pracach Młodaka i in. (2022) oraz Pietrzaka i in. (2022) – z uwagi na ograniczone spektrum odniesienia zastosowanie TRS może się okazać niewystarczające.

Mapa. Rozkłady przestrzenne jednoczesnych ukryć

A. Wszystkie kategorie płci lub grup wieku

B. Wszystkie kategorie płci i grup wieku



Źródło: opracowanie własne na podstawie bazy danych z NSP 2021.

- przyjęcie, że zbiór jest tablicą wielowymiarową. Można wówczas wykorzystać narzędzia SDC, stosowane głównie w przypadku tablic częstości, przede wszystkim metodę kluczy komórkowych. Alternatywnie warto w tym kontekście rozważyć metodę kontrolowanego dopasowania tablicy (ang. *controlled tabular adjustment* – CTA) lub podejście hiperkostkowe (ang. *hypercube*), oparte na heurystycznej ochronie całościowej polegającej na ochronie podtablicy i bazującej na znajdowaniu i ukrywaniu struktur geometrycznych (kostek n-wymiarowych), co jest niezbędne do ochrony komórek wrażliwych wskazanych podczas ukrywania pierwotnego (Hundepool i in., 2012; Młodak i in., 2023).

Trzeba się liczyć z możliwością ujawnienia wrażliwego atrybutu osoby w wyniku połączenia dwóch różnych zbiorów. Z uwagi na to, że oprócz danych dotyczących ludności rezydującej udostępniane są także informacje dotyczące ludności według definicji krajowej (dla tych samych zmiennych), powiązanie obu zbiorów może pozwolić na uzyskanie dodatkowych wrażliwych informacji o osobie. Należy przypomnieć, że według definicji krajowej ludność naszego kraju tworzą stali mieszkańcy Polski, w tym osoby, które przebywają czasowo za granicą (bez względu na okres przebywania), ale zachowały stałe zameldowanie w Polsce. W tym ujęciu do ludności nie są natomiast zaliczani imigranci przebywający w Polsce czasowo. Z kolei ludność rezydująca to ogół osób mieszkających/przebywających na danym terenie przez co najmniej 12 miesięcy. Do rezydentów danej jednostki administracyjnej zalicza się wszystkie osoby mieszkające lub zamierzające mieszkać w tej jednostce nie krócej niż rok. Oznacza to, że w liczbie rezydentów danej jednostki administracyjnej ujęci są:

- stali mieszkańcy (osoby zameldowane lub stale mieszkające bez zameldowania), z wyjątkiem tych, którzy wyjechali na co najmniej 12 miesięcy do innego miejsca w kraju lub za granicę;
- osoby przybyłe z innego miejsca w kraju lub z zagranicy (imigranci bez karty stałego pobytu) na co najmniej 12 miesięcy.

Gdy dla danej zmiennej różnica między odpowiednimi agregatami w obu typach ludności jest niewielka, może się to przyczynić do identyfikacji osoby i ujawnienia jej danych wrażliwych. Oto przykład: jeśli w tym samym agregacie liczba ludności rezydującej jest o 1 większa od liczby ludności według definicji krajowej, to w konkretnych uwarunkowaniach lokalnych (np. brak osób, które wyjechały za granicę) może to oznaczać, że ową jedną osobą był imigrant przebywający w Polsce czasowo, ale dłużej niż 12 miesięcy. W konsekwencji można odczytać z danych jego płeć i wiek. Podczas przeprowadzania procesu SDC należy zatem uwzględnić i takie zagrożenie.

W tabl. 2 zaprezentowano możliwą skalę problemu przy założeniu, że zagrożenie dla poufności danych pojawia się przede wszystkim wtedy, gdy wartość bezwzględna wspomnianej różnicy jest większa od 0, ale mniejsza od 3.

Tabl. 2. Różnice między liczbą ludności rezydującej a liczbą ludności według definicji krajowej dla zmiennych prezentowanych w przekroju siatki kilometrowej

Zmienne	Różnice							
	–2		–1		1		2	
		w %		w %		w %		w %
SEX1	15200	7,71	28654	14,54	7599	3,86	1568	0,80
SEX2	15347	7,79	26964	13,68	7901	4,01	1609	0,82
AGE1	10523	5,34	17699	8,98	2809	1,43	1012	0,51
AGE2	16063	8,15	27250	13,83	8772	4,45	2028	1,03
AGE3	2614	1,33	8607	4,37	3909	1,98	397	0,20

Źródło: opracowanie własne na podstawie bazy danych z NSP 2021.

Jak widać, liczba i odsetek problemowych różnic są dość duże – dotyczy to zwłaszcza opisanej wcześniej sytuacji, gdy liczba ludności według definicji krajowej jest o 1 większa od liczby ludności rezydującej. Problem wydaje się wprawdzie w naj-większym stopniu dotyczyć mniejszych zbiorów, ale w określonych okolicznościach może okazać się istotny także dla tych większych.

Należy również pamiętać, że w przypadku ochrony zbioru o większej liczbie zmiennych – która wymaga wprowadzenia szerszego zakresu ukryć lub zniekształceń – wzrośnie też strata informacji spowodowana zastosowaniem SDC. Tymczasem osiągnięcie balansu między minimalizacją ryzyka ujawnienia a minimalizacją straty informacji (co oznacza maksymalizację użyteczności danych dla użytkownika) to kluczowy cel tego procesu. Istotne wyzwanie stanowi zatem wypracowanie kompromisu między tymi dwoma dążeniami. Zastosowanie rekomendowanych przez Eurostat metod TRS i CKM, względnie podejść dodatkowych lub alternatywnych, będzie wymagało dostępu do zbioru danych jednostkowych w ostatecznej postaci i w pełni funkcjonalnego środowiska operacyjnego obsługującego specjalistyczne pakiety programu R.

6. Podsumowanie

Skuteczna ochrona danych udostępnianych w przekroju siatki kilometrowej wymaga zaawansowanych technik wykorzystujących narzędzia kontroli ujawniania danych. Ochrona pojedynczego zasobu nie wiąże się z zastosowaniem szczególnie złożonych rozwiązań. Większe wyzwanie pojawia się w sytuacji, gdy należy ochronić kilka powiązanych ze sobą baz danych. Wówczas ukrywanie danych w każdym pliku niezależnie może nie wystarczyć, jako że występowanie logicznych i matematycznych powiązań między zbiorami – zwłaszcza gdy zmienne w obu zbiorach bazują na jednolitych fundamentach definicyjnych – stwarza ryzyko wtórnego odtworzenia danych wrażliwych, a tym samym identyfikacji jednostki. W takiej sytuacji konieczne

okazują się dodatkowe ukrycia dokonane na podstawie analizy możliwości reidentyfikacji jednostki oraz maksymalnego możliwego zachowania użyteczności danych. Być może niezbędne w tej sytuacji będzie np. użycie wyższych poziomów agregacji.

Inną kwestią pozostaje gęstość samej siatki. Obecnie najczęściej stosuje się komórki o boku 1 km×1 km. W wielu krajach w użyciu bywają jednak siatki gęstsze – o komórkach 500 m×500 m lub 200 m×200 m. Można też używać siatek o większych komórkach (np. 2 km×2 km lub 5 km×5 km). Ponadto metodologiczne podstawy konstrukcji takich siatek nie zawsze są spójne. Dodatkowo mogą one tworzyć strukturę hierarchiczną i wówczas mamy do czynienia z siatką wielopoziomową. W takim przypadku należy rozważyć zastosowanie podejścia zaproponowanego przez Branchu i in. (2018). Chodzi o metodę zakłócającą wartości zagregowane w sposób deterministyczny, która opiera się na dwóch krokach. W pierwszym kroku należy utworzyć podstawowe grupy komórek siatki w następujący sposób:

- połączyć komórki zawierające dane obarczone ryzykiem ujawnienia z innymi komórkami, tak aby utworzyć grupy komórek (mogą to też być komórki wyższego poziomu), w których dane pozostaną bezpieczne w myśl przyjętej reguły ryzyka ujawnienia (np. k-anonimowości);
- jeśli to możliwe, zastosować kilka siatek dla tego samego obszaru, co pomaga ograniczyć poszukiwanie najlepszych komórek do połączenia w określonym obszarze (na podstawie odpowiednio zdefiniowanej odległości, ale przy mniejszym nakładzie obliczeniowym);
- algorytm rozpoczyna się od poziomu najbardziej zgrubnego (czyli siatki, w której wszystkie komórki są bezpieczne, a zatem ryzyko ujawnienia jest małe), po czym stopniowo zniża się do poziomu najdokładniejszego.

W drugim kroku należy proporcjonalnie nałożyć zakłócenia. Oznacza to, że każda liczba na poziomie grupy otrzymanej w pierwszym kroku (np. liczba kobiet w wieku 65 lat i więcej) rozkłada się na komórki grupy proporcjonalnie do liczby ludności w każdej komórce siatki – wówczas ochrona jest silna. Ryzyko identyfikacji jednostki pojawia się jedynie wówczas, gdy rozkład danego atrybutu jest identyczny we wszystkich komórkach grupy (to bardzo rzadki przypadek). W takiej sytuacji komórki te należy złączyć z komórkami najbliższej grupy i poddać stosownemu zakłóceniu.

Zaletą prezentowanej metody (stosowanej np. we francuskim Institut national de la statistique et des études économiques – pol. Narodowy Instytut Statystyki i Badań Ekonomicznych) jest zachowanie addytywności w komórkach siatki. Ponadto zakłócenia są monitorowane geograficznie, jako że w wielu przypadkach odległość pomiędzy komórkami siatki kilometrowej nie jest zbyt duża. Dla zminimalizowania ryzyka ujawnienia informacji chronionych dokładny skład grup nie jest zazwyczaj ujawniany.

Analiza możliwości zastosowania w razie potrzeby tego rodzaju podejść oraz ich powiązania z omówionymi w artykule stanowi wyzwanie dla statystyki na najbliższe lata. Co ważne, w każdym przypadku konieczna jest ocena ryzyka ujawnienia i oczekiwanej straty informacji na skutek zastosowania opisanych mechanizmów.

Podziękowania

Autorzy składają podziękowania pani Dorocie Szałtys z Departamentu Badań Demograficznych GUS za cenne uwagi zgłoszone na etapie przygotowywania artykułu.

Bibliografia

Branchu, M., Costemalle, V., & Fontaine, M. (2018). Données carroyées et confidentialité. *Journées de Méthodologie Statistiques de l'Insee*, *13*, 1–23. http://www.jms-insee.fr/2018/S23_3_ACTE_BRANCHU_JMS2018.pdf.

Dygaszewicz, J. (2012). Spisy powszechne jako źrodło danych do analiz geoprzestrzennych. *Archiwum Fotogrametrii, Kartografii i Teledetekcji*, *23*, 91–100. https://ptfit.sgp.geodezja.org.pl/wydawnictwa/archiwum_vol_23/Dygaszewicz.pdf.

Eurostat. (2019). *EU legislation on the 2021 population and housing censuses. Explanatory notes*. Publications Office of the European Union. https://ec.europa.eu/eurostat/docments/3859598/9670557/KS-GQ-18-010-EN-N.pdf/c3df7fcb-f134-4398-94c8-4be0b7ec0494?t=1552653277000.

Gussenbauer, J. (2023, 30 sierpnia). *Target Record Swapping*. https://cran.r-project.org/web/packages/sdcMicro/vignettes/recordSwapping.html.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & de Wolf, P.-P. (2012). *Statistical Disclosure Control*. John Wiley & Sons. https://doi.org/10.1002/9781118348239.

Młodak, A., Pietrzak, M., & Józefowski, T. (2022). The trade-off between the risk of disclosure and data utility in SDC: A case of data from a survey of accidents at work. *Statistical Journal of the IAOS*, *38*(4), 1503–1511. https://doi.org/10.3233/SJI-220936.

Młodak, A., Pietrzak, M., Klimanek, T., Józefowski, T., & Lańduch, P. (2023). *Poufność a użyteczność informacji statystycznych. Dylematy ochrony udostępnianych danych*. Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu. https://doi.org/10.18559/978-83-8211-168-2.

Pietrzak, M., Józefowski, T., Klimanek, T., & Młodak, A. (2022). An Optimized Selection of Statistical Disclosure Control Methods – A Case Study Involving Microdata from the Polish Survey of Accidents at Work. W: K. Jajuga, G. Dehnel, M. Walesiak (Eds.), *Modern Classification and Data Analysis: Methodology and Applications to Micro- and Macroeconomic Problems* (pp. 63–78). Springer. https://doi.org/10.1007/978-3-031-10190-8_6.

Rozporządzenie Rady Ministrów z dnia 15 października 2012 r. w sprawie państwowego systemu odniesień przestrzennych (Dz.U. 2012 poz. 1247 ze zm.).

Rozporządzenie (WE) nr 1059/2003 Parlamentu Europejskiego i Rady z dnia 26 maja 2003 r. w sprawie ustalenia wspólnej klasyfikacji Jednostek Terytorialnych do Celów Statystycznych (NUTS) (Dz.U. UE L 154/1).

Rozporządzenie wykonawcze Komisji (UE) 2018/1799 z dnia 21 listopada 2018 r. w sprawie ustanowienia tymczasowego bezpośredniego działania w dziedzinie statystyki dotyczącego rozpowszechniania wybranych tematów spisu ludności i mieszkań w 2021 r. i geokodowanych w siatce kilometrowej (Dz.U. L 296 z 22.11.2018).

System informacji przestrzennej. (n.d.). W: *Internetowy leksykon geomatyczny*. https://www.ptip.info/LMB/system-informacji-przestrzennej.

United Nations. (2017). *Principles and Recommendations for Population and Housing Censuses: the 2020 Round* (Revision 3). https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Principles_and_Recommendations/Population-and-Housing-Censuses/Series_M67rev3-E.pdf.