Urząd Statystyczny
w Poznaniu

Ochrona tajemnicy statystycznej w siatce kilometrowej – metody i wyzwania dla danych spisowych

Tomasz Klimanek[1,2], Tomasz Józefowski[1,2], Andrzej Młodak[1,3]

[1] Urząd Statystyczny w Poznaniu
[2] Uniwersytet Ekonomiczny w Poznaniu
[3] Akademia Kaliska im. Prezydenta Stanisława Wojciechowskiego

Warszawa, 23-24 marca 2023 r.

Plan prezentacji

1. ROZPORZĄDZENIE WYKONAWCZE KOMISJI (UE) 2018/1799 z dnia 21 listopada 2018 r.
2. Metody rekomendowane przez Eurostat w spisach
3. Ludność rezydująca według płci i grup wieku w siatce kilometrowej
4. Wyzwania związane z udostępnianiem w siatce kilometrowej ludności rezydującej i krajowej według płci i grup wieku
5. Podsumowanie

ROZPORZĄDZENIE WYKONAWCZE KOMISJI (UE) 2018/1799 z dnia 21 listopada 2018 r.

- Kategorie tematów spisu, które mają być przedstawione w podziale według kilometrowej siatki odniesienia

Dane przekazywane	**Dane planowane do przekazania**
SEX.0.: Ludność ogółem	CAS.L.1.1.: Osoby pracujące
SEX.1.: Mężczyzna	POB.L.1.: Miejsce urodzenia w kraju zgłaszającym
SEX.2.: Kobieta	POB.L.2.1.: Miejsce urodzenia w innym państwie członkowskim UE
AGE.G.1.: Poniżej 15 lat	POB.L.2.2.: Miejsce urodzenia gdzie indziej
AGE.G.2.: Od 15 do 64 lat	ROY.1.: Bez zmiany miejsca zamieszkania na rok przed spisem
AGE.G.3.: 65 lat i więcej	ROY.2.1.: Miejsce zamieszkania na rok przed spisem: przemieszczenie na terytorium kraju zgłaszającego
	ROY.2.2.: Miejsce zamieszkania na rok przed spisem: przemieszczenie spoza terytorium kraju zgłaszającego

Metody rekomendowane przez Eurostat w spisach

- Targeted Record Swapping (TRS) – celowana wymiana rekordów
  - metoda pretablicowa, rodzaj wymiany rekordów, która zamienia pewne nierówne wartości zmiennych – najczęściej geograficznych, choć nie tylko – pomiędzy odpowiednio sparowanymi jednostkami,
  - w metodzie TRS znajduje się rekordy z wysokim ryzykiem ujawnienia i wymienia się w nich określone dane na dane pochodzące z innego, podobnego do danego, rekordu,
  - najistotniejsze założenia:
    - wymiana dokonywana jest jedynie między gospodarstwami domowymi (lub innymi podobnie predefiniowanymi grupami jednostek),
    - na każdym poziomi wymienianych zmiennych wyznaczane są jednostki o wysokim poziomie ryzyka ujawnienia,
    - za pomocą losowania probabilistycznego z wysokim prawdopodobieństwem losuje się próbkę gospodarstw domowych (lub innych predefiniowanych grup jednostek) o wysokim poziomie ryzyka ujawnienia.

Metody rekomendowane przez Eurostat w spisach

- Targeted Record Swapping (TRS) – celowana wymiana rekordów
  - najistotniejsze założenia:
    - każde gospodarstwo domowe (grupa) z tej próbki jest parowane z innym, podobnym w kontekście określonych zmiennych,
    - dokonuje się wymiany danych z zakresu zmiennych docelowych (np. geograficznych) pomiędzy sparowanymi gospodarstwami domowymi (grupami).
  - TRS może być wykonana przy użyciu pakietu recordSwapping środowiska R, dostępnego na platformie GitHub (ostatnio także włączonego jako procedura do pakietu sdcMicro) – https://github.com/sdcTools/recordSwapping, w kodzie źródłowym wykorzystano skrypty napisane w języku C++.
  - ocena ryzyka oparta jest głównie na regule *k*–anonimowości.

Metody rekomendowane przez Eurostat w spisach

- Cell-key method (CKM) – metoda kluczy komórkowych
  - post–tablicowa metoda zakłócania wartości w komórkach tablicy z wykorzystaniem pewnych liczb losowych (kluczy rekordów – ang. *record keys*) przyporządkowywanych do rekordów, w oparciu o które tworzona jest tablica oraz specjalnej tabeli zakłóceń (*p*–table – *p*–tabeli)
  - kroki:
    - do każdego rekordu przyporządkowujemy liczbę losową (klucz rekordu) będącą liczbą naturalną z zakresu od 1 do n, gdzie n jest predefiniowaną maksymalną możliwą wartością klucza,
    - tworzymy tablicę docelową i, dla każdej jej komórki, sumujemy klucze tworzących ją rekordów modulo n,
    - używamy predefiniowaną (optymalnie dobraną w danym przypadku) *p*–tablicę (zawierającą liczby 1 i -1) aby otrzymać wartość zakłócenia,
    - do każdej komórki docelowej tablicy dodajemy odpowiadające jej zakłócenie z *p*–tabeli.

Metody rekomendowane przez Eurostat w spisach

- Cell-key method (CKM) – metoda kluczy komórkowych
  - może być stosowana przy użyciu pakietu cellKey środowiska R,
  - pakiet cellKey jest ciągle testowany, dlatego też znajduje się jedynie na platformie GitHub (https://github.com/sdcTools/cellKey),
  - testowano CKM na danych z Narodowego Spisu Powszechnego Ludności i Mieszkań 2011.

Ludność rezydująca według płci i grup wieku w siatce kilometrowej

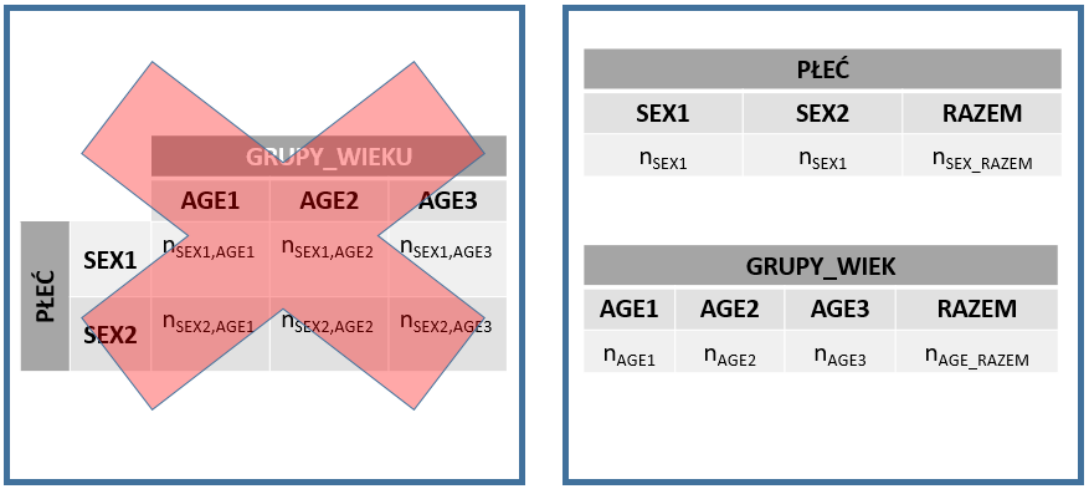
Przyjęte założenia:

- Nie są rozpatrywane niezamieszkałe oczka siatki na terytorium kraju.
- Znana jest ogólna liczba ludności rezydującej dla danego oczka siatki (Rozp. KE 2018/1799 z 21 listopada 2018 r.).
- Udostępnione są trzy zmienne na poziomie osób tworzących populację ludności rezydującej:
  - identyfikator oczka siatki kilometrowej,
  - płeć, przyjmująca warianty:
    - ***1*** dla mężczyzn,
    - ***2*** dla kobiet.
  - grupa wieku, przyjmująca warianty:
    - ***1*** dla osób w wieku poniżej 15 lat,
    - ***2*** dla osób w wieku 15-64 lata
    - ***3*** dla osób w wieku 65 lat i więcej.

Ludność rezydująca według płci i grup wieku w siatce kilometrowej

Przyjęte założenia - cd.:

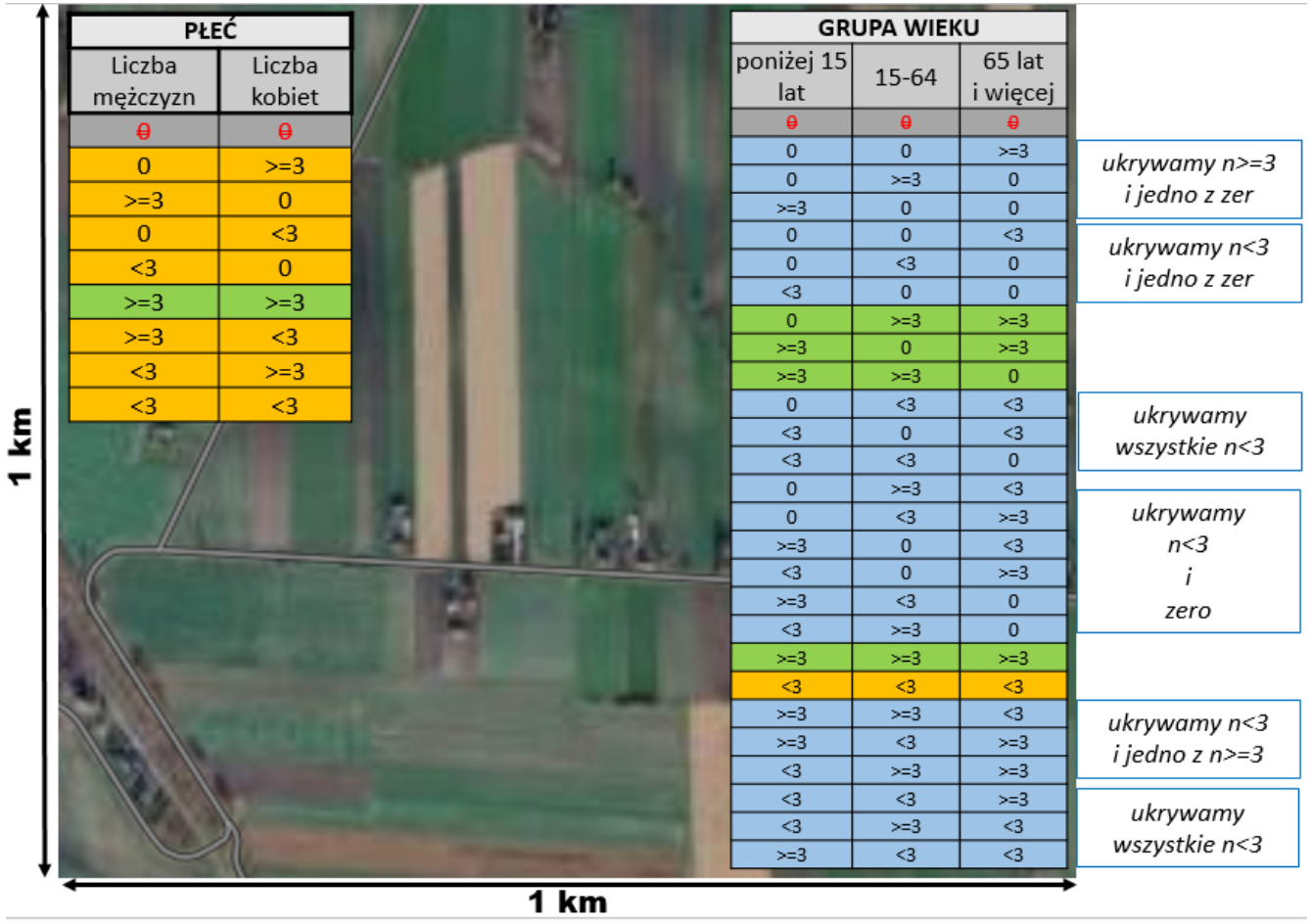
- Publikowane są jedynie rozkłady brzegowe zmiennych płeć i grupa wieku w oczkach siatki kilometrowej.



Rys. 1: Rozkłady publikowane w Portalu Geostatystycznym

- ochrona tajemnicy statystycznej za pomocą pierwotnego i wtórnego ukrywania komórek (*ang. primary/secondary supression*) z zastosowaniem zasad:
  - ***k-anonimowości*** $k = 3$
  - ***l-różnorodności*** $l = 2$

Ludność rezydująca według płci i grup wieku w siatce kilometrowej



PŁEĆ	
Liczba mężczyzn	Liczba kobiet
~~0~~	~~0~~
0	>=3
>=3	0
0	<3
<3	0
>=3	>=3
>=3	<3
<3	>=3
<3	<3

GRUPA WIEKU			
poniżej 15 lat	15-64	65 lat i więcej	
~~0~~	~~0~~	~~0~~	
0	0	>=3	*ukrywamy n>=3 i jedno z zer*
0	>=3	0	
>=3	0	0	
0	0	<3	*ukrywamy n<3 i jedno z zer*
0	<3	0	
<3	0	0	
0	>=3	>=3	
>=3	0	>=3	
>=3	>=3	0	
0	<3	<3	*ukrywamy wszystkie n<3*
<3	0	<3	
<3	<3	0	
0	>=3	<3	*ukrywamy n<3 i zero*
0	<3	>=3	
>=3	0	<3	
<3	0	>=3	
>=3	<3	0	
<3	>=3	0	
>=3	>=3	>=3	
<3	<3	<3	
>=3	>=3	<3	*ukrywamy n<3 i jedno z n>=3*
>=3	<3	>=3	
<3	>=3	>=3	
<3	<3	>=3	*ukrywamy wszystkie n<3*
<3	>=3	<3	
>=3	<3	<3	

Rys. 2: Schematy pierwotnego/wtórnego ukrywania kategorii płci i grup wieku w oczku siatki kilometrowej

Ludność rezydująca według płci i grup wieku w siatce kilometrowej

W przypadku kiedy dwa warianty grup wieku były bezpieczne ($n >= 3$) a trzeci był niebezpieczny ($n < 3$) i jednocześnie warianty bezpieczne były równoliczne należało zastosować odpowiedni mechanizm ukrywania jednego z wariantów bezpiecznych. Rozważano podejścia oparte na:

- konsekwentnym ukrywaniu określonego przekroju,

$$n_{p_1} < 3 \wedge n_{p_3} >= 3 \implies n_{p_2} = \text{"\#"} \quad (1)$$

$$n_{p_2} < 3 \wedge n_{p_1} >= 3 \implies n_{p_3} = \text{"\#"} \quad (2)$$

$$n_{p_3} < 3 \wedge n_{p_2} >= 3 \implies n_{p_1} = \text{"\#"} \quad (3)$$

gdzie:

- n_{p_1}, n_{p_2}, n_{p_3} to liczebności grup wieku odpowiednio do 15 lat, 15-64 lat oraz 65 lat i więcej,
- "#" oznacza przekrój z ochroną tajemnicy statystycznej.

Ludność rezydująca według płci i grup wieku w siatce kilometrowej

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

$$P(n_{p_2} = \text{"\#"} \mid n_{p_1} < 3) = P(n_{p_3} = \text{"\#"} \mid n_{p_1} < 3) = 0.5 \quad (4)$$

$$P(n_{p_1} = \text{"\#"} \mid n_{p_2} < 3) = P(n_{p_3} = \text{"\#"} \mid n_{p_2} < 3) = 0.5 \quad (5)$$

$$P(n_{p_1} = \text{"\#"} \mid n_{p_3} < 3) = P(n_{p} = \text{"\#"} \mid n_{p_3} < 3) = 0.5 \quad (6)$$

gdzie:

- n_{p_1}, n_{p_2}, n_{p_3} to liczebności grup wieku odpowiednio do 15 lat, 15-64 lat oraz 65 lat i więcej,
- $\#$ oznacza przekrój z ochroną tajemnicy statystycznej.

Ludność rezydująca według płci i grup wieku w siatce kilometrowej

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

$$P(n_{p_1} = \text{"\#"} \mid n_{p_2} < 3) = 1 - \frac{n_{p_1}}{n_{p_1} + n_{p_3}} \quad \text{i} \quad P(n_{p_3} = \text{"\#"} \mid n_{p_2} < 3) = 1 - \frac{n_{p_3}}{n_{p_1} + n_{p_3}} \tag{7}$$

$$P(n_{p_1} = \text{"\#"} \mid n_{p_3} < 3) = 1 - \frac{n_{p_1}}{n_{p_1} + n_{p_2}} \quad \text{i} \quad P(n_{p_2} = \text{"\#"} \mid n_{p_3} < 3) = 1 - \frac{n_{p_2}}{n_{p_1} + n_{p_2}} \tag{8}$$

$$P(n_{p_2} = \text{"\#"} \mid n_{p_1} < 3) = 1 - \frac{n_{p_2}}{n_{p_2} + n_{p_3}} \quad \text{i} \quad P(n_{p_3} = \text{"\#"} \mid n_{p_1} < 3) = 1 - \frac{n_{p_3}}{n_{p_2} + n_{p_3}} \tag{9}$$

gdzie:

- n_{p_1}, n_{p_2}, n_{p_3} to liczebności grup wieku odpowiednio do 15 lat, 15-64 lata oraz 65 lat i więcej.
- $\#$ oznacza przekrój z ochroną tajemnicy statystycznej.

Ludność rezydująca według płci i grup wieku w siatce kilometrowej

Ostatecznie zastosowano wariant oparty na rozkładzie częstości grup wieku w populacji.

Podobne podejście zastosowano w przypadku kiedy liczebności dwóch wariantów grup wieku były równe zero a trzeci był niebezpieczny ($n < 3$).

Ludność rezydująca według płci i grup wieku w siatce kilometrowej

Wybrane wyniki

Liczba rezydentów w gridach	Liczba gridów	Odsetek gridów (%)	Łączna liczba rezydentów	Odsetek łącznej liczby rezydentów (%)
1 rezydent	3424	1.74	3424	0.01
2 rezydentów/tki	4213	2.14	8426	0.02
3 i więcej rezydentów/tek	189493	96.13	37007477	99.97
potentcjalnie: safe	189493	96.13	37007477	99.97
potentcjalnie: unsafe	7637	3.87	11850	0.03
Razem	197130	100.00	37019327	100.00

Rys. 3: Rozkład liczby rezydentów w oczkach siatki kilometrowej

Bezwzględna strata informacji wskutek ukrywania niebezpiecznych kategorii płci	Względna strata informacji wskutek ukrywania niebezpiecznych kategorii płci
-91719	0.25%

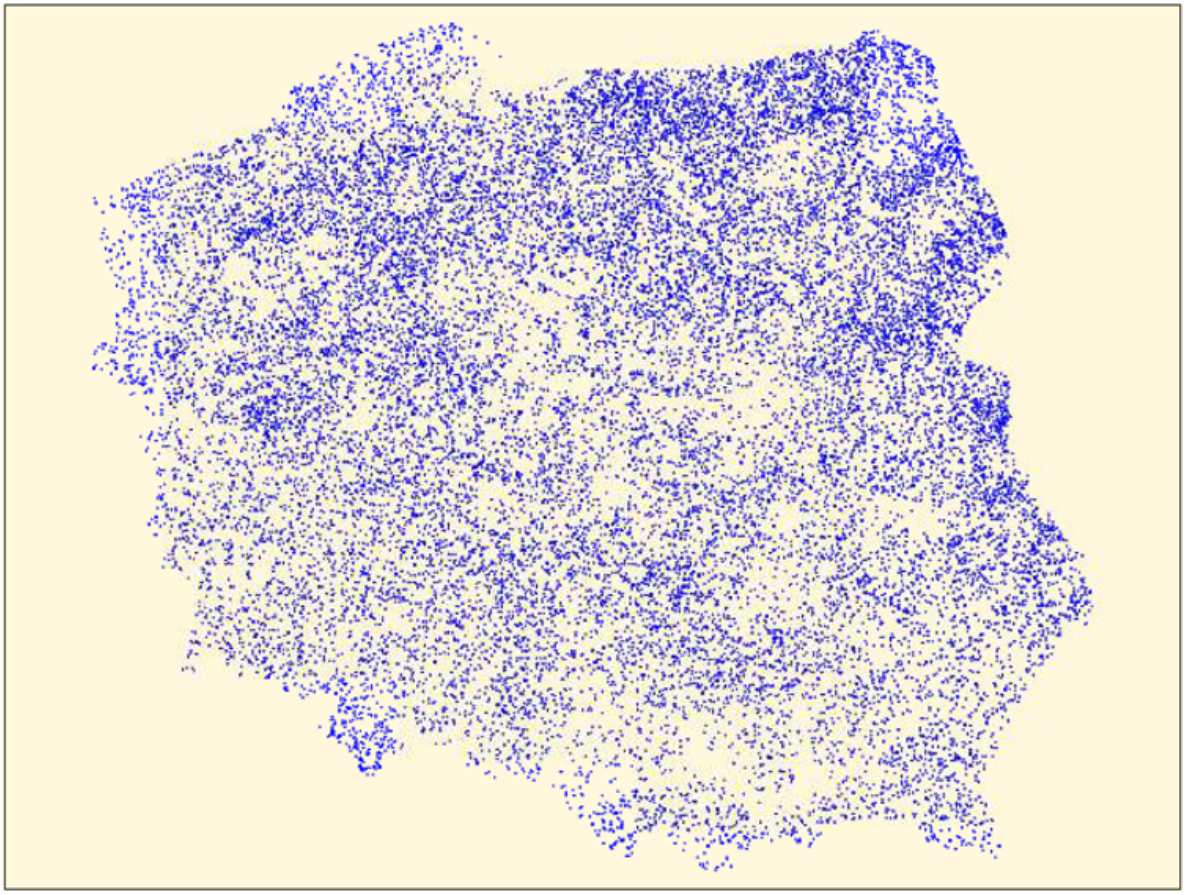
Rys. 4: Strata informacji wskutek pierwotnego/wtórnego ukrywania dla zmiennej płeć

Bezwzględna strata informacji wskutek ukrywania niebezpiecznych kategorii grup wieku	Względna strata informacji wskutek ukrywania niebezpiecznych kategorii grup wieku
-261560	0.71%

Rys. 5: Strata informacji wskutek pierwotnego/wtórnego ukrywania dla zmiennej grupy wieku

Ludność rezydująca według płci i grup wieku w siatce kilometrowej

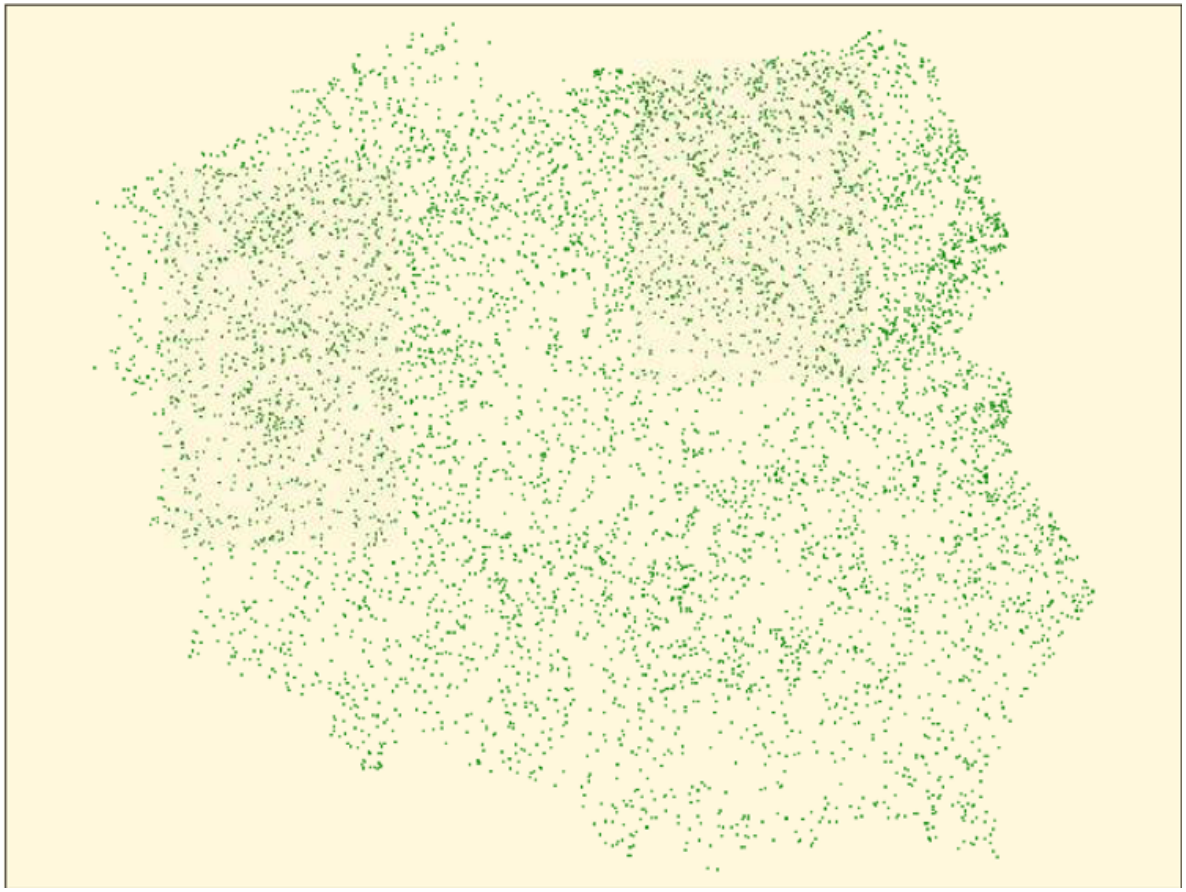
Wybrane wyniki



Rys. 6: Rozkłady przestrzenne ukryć wszystkich jednocześnie kategorii płci **lub** wieku

Ludność rezydująca według płci i grup wieku w siatce kilometrowej

Wybrane wyniki



Rys. 7: Rozkłady przestrzenne ukryć wszystkich jednocześnie kategorii płci **i** wieku

Wyzwania związane z udostępnianiem w siatce kilometrowej ludności rezydującej i krajowej według płci i grup wieku

- Problem związany z różnicami między liczbą ludności rezydującej a liczbą ludności według definicji krajowej

Zmienna	Różnica między liczbą ludności rezydującej a liczbą ludności według definicji krajowej			
	-2		-1	
	liczba	%	liczba	%
SEX1	15200	7,71	28654	14,54
SEX2	15347	7,79	26964	13,68
AGE1	10523	5,34	17699	8,98
AGE2	16063	8,15	27250	13,83
AGE3	2614	1,33	8607	4,37

Wyzwania związane z udostępnianiem w siatce kilometrowej ludności rezydującej i krajowej według płci i grup wieku

- Problem związany z różnicami między liczbą ludności rezydującej a liczbą ludności według definicji krajowej

Zmienna	Różnica między liczbą ludności rezydującej a liczbą ludności według definicji krajowej			
	1		2	
	liczba	%	liczba	%
SEX1	7599	3,86	1568	0,80
SEX2	7901	4,01	1609	0,82
AGE1	2809	1,43	1012	0,51
AGE2	8772	4,45	2028	1,03
AGE3	3909	1,98	397	0,20

Podsumowanie

- W przypadku większej liczby zmiennych zakres ukrywania wtórnego spowodowany liczniejszym występowaniem niebezpiecznych kombinacji wartości zmiennych będzie większy; wzrośnie też strata informacji,
- Problemy te pogłębią się w przypadku udostępniania tablic wielowymiarowych. Konieczne jest wtedy zastosowanie zaawansowanych narzędzi SDC przewidzianych do ochrony mikrodanych lub tablic częstości/wielkości –- np. TRS lub CKM, odpowiednio – co pozwoli na efektywniejsze osiągnięcie balansu między minimalizacją ryzyka ujawnienia a minimalizacją straty informacji
- Zastosowanie rekomendowanych przez Eurostat metod SDC (TRS i CKM) będzie wymagało dostępu do zbioru danych jednostkowych w ostatecznej postaci oraz w pełni funkcjonalnego środowiska operacyjnego obsługującego odpowiednie specjalistyczne pakiety R.

Urząd Statystyczny
w Poznaniu

Dziękujemy za uwagę!