

Urząd Statystyczny
w Poznaniu

Kontrola ujawniania danych w Portalu Geostatystycznym GUS – wyzwania i problemy

Andrzej Młodak, Tomasz Józefowski

2.07.2025

Wprowadzenie

- Poufne informacje o indywidualnych respondentach podlegają ochronie na podstawie prawa polskiego i Europejskiego, a także przyjętych zasad etycznych.
- Dane o osobach fizycznych lub prawnych identyfikujące bezpośrednio te osoby lub podmioty, bądź pozwalające na ich pośrednią identyfikację, jako dane jednostkowe identyfikowalne, podlegają **tajemnicy statystycznej**.
- Wynikowe dane statystyczne (w postaci mikrodanych, tablic statystycznych bądź wyników analiz) przed opublikowaniem lub udostępnieniem muszą zostać w odpowiedni sposób zweryfikowane i poddać ochronie – jest to przedmiotem procesu **kontroli ujawniania danych statystycznych** (ang. *Statistical Disclosure Control* – SDC).

Wprowadzenie

- Zarówno anonimizacja, jak i pseudonimizacja nie zapewniają danym jednostkowym identyfikowalnym wystarczającej ochrony, ponieważ nadal możliwa może okazać się pośrednia identyfikacja poszczególnych osób fizycznych lub prawnych na podstawie opublikowanych bądź udostępnionych danych wynikowych.
- Z tego powodu koniecznym jest poddanie wszelkich zasobów, które mają zostać opublikowane bądź udostępnione użytkownikom zewnętrznym, odpowiednio zaplanowanemu procesowi kontroli ujawniania danych statystycznych.
- Fundamentalnym zadaniem jest tutaj osiągnięcie optymalnej równowagi pomiędzy minimalizacją ryzyka ujawnienia a maksymalizacją użyteczności ujawnianych danych.

Wprowadzenie

- W wystąpieniu przypomnimy podstawowe założenia algorytmu kontroli ujawniania danych przygotowanego dla potrzeb Portalu Geostatystycznego GUS.
- Przedstawimy założenia i schemat procesu optymalnej ochrony danych oraz wskażemy na podstawowe problemy związane z praktyczną realizacją oraz możliwości ich rozwiązania.
- Zarysujemy ponadto najistotniejsze wyzwania, jakie stoją przed statystyką w tym zakresie – metodologiczne i praktyczne.

Układ prezentacji

- Kontrola ujawniania danych – podstawy
- Udostępnianie mikrodanych w Portalu Geostatystycznym
- Kontrola ujawniania danych w Portalu Geostatystycznym
- Zastosowane metody
- Pierwsze testy
- Planowane prace wdrożeniowe
- Wnioski

Kontrola ujawniania danych – podstawy

- **Kontrola ujawniania danych statystycznych** to proces polegający na weryfikacji wynikowych danych statystycznych przed ich udostępnieniem lub opublikowaniem celem wyeliminowania lub zminimalizowania ryzyka ujawnienia tożsamości respondentów lub ich cech przez zewnętrznych użytkowników tych zasobów, przy jednoczesnym zminimalizowaniu poniesionej wskutek tegoż procesu straty informacji.
- Terminy: *ochrona tajemnicy statystycznej* i *kontrola ujawniania danych statystycznych* traktuje się jako synonimy, pomimo iż pierwszy z nich ma węższe znaczenie i stanowi raczej jeden z kroków realizowanych w procesie kontroli ujawniania danych statystycznych.
 - W procesie tym bowiem nie tylko chroni się w danych wynikowych tajemnicę statystyczną z wykorzystaniem odpowiednich metod i technik, ale również bada się wpływ tych rozwiązań na poufność i użyteczność owych danych.

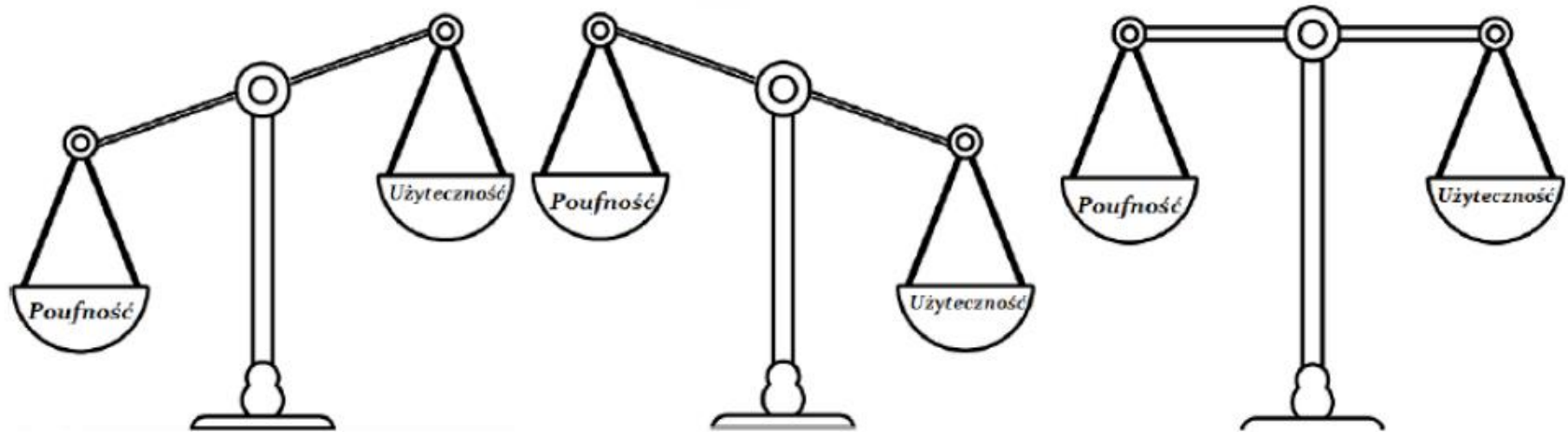
Kontrola ujawniania danych – podstawy

- **Ujawnienie** – zachodzi wówczas, gdy użytkownik rozpoznaje respondenta lub dowiaduje się o nim czegoś, czego wcześniej nie wiedział, na podstawie udostępnionych wynikowych danych statystycznych.
- **Strata informacji** – to ubytek zasobu informacyjnego wynikowych danych statystycznych (tj. mikrodanych, tablic statystycznych czy innej postaci danych wynikowych) będący efektem zastosowania metod kontroli ujawniania danych statystycznych; bywa uważana za pojęcie subiektywne, choć jej skalę w wymiarze obiektywnym można dość precyzyjnie oceniać na różne sposoby (w zależności od typu danych wynikowych).

Kontrola ujawniania danych – podstawy

- Poziom ryzyka ujawnienia informacji poufnych przekłada się na zapewnianą poufność danych wynikowych – im mniejsze ryzyko ujawnienia, tym większa zapewniana poufność danych wynikowych.
- Strata informacji powstała wskutek nałożenia tajemnicy statystycznej na dane wynikowe znajduje odzwierciedlenie w ich użyteczności – im mniejsza ponoszona strata informacji, tym większa użyteczność danych wynikowych.
- Przeprowadzając proces kontroli ujawniania danych statystycznych nie można pominąć żadnego z dwóch wymienionych wyżej aspektów – należy zapewnić odpowiedni balans pomiędzy zapewnianą poufnością danych wynikowych (czyli minimalizacją ryzyka ujawnienia informacji poufnych), a ich użytecznością (czyli ponoszoną wskutek nałożenia tajemnicy statystycznej stratą informacji).

Kontrola ujawniania danych – podstawy



Kontrola ujawniania danych – podstawy

- Klasyfikacja metod
 - metody niezakłóceńowe (maskowanie danych wrażliwych)
 - anonimizacja (usunięcie identyfikatorów)
 - podpróbkiowanie (ang. subsampling)
 - przekodowanie (ang. recoding)
 - lokalne ukrywanie danych (ang. local suppression)
 - metody zakłóceńowe (zaburzanie danych)
 - dodawanie szumu (addytywne, addytywne z transformacją liniową, multiplikatywne)
 - mikroagregacja (z wykorzystaniem analizy skupień do wyodrębniania jednorodnych grup jednostek)
 - wymiana danych (wymiana rang, celowana wymiana rekordów)
 - zaokrąglanie
 - próbkiowanie wtórne
 - metoda postrandomizacyjna (PRAM)

Udostępnianie mikrodanych w Portalu Geostatystycznym

- Portal Geostatystyczny Głównego Urzędu Statystycznego (<https://portal.geo.stat.gov.pl/pl/home/>) powstał z myślą o zaspokojeniu zapotrzebowania użytkowników na szczegółowe i wysokiej jakości dane na różnych poziomach przestrzennych, w tym umożliwienia użytkownikowi przeprowadzenia własnych badań i analiz.
- Został on opracowany w ramach projektu PDS (Przestrzenne Dane Statystyczne w Systemie Informacyjnym Państwa), mającego na celu poszerzenie zakresu i dostępności informacji statystycznych oraz metod analizy geostatystycznej wykorzystujących zasoby statystyki publicznej.

Udostępnianie mikrodanych w Portalu Geostatystycznym

- Szczegółowe i wszechstronne upowszechnianie danych statystycznych oraz umożliwienie zaawansowanych analiz wymaga szczególnie starannej ochrony tajemnicy statystycznej.
- Dlatego też, poza spełnieniem wymagań wynikających z aktualnie obowiązujących przepisów prawa, konieczne jest zastosowanie dodatkowych, profesjonalnych narzędzi ochrony przed identyfikacją jednostki i nieuprawnionym ujawnieniem informacji wrażliwych na jej temat.
- Niezbędne jest także, aby osoby odpowiedzialne za ochronę tajemnicy statystycznej w jednostkach autorskich nabyły odpowiednie kompetencje

Kontrola ujawniania danych w Portalu Geostatystycznym

- Przyjęta podstawowa koncepcja zakłada, że w systemie (jeśli mówimy o indywidualnych danych statystyki publicznej) do analizy udostępniane są jedynie zbiory danych zaburzonych, otrzymane przy zastosowaniu zakłóceńowych metod kontroli ujawniania danych
- Stosowana parametryzacja metod jest poufna, użytkownik natomiast może otrzymać informację o oczekiwanej stracie informacji na skutek zaburzania
- Każdy zbiór danych statystycznych, który zostałby udostępniony użytkownikom zewnętrznym za pośrednictwem systemu, musi podlegać tej czynności. Przeprowadzić ją winien analityk posiadający stosowną wiedzę i doświadczenie

Kontrola ujawniania danych w Portalu Geostatystycznym

- Wykonanie zaburzania danych dla każdego zbioru polega na utworzeniu odpowiedniego skryptu na podstawie dostarczonego przez wykonawcę szablonu, w którym należy określić rolę poszczególnych zmiennych zbioru wejściowego w procesie zakłócania oraz parametry tego procesu związane z zastosowanymi metodami.
- Dlatego też rdzeniem procesu SDC w systemie jest skrypt R oparty na zasobach pakietu `sdcMicro` zagnieżdżony w SAS/Python i umożliwiający zastosowanie odpowiednich metod zakłóceńowych przy ryzyku ujawnienia kontroli i utraty informacji.

Zastosowane metody

- **Zakłócenia dla zmiennych kategorialnych (kluczowych)**
 - mikroagregacja z wykorzystaniem odległości Gowera; parametry:
 - poziom agregacji
 - przypisanie zastępczej wielkości (catFun)
 - „sampleCat” – przypisanie jednostce należącej do danego skupienia odpowiedniej kategorii w sposób losowy z prawdopodobieństwem odpowiadającym częstości występowania tejże kategorii wśród najbliższych sąsiadów owej jednostki
 - „maxCat” – wybór poziomu, który występuje najczęściej lub losowego jeśli to maksimum częstości zachodzi dla dwóch lub więcej wielkości

Zastosowane metody

- **Zakłócenia dla zmiennych kategoryalnych (kluczowych)**
 - mikroagregacja z wykorzystaniem odległości Gowera (zob. np. Młodak i in. (2023)); parametry:
 - `dist_var` – zmienne w oparciu o które wyznacza się odległość między obiektami
 - `by` – grupowanie w obrębie wskazanych predefiniowanych grup (np. obszarów przestrzennych)
 - `addRandom` – dodawanie zakłócenia do odległości
 - post-randomizacja (PRAM); parametry:
 - `pd` – minimalne elementy diagonalne wygenerowanej macierzy przejścia
 - `alpha` – dopuszczalne wielkości zakłóceń elementów macierzy przejścia poza przekątną w relacji do `pd`

Zastosowane metody

- **Zakłócenia dla zmiennych ciągłych**
 - dodawanie szumu
 - metoda – `correlated2` (dodawanie szumu z zachowaniem kowariancji zmiennych)
 - parametry
 - `noise` – wielkość szumu (w procentach)
 - `delta` – dodatkowy parametr stosowany wyłącznie dla tej metody, zależny od poziomu zakłóceń, który określa poziom zachowania oryginalnej wartości przed zaszumieniem a jednocześnie wyznacza skalę szumu
 - mikrogregacja oparta na odległości Gowera (możliwa z parametrem `numFun` – np. `numFun=„mean”`)

Zastosowane metody

- **Ocena straty informacji**

- przebiegała w dwóch głównych kierunkach:
 - miara zakłócenia rozkładu,
 - miara zaburzenia powiązań pomiędzy zmiennymi ciągłymi.
- pierwsza miara opiera się na idei odległości Gowera pomiędzy wartościami pierwotnymi i zaburzonymi, z tą różnicą, że dla zmiennych ciągłych odległości cząstkowe obliczane są za pomocą funkcji cyklometrycznej i przyjmuje wartości z zakresu $[0,1]$. Im większa wartość miary tym większa strata informacji. Utratę informacji można mierzyć zarówno na poziomie globalnym, jak i dla poszczególnych zmiennych. Jest to rozwiązanie zaproponowane przez Młodaka, Pietrzaka i Józefowskiego (2022) i zaimplementowane w pakiecie `sdcMicro` jako funkcja `IL_variables()`.

Zastosowane metody

- **Ocena straty informacji**

- miarą utraty informacji na powiązaniach pomiędzy zmiennymi jest podejście oparte na odległościach pomiędzy diagonalnymi elementami macierzy odwrotnej do macierzy korelacji i przyjmuje wartości z $[0,1]$ (ponownie im większa wartość, tym większa strata).
- jest to rozwiązanie zaproponowane przez Młodaka, Pietrzaka i Józefowskiego (2022) i zaimplementowane w pakiecie `sdcMicro` jako funkcja `IL_correl()`.

Pierwsze testy

- Algorytm był testowany na częściowym zbiorze danych z PSR 2020, zawierającym 1 317 400 rekordów i 81 zmiennych opisujących podstawowe cechy gospodarstw rolnych i informacje na temat hodowli zwierząt gospodarskich
- Działania prowadzono na 55 zmiennych pierwotnych, identyfikatory województwa i powiatu połączono w jeden, wymianę danych w tym zakresie dopuszczano dla gmin w obrębie powiatu
- Pełna realizacja algorytmu zajęła prawie 35 godzin. Tak długi czas wynikał głównie z dużej liczby zmiennych, złożoności skryptu i zastosowanego środowiska obliczeniowego.
- W przypadku zmiennych kategoryalnych kategorie o najmniejszej częstotliwości występują w ponad 6 tysiącach rekordów. Zasada k -anonimowości dla $k=2, 3$ i 5 praktycznie nie jest zatem łamana. Dlatego ryzyko wynikające ze zmiennych kategorycznych jest znikome.

Pierwsze testy

- W przypadku zmiennych ciągłych ryzyko ujawnienia ocenia się jako odsetek obserwacji mieszczących się w przedziale wycentrowanym na zaburzonej wartości, podczas gdy górna granica takiego przedziału odpowiada najgorszemu scenariuszowi, w którym intruz ma pewność, że każdy najbliższy sąsiad tej wielkości faktycznie jest nią; stosowany w ujęciu porównawczym (po procesie SDC w porównaniu do sytuacji przed)
- W przypadku surowych danych ryzyko mieści się w przedziale od 0% do 100%. Po procesie SDC przedział ryzyka skurczył się z $[0,00\%, 100,00\%]$ do $[0,00\%, 0,00\%]$. Zatem ochrona jest idealna.
- Strata na sile związku między zmiennymi ciągłymi wyniosła 7,9%. Wobec tego zależności pomiędzy rzeczonymi zmiennymi zostają w przeważającej mierze zachowane. Wpływ na to może mieć zastosowana metoda nakładania szumu
- Więcej szczegółów na temat testu – Młodak i Józefowski (2025).

Planowane prace wdrożeniowe

- Docelowo planuje się załadowanie do Portalu Geostatystycznego zbioru zawierającego obok danych o pogłowie zwierząt gospodarskich także informacje o powierzchniach zasiewów, zużyciu nawozów, działaniach w zakresie ochrony roślin, maszynach i urządzeniach rolniczych, budynkach gospodarskich, działalności gospodarczej gospodarstw rolnych ich dochodach oraz aktywności ekonomicznej ich członków (242 zmienne)
- Na tym zbiorze należy efektywnie przeprowadzić proces kontroli ujawniania danych, bazując na doświadczeniach testowych.

Planowane prace wdrożeniowe

- Problemy, które należy rozwiązać w tym zakresie
 - ustalenie które zmienne są funkcjami innych zmiennych zawartych w zbiorze (dla utrzymania spójności będą one tymczasowo pominięte)
 - ewentualne sprawdzenie czy pewne dane źródłowe nie zostały już udostępnione w inny sposób, a jeśli tak, to ocena możliwości zachowania spójności z bieżącym zbiorem po realizacji procesu SDC
 - określenie roli poszczególnych zmiennych w kontekście SDC
 - dobór metod SDC i ich parametrów
 - korekta zaburzeń wedle definicji zmiennych
 - optymalizacja oceny ryzyka ujawnienia i straty informacji

Planowane prace wdrożeniowe

- Proponowana procedura SDC
 1. Przygotowanie zbioru, w tym wskazanie zmiennych, które na pewno nie będą udostępnione
 2. Wskazanie przez gestora zmiennych przewidywanych do udostępnienia, które są identyfikatorami (pośrednimi lub bezpośrednimi) oraz ustalenie na jakiej skali pomiarowej wyrażają się ich dane, tzn. które zmienne są kategoryjne (tzn. których dane wyrażają się na skali nominalnej, a których na skali porządkowej – w tym drugim przypadku ze wskazaniem zastosowanego porządku, o ile nie jest oczywisty), a które ciągłe (czyli o danych na skali różnicowej lub ilorazowej), a także tych, których wartości są funkcjami (np. sumami) innych (te nie będą zaburzane, zostaną naliczone przez gestora wtórnie, po zaburzeniu).
 3. Ocena ryzyka ujawnienia *ex ante* (możliwa dodatkowa eliminacja zmiennych szczególnie wrażliwych).

Planowane prace wdrożeniowe

- Proponowana procedura SDC

4. Wykonanie zaburzania:

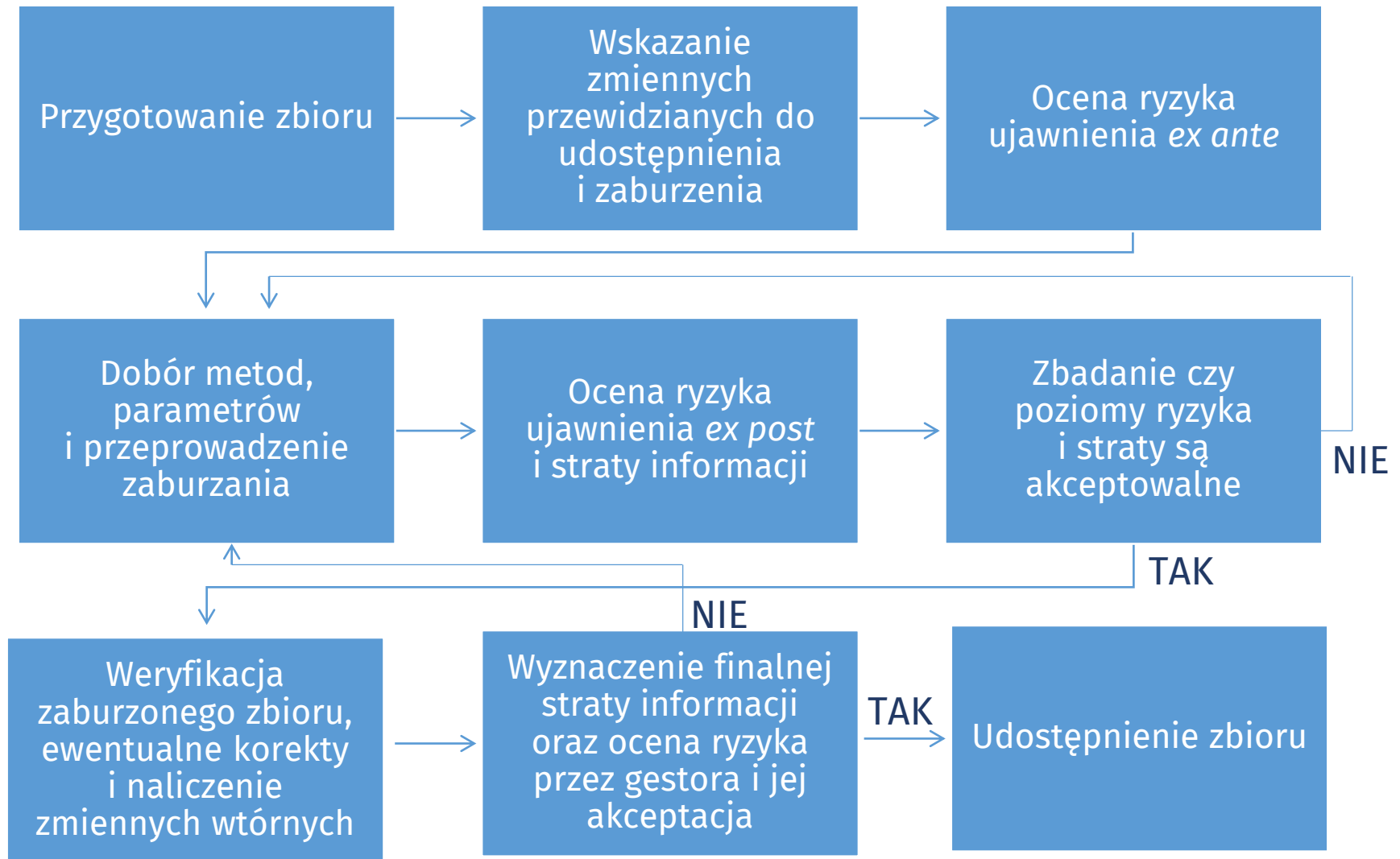
- wybór metod SDC, które zostaną zastosowane, spośród dostępnych opcji
- ustalenie parametrów konfiguracyjnych powyższych metod
- nałożenie zakłóceń

5. Wyznaczenie ryzyka ujawnienia *ex post* i straty informacji. Ewentualny powrót do kroku 4 w przypadku zbyt dużego ryzyka ujawnienia lub zbyt dużej straty informacji (ogółem lub na poziomie poszczególnych zmiennych); możliwa też najpierw dodatkowa eliminacja zmiennych

Planowane prace wdrożeniowe

- Proponowana procedura SDC
 6. Weryfikacja zaburzonego zbioru:
 - korekty w przypadku odbiegania zaburzonych wartości od zakresu wyznaczonego przez definicję danej zmiennej lub innymi założeniami
 - powtórne wyznaczenie wartości zmiennych pochodnych na podstawie zaburzonych wartości zmiennych pierwotnych
 7. Wyznaczenie finalnej straty informacji; w razie zbyt dużych strat można wrócić do kroku 4
 8. Przeprowadzenie oceny ryzyka ujawnienia danych przez gestora; ewentualny powrót do kroku 4
 9. Zgoda gestora i udostępnienie zbioru

Schemat procesu SDC



Wnioski

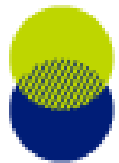
- Przeprowadzone prace testowe wykazały, że zastosowanie perturbacyjnych metod SDC może być bardzo przydatne.
- Problem stanowi długi czas realizacji algorytmu zaburzenia; pożądana jest wydajniejsza infrastruktura informatyczna oraz optymalizacja zestawu zmiennych przeznaczonych do udostępniania.
- Bardzo ważne jest wprowadzenie korekt *ex post* uspójniających zaburzone wartości z zakresem definicyjnym i innymi ograniczeniami, co może mieć również pewny (raczej umiarkowany) wpływ na ostateczną utratę informacji.

Wnioski

- Ustalenie akceptowalnych poziomów ryzyka ujawnienia i straty informacji powinno efektem ustaleń uwzględniających istotność zmiennych, planowany sposób udostępniania i przewidywanych użytkowników.
- Niezbędne jest zapewnienie właściwej koordynacji udostępniania danych, celem zagwarantowania właściwej ochrony udostępnionych lub planowanych do udostępnienia pokrewnych danych z różnych badań, których powiązanie z bieżącymi może zwiększać ryzyko identyfikacji jednostki.
- Proponowany schemat działań może być zastosowany do każdego badania statystycznego, mikrodane z którego zamierza się udostępniać.

Bibliografia

- Młodak A., Józefowski T. (2025), Application of Statistical Disclosure Control methods to protect the confidentiality of the 2020 agricultural census microdata, *Statistics in Transition-new series*, Vol. 26, No. 2 (w druku).
- 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, Poznań.
- 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*, Vol. 38, str. 1503–1511.



Urząd Statystyczny
w Poznaniu

Dziękujemy za uwagę