Wiadomości Statystyczne. The Polish Statistician, 2020, vol. 65, 9, 7–27 DOI: 10.5604/01.3001.0014.4121
Studia metodologiczne / Methodological studies
Otrzymano/received: 3.06.2020, zaakceptowano/accepted: 6.08.2020

Strata informacji wskutek przeprowadzenia kontroli ujawniania danych wynikowych

Andrzej Młodak[a]

Streszczenie. W pracy omówiono najważniejsze metody, za pomocą których można ocenić stratę informacji spowodowaną przeprowadzaniem kontroli ujawniania danych (ang. *statistical disclosure control*, SDC). Kontrola ta ma na celu ochronę przed identyfikacją jednostki i dotarciem do dotyczących jej wrażliwych informacji przez osoby nieupoważnione. Zastosowanie metod zarówno opartych na ukrywaniu określonych danych, jak i prowadzących do ich zniekształcania powoduje stratę informacji, która ma wpływ na jakość danych wynikowych, w tym rozkładów zmiennych, kształt ich związków oraz estymacji. Celem artykułu jest krytyczna analiza mocnych i słabych stron metod oceny straty informacji na skutek zastosowania SDC. Przedstawiono również nowatorskie propozycje prowadzące do uzyskania efektywnych i dobrze interpretowalnych mierników, m.in. nową możliwość wykorzystania funkcji cyklometrycznej (arcus tangens) do wyznaczenia odchylenia wartości od tych oryginalnych po przeprowadzeniu SDC. Ponadto zastosowano odwróconą macierz korelacji do oceny wpływu SDC na siłę związków między zmiennymi. Pierwsza z przedstawionych metod umożliwia uzyskanie efektywnych i dobrze interpretowalnych mierników, druga – maksymalne wykorzystanie wzajemnych powiązań między zmiennymi (także tych trudno uchwytnych za pomocą klasycznych metod statystycznych) w celu lepszej analizy skutków kontroli w tym zakresie.

Empiryczna weryfikacja użyteczności sugerowanych metod potwierdziła m.in. przewagę funkcji cyklometrycznej w pomiarze odległości w zakresie uwypuklania odchyleń od danych oryginalnych, a także potrzebę umiejętnej korekcji jej spłaszczenia przy dużej wartości argumentów.

Słowa kluczowe: kontrola ujawniania danych, SDC, strata informacji, funkcja cyklometryczna, odwrócona macierz korelacji

JEL: C19, C63, C80, D82

Information loss resulting from statistical disclosure control of output data

Abstract. The most important methods of assessing information loss caused by statistical disclosure control (SDC) are presented in the paper. The aim of SDC is to protect an individual against identification or obtaining any sensitive information relating to them by anyone unauthorised. The application of methods based either on the concealment of specific data or on their perturbation results in information loss, which affects the quality of output data, including the distributions of variables, the forms of relationships between them, or any estimations. The aim of this paper is to perform a critical analysis of the strengths and weaknesses of the particular types of methods of assessing information loss resulting from SDC. Moreover, some

[a] Akademia Kaliska im. Prezydenta Stanisława Wojciechowskiego, Międzywydziałowy Zakład Matematyki i Statystyki; Urząd Statystyczny w Poznaniu, Ośrodek Statystyki Małych Obszarów.
ORCID: https://orcid.org/0000-0002-6853-9163.

novel ideas on how to obtain effective and well-interpretable measures are proposed, including an innovative way of using a cyclometric function (arcus tangent) to determine the deviation of values from the original ones, as a result of SDC. Additionally, the inverse correlation matrix was applied in order to assess the influence of SDC on the strength of relationships between variables. The first presented method allows obtaining effective and well--interpretable measures, while the other makes it possible to fully use the potential of the mutual relationships between variables (including the ones difficult to detect by means of classical statistical methods) for a better analysis of the consequences of SDC.

Among other findings, the empirical verification of the utility of the suggested methods confirmed the superiority of the cyclometric function in measuring the distance between the curved deviations and the original data, and also heighlighted the need for a skilful correction of its flattening when large value arguments occur.

Keywords: statistical disclosure control, SDC, information loss, cyclometric function, inverse correlation matrix

1. Wprowadzenie

Właściwa i efektywna ochrona tajemnicy statystycznej polega m.in. na dokonaniu weryfikacji danych statystycznych przed ich udostępnieniem w celu wyeliminowania – lub przynajmniej zminimalizowania – ryzyka ujawnienia bądź odtworzenia przez użytkowników udostępnianych zasobów wrażliwych informacji o jednostkach statystycznych, czego konsekwencją mogłaby być identyfikacja konkretnych jednostek. Immanentną cechę postępowania prowadzącego do osiągnięcia tego celu, które znane jest powszechnie jako kontrola ujawniania danych (ang. *statistical disclosure control*, SDC), stanowi wprowadzanie niepewności w odniesieniu do prawdziwej wartości udostępnianych zmiennych. Niepewność ta w naturalny sposób wiąże się z ukryciem lub zmianą wartości w rekordzie zbioru mikrodanych czy w komórce tablicy. W konsekwencji na skutek zastosowania SDC powstaje określona strata informacji źródłowej. Taka strata może wpłynąć na jakość udostępnianych danych oraz obliczeń i szacunków dokonywanych przez ich użytkownika. Mówiąc bardziej precyzyjnie, strata stanowi – niekiedy istotną – składową obciążenia wyników uzyskiwanych przez użytkownika (np. estymacji określonych wielkości). Dotyczy ona wprawdzie bezpośrednio informacji źródłowych (ponieważ strata wynika z usunięcia lub zniekształcenia określonych danych), ale może mieć znaczny wpływ na kształt rozkładów określonych zmiennych i związków pomiędzy nimi. Użytkownik powinien więc razem z danymi otrzymywać informację na temat oczekiwanej wielkości rozpatrywanej straty spowodowanej przez SDC. Minimalizacja tej straty jest kluczowym – obok minimalizacji ryzyka identyfikacji jednostki i ujawnienia informacji wrażliwych – kryterium optymalizacji SDC. Optymalizacja ta oznacza bowiem udostępnianie mikrodanych oraz publikowanie tablic i analiz w możliwie największym niezmienionym zakresie. Ma to duże znaczenie nie tylko dla bezpośredniej wartości informacyjnej udostępnianych zasobów, lecz także dla zapewnienia odpowiedniej jakości szacunków i analiz. Konieczny staje się więc efektywny pomiar wielkości

straty informacji spowodowanej zastosowaniem SDC. Problem ten badany jest stosunkowo rzadko. Podwaliny pod badania położyli i pierwszych kompleksowych analiz dokonali Domingo-Ferrer, Mateo-Sanz i Torra (2001). Wszechstronnie takie metody scharakteryzowali m.in. Shlomo i Young (2006) oraz Hundepool i współpracownicy (2012).

Celem artykułu jest krytyczna analiza mocnych i słabych stron metod oceny straty informacji na skutek zastosowania SDC. Przedstawiono również nowatorskie propozycje prowadzące do uzyskania efektywnych i dobrze interpretowalnych mierników, które obejmują wykorzystanie funkcji cyklometrycznej (arcus tangens) do wyznaczenia odchylenia wartości od tych oryginalnych po przeprowadzeniu SDC oraz zastosowanie odwróconej macierzy korelacji do oceny wpływu takiej kontroli na siłę związków między zmiennymi. Pierwsza z nich umożliwia uzyskanie efektywnych i dobrze interpretowalnych mierników przy zachowaniu klarownego, z punktu widzenia potencjalnego użytkownika, przekazu dotyczącego oczekiwanej straty informacji, druga – maksymalne wykorzystanie wzajemnych powiązań między zmiennymi (także tych trudno uchwytnych za pomocą klasycznych metod statystycznych) w celu lepszej analizy skutków kontroli w tym wymiarze.

2. Istota straty informacji

Pojęcie straty informacji na skutek zastosowania metod ochrony poufności ma charakter subiektywny. Ważne w tym kontekście stają się cele analiz, w jakich znajdą zastosowanie wynikowe mikrodane i opublikowane tablice. Strata informacji może być oceniana w różny sposób przez pryzmat decyzji, które muszą zostać podjęte w celu zachowania poufności, oraz konsekwencji ich wyboru. Shlomo i Young (2006) wskazują na trzy główne grupy ocen straty wyodrębnione w zależności od statystycznego aspektu rozpatrywanych danych. Czynią to wprawdzie w odniesieniu do tablic częstości, ale klasyfikację tę bez trudu można zastosować również do innego rodzaju informacji statystycznych (mikrodanych, tablic wielkości, prezentacji wynikowych itp.). Wyróżnia się następujące grupy miar oceny straty informacji:

- miary zakłócenia rozkładu, które są oparte na metrykach odległości pomiędzy rzeczywistymi a zmienionymi wartościami zmiennych. Jeśli jednostką podstawową jest obszar przestrzenny, to dla każdego takiego obszaru z osobna mierzy się odległość pomiędzy nimi, po czym oblicza się średnią z tych odległości;
- miary wpływu na wariancję szacunków, w przypadku których pod uwagę bierze się różnice wariancji dla przeciętnych wartości określonych podzbiorów lub całego zbioru (w przypadku tablic – kolumn, wierszy lub całej tablicy). Innym sposobem jest wieloczynnikowa analiza wariancji ANOVA dla wybranej zmiennej zależnej względem wybranych niezależnych zmiennych kategorialnych. Miarą straty jest porównanie, jak zmieniają się komponenty współczynnika determinacji R^2

(poprzez podział na wariancję wewnątrzgrupową i międzygrupową) dla zbioru źródłowego opartego na danych rzeczywistych oraz dla zbioru zmienionego poprzez zastosowanie danej metody ochrony poufności (lub dla tablicy źródłowej opartej na danych rzeczywistych oraz dla tablicy zmienionej w wyniku zastosowania danej metody ochrony poufności). Zastosowanie to może spowodować utratę homogeniczności grup, co oznacza, że wariancja międzygrupowa może maleć, a wewnątrzgrupowa – rosnąć. Możliwa jest też sytuacja odwrotna, czyli wzrost wariancji międzygrupowej przy spadku wariancji wewnątrzgrupowej, np. w przypadku mikroagregacji[1]: gdy w każdym skupieniu jednostek każda oryginalna wartość zastępowana jest przez średnią grupową odpowiedniej zmiennej dla tego skupienia, wówczas zmienność w takim skupieniu jest równa 0 (jednakowe wartości zmiennej), podczas gdy różne skupienia mogą być między sobą bardzo zróżnicowane (zob. np. Hundepool i in., 2012);

- miary wpływu na siłę związku, które bazują na analizie oddziaływania SDC na kierunek i siłę związku między określonymi zjawiskami w porównaniu z tymi cechami dla oryginalnych zmiennych. W tym celu można wykorzystać współczynniki korelacji, często jednak wykonuje się także test niezależności pomiędzy odpowiednimi wymiarami w stosownych przekrojach. Oznacza to badanie niezależności przy użyciu pewnej tablicy kontyngencji. Możliwe są także inne podejścia w tym zakresie.

Można zauważyć, że mierniki zakłócenia rozkładu są użyteczne, gdy osobę korzystającą z udostępnionych danych interesują rozkłady określonych zjawisk, np. w czasie i przestrzeni. Wpływ SDC na wariancję szacunków ma istotne znaczenie w przypadku estymacji określonych wielkości dla populacji. Z kolei ocena zmiany siły związków na skutek zastosowania narzędzi SDC może być szczególnie ważna podczas analizy współzależności zjawisk. Warto jednak zauważyć, że niektóre rodzaje badań mogą wymagać mierników straty ukierunkowanych nie tylko w jedną stronę (np. podczas estymacji dokonywanej z wykorzystaniem estymatorów regresyjnych pożądany będzie pomiar straty informacji w kierunku wpływu zarówno na wariancję szacunków, jak i na siłę związku między określonymi zjawiskami). W kolejnych częściach pracy szczegółowo ukazane zostaną konkretne metody oceny straty informacji należące do tych grup.

[1] Mikroagregacja (ang. *microaggregation*) – rodzina zakłóceniowych narzędzi SDC zapewniających ochronę poufności danych w ujęciu makro poprzez odpowiednie działania na poziomie mikro. U podstaw stosowania mikroagregacji leżą zasadnicze reguły publikacyjne, dopuszczające publikowanie zbiorów mikrodanych, gdy zawierają one co najmniej k rekordów, a żaden z nich nie dominuje pod danym względem (tzn. jego udział w danej wielkości ogółem dla grupy nie jest większy niż $p\%$). Parametry naturalne k i p ($0 < p < 100$) są zazwyczaj arbitralnie ustalone. Ich wspólną cechę stanowi podział obiektów na wewnętrznie jednorodne, co najmniej k-elementowe podzbiory. Dla każdej takiej grupy obliczane są średnie wartości zmiennych z danymi wrażliwymi – i tymi średnimi zastępuje się oryginalne wartości (zob. np. Hundepool i in., 2012).

3. Miary zakłócenia rozkładu

Pomiar straty informacji w kontekście zakłócenia rozkładu oryginalnych zmiennych wskutek zastosowania SDC opiera się na unormowanych różnicach między odpowiednimi wartościami w zbiorze danych oryginalnych oraz w zbiorze danych zniekształconych. Należy przy tym uwzględnić skalę pomiarową, na jakiej mierzone są poszczególne obserwacje, oraz dopuszczalność wykonywania na nich odpowiednich działań arytmetycznych.

Niech zatem zbiory danych liczą n obserwacji i m zmiennych (gdzie n i m to liczby naturalne). Oznaczmy przez x_{ij} wartość obserwacji zmiennej X_j dla jednostki i, a przez x_{ij}^* – odpowiednią wartość w zbiorze powstałym w efekcie zastosowania metod SDC, $i = 1, 2, \ldots, n,\ \ j = 1, 2, \ldots, m$. Ogólna postać miary straty może być zatem następująca:

$$\lambda = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n} d(x_{ij}, x_{ij}^*)}{mn}, \tag{1}$$

gdzie $d(\cdot,\cdot)$ – miara odległości spełniająca klasyczne warunki zwrotności, symetrii i nierówności trójkąta, przyjmująca wartości należące do przedziału [0, 1].

Wartości wskaźnika λ także należą wtedy do tego przedziału, przy czym im wyższa wartość, tym bardziej dotkliwa strata informacji. Sytuacja, gdy $\lambda = 0$, tzn. gdy w zbiorze nie ma żadnych zmian, jest w SDC oczywiście tylko teoretyczna.

Definicja miary d zależy od skali pomiarowej, na której obserwowane są wartości zmiennej X_j. Jeśli obserwacje tej zmiennej mierzone są na skali nominalnej, to odległość ta wynosi

$$d\left(x_{ij}, x_{ij}^*\right) = \begin{cases} 0, \text{gdy } x_{ij} = x_{ij}^*, \\ 1, \text{gdy } x_{ij} \neq x_{ij}^*. \end{cases} \tag{2}$$

Jeżeli wartości zmiennej X_j mierzone są na skali porządkowej, wówczas rzeczoną miarą jest relacja

$$d\left(x_{ij}, x_{ij}^*\right) = \frac{r(x_{ij}, x_{ij}^*)}{k_j - 1}, \tag{3}$$

gdzie:

$r(x_{ij}, x_{ij}^*)$ – liczba tych kategorii zmiennej X_j, o którą różnią się wartości x_{ij} i x_{ij}^* w danym porządku (np. jeżeli zmienna przyjmuje kategorie A, B, C, D

– porządek $<$ jest tu alfabetyczny: $A < B < C < D$ – to kategoria C różni się od kategorii A o dwie pozycje, natomiast od B – o jedną),

k_j – liczba kategorii, które może przyjmować zmienna X_j.

Dla zmiennej ciągłej odległość ta może być np. znormalizowaną wartością bezwzględną lub znormalizowanym kwadratem różnicy między wartością ze zbioru oryginalnego a odpowiednią wartością ze zbioru ukształtowanego w wyniku SDC, czyli

$$d\left(x_{ij}, x_{ij}^*\right) = \frac{\left|x_{ij} - x_{ij}^*\right|}{\max\limits_{k=1,\,2,\,\ldots,\,n} \left|x_{kj} - x_{kj}^*\right|} \tag{4}$$

lub

$$d\left(x_{ij}, x_{ij}^*\right) = \frac{\left(x_{ij} - x_{ij}^*\right)^2}{\max\limits_{k=1,\,2,\,\ldots,\,n} \left(x_{kj} - x_{kj}^*\right)^2}, \tag{5}$$

gdzie $i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m$.

Stosowanie tych miar pociąga za sobą pewne problemy. Jednym z narzędzi SDC dla zmiennych kategorialnych jest przekodowywanie. Liczby kategorii przekodowywanej zmiennej w zbiorze oryginalnym i w nowym, powstałym w efekcie SDC, będą różne. Należy więc zadbać przede wszystkim o to, aby numery kategorii pozostawionych bez zmian były identyczne w obu wariantach. Na przykład jeśli przed przekodowaniem zmienna X_j liczyła $k_j = 8$ kategorii oznaczonych jako 1, 2, 3, 4, 5, 6, 7, 8, a w wyniku przekodowania połączono kategorie 2 i 3 oraz 6 i 7, to nowe kategorie powinny mieć odpowiednio numery 1, 2, 4, 5, 6, 8. Wtedy podejście (3) stosuje się i w tym przypadku.

Miary (4) i (5) mają też inny znamienny mankament. Miernik straty informacji powinien być funkcją rosnącą ze względu na poszczególne cząstkowe straty informacji. Oznacza to, że np. jeśli dla pewnego $i \in \{1, 2, \ldots, n\}$ wartość $\left|x_{ij} - x_{ij}^*\right|$ zwiększy się, a wszystkie wartości $\left|x_{hj} - x_{hj}^*\right|$ dla $h \neq i$ pozostaną takie same, to wartość miernika powinna wzrosnąć. Tymczasem w przypadku formuł (4) i (5) tak nie będzie. Jeśli bowiem dla i wskazana bezwzględna różnica (lub kwadrat różnicy, odpowiednio) między wartością oryginalną a wartością po przeprowadzeniu SDC osiągnie maksimum, to cząstkowa strata informacji dla i pozostanie bez zmian – wyniesie 1, a dla pozostałych okaże się mniejsza. W efekcie otrzymamy mniejszą wartość miernika, podczas gdy strata informacji tak naprawdę się zwiększyła.

Można temu zaradzić, korzystając z tego, że wspomniane odległości są nieujemne, i oprzeć się na ograniczonej i rosnącej funkcji arcus tangens. Przyjmujemy zatem

$$d\left(x_{ij}, x_{ij}^*\right) = \frac{2}{\pi}\operatorname{arctg}\left|x_{ij} - x_{ij}^*\right|, \tag{6}$$

gdzie $i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m$.

Kolejny problem pojawia się w przypadku ukrywania, w zbiorze poddanym SDC powstają bowiem luki w danych. Gdy dane o zmiennej X_j wyrażone są na skali nominalnej, to jeżeli obserwacja x_{ij}^* w (2) jest ukryta, wówczas przypisujemy $d\left(x_{ij}, x_{ij}^*\right) = 1$. Jeśli obserwacje zmiennej X_j wyrażają się na skali porządkowej, wtedy dla celów obliczeniowych we wzorze (3) ukrytej wartości x_{ij}^* przyporządkowujemy arbitralnie $x_{ij}^* \coloneqq 1$, gdy x_{ij} jest bliższe k_j lub $x_{ij}^* \coloneqq k_j$, gdy x_{ij} jest bliższe 1, a gdy zmienna X_j ma charakter ciągły, wtedy w obliczeniach podstawiamy $x_{ij}^* \coloneqq \max\limits_{k=1,\,2,\,\ldots,\,n} x_{kj}$, gdy $x_{kj} \leq \operatorname{med}\limits_{k=1,\,2,\,\ldots,\,n} x_{kj}$ oraz $x_{ij}^* \coloneqq \min\limits_{k=1,\,2,\,\ldots,\,n} x_{kj}$, gdy $x_{kj} > \operatorname{med}\limits_{k=1,\,2,\,\ldots,\,n} x_{kj}$, $j = 1, 2, \ldots, m$. Pozwala to na uzyskanie wyraźnego obrazu powstałych różnic, zarówno w przypadku zmiennych kategorialnych, jak i ciągłych. Oczywiście pewien wpływ na poziom straty informacji w tym kontekście mógłby mieć także charakter stosowanego algorytmu SDC (który decyduje o postaci i zakresie ukryć), jednak ze względu na to, że propozycja ta nie odnosi się doń bezpośrednio, wpływ ten nie powinien być znaczny. Podejście oparte na różnicach jednostkowych jest mniej przydatne do oceny wpływu rozpatrywanej kontroli na wartości zagregowane, m.in. ze względu na konieczność zapewnienia kalibracji.

Koncepcja (1) jest uogólnieniem i normalizacją miar proponowanych przez Domingo-Ferrera i in. (2001). Dla zmiennych ciągłych sugerują oni także miary (niekoniecznie ograniczone od góry) oparte na różnicy między średnimi:

$$\lambda = \frac{1}{m}\sum_{j=1}^{m}\left|\bar{x}_j - \bar{x}_j^*\right| \tag{7}$$

lub

$$\lambda = \frac{1}{m}\sum_{j=1}^{m}\left(\bar{x}_j - \bar{x}_j^*\right)^2 \tag{8}$$

albo

$$\lambda = \frac{1}{m}\sum_{j=1}^{m}\frac{\left|\bar{x}_j - \bar{x}_j^*\right|}{\left|\bar{x}_j\right|}, \tag{9}$$

gdzie $\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n$, a $\bar{x}_j^* = \sum_{i=1}^{n} x_{ij}^*/n$ stanowią średnie arytmetyczne wartości zmiennej X_j odpowiednio przed zastosowaniem i po zastosowaniu SDC, $j = 1, 2, \ldots, m$.

Do miar nieznormalizowanych może należeć też miara oparta na formule (1), w której[2]

$$d\left(x_{ij}, x_{ij}^*\right) = \frac{\left|x_{ij} - x_{ij}^*\right|}{\max\limits_{k=1,\,2,\,\ldots,\,n} x_{kj}}$$

przy założeniu, że zmienna X_j przyjmuje wyłącznie wartości nieujemne, nie wszystkie równe 0, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$.

Z kolei Młodak (2019) zaproponował wykorzystanie jako miary straty informacji miernika kompleksowego wyznaczonego przy użyciu metody TOPSIS (opartej na wzorcu i antywzorcu rozwojowym), którego konstrukcja opiera się na cząstkowych odległościach między obserwacjami dla poszczególnych zmiennych traktowanych jako cechy diagnostyczne.

Warto też wspomnieć, że w przypadku stosowania metody postrandomizacyjnej (ang. *post-randomization method*, PRAM)[3] stratę informacji można obliczać także za pomocą wskaźnika entropii postaci (zob. np. Domingo-Ferrer i in., 2001)

$$EBIL_j = \sum_{i=1}^{n} \mathcal{H}\left(X_j \middle| X_j^* = q\right),$$

gdzie:

$\mathcal{H}\left(X_j \middle| X_j^* = x_{ij}^*\right) = -\sum_{r=1}^{k_j} P\left(X_j = r \middle| X_j^* = q\right) \log P\left(X_j = r \middle| X_j^* = q\right)$,

k_j – liczba kategorii zmiennej X_j,

$\mathbf{P}_j = \left[P\left(X_j = r \middle| X_j^* = q\right)\right]$, $r, q = 1, 2, \ldots, k_j$ – macierz prawdopodobieństw przejść Markowa w PRAM dla tej zmiennej.

Nazwa EBIL pochodzi od anglojęzycznego określenia *entropy-based information loss measur*e, które oznacza miarę straty informacji opartą na entropii. Im wyższa wartość entropii, tym większa jest strata informacji. Szerzej o miarach entropii w kontekście oceny straty informacji na skutek SDC pisze np. Antal (2016).

[2] Miarę tę zaproponowała w dyskusji Karolina Warno.

[3] Jest to probabilistyczna metoda SDC generująca określone zakłócenia. Wartości zmiennych kategorialnych dla pewnych rekordów zostają tutaj zamienione na inne z wykorzystaniem specyficznego mechanizmu probabilistycznego, a konkretnie – macierzy przejść Markowa. PRAM łączy w sobie dodawanie tzw. szumu, ukrywanie danych oraz przekodowywanie (zob. np. Hundepool i in., 2012).

Wszystkie opisane wyżej podejścia mają mocne i słabe strony. Do tych pierwszych należy przede wszystkim pełne odzwierciedlenie odmienności między danymi oryginalnymi a tymi udostępnianymi po przeprowadzeniu SDC, zwłaszcza przez miarę postaci (1). Dzięki bezwzględnym różnicom między poszczególnymi obserwacjami każde ukrycie bądź zniekształcenie informacji staje się istotnym składnikiem straty łącznej. Jest to bardzo użyteczne do oceny skali wprowadzonych ingerencji oraz ich wpływu na rozkłady poszczególnych zmiennych. Jednak w szacowaniu określonych wielkości dla populacji oraz – dla danych wyrażonych na skali ilorazowej – w wyznaczaniu wskaźników strata w tym ujęciu najczęściej będzie przeszacowana. Wynika to z tego, że jednostkowe różnice w procesie sumowania mogą się w określonym stopniu wzajemnie niwelować, wskutek czego faktyczne odchylenia szacunków od ich wielkości, które mogłyby zostać uzyskane, gdyby nie zastosowano SDC, będą niezbyt duże. Podobnie rzecz ma się w przypadku wskaźników: analogiczne straty jednostkowe dla obu zmiennych stanowiących podstawę wyznaczania wskaźnika powodują, że wartości tego wskaźnika będą bardzo bliskie wartościom, które można by otrzymać przy użyciu danych oryginalnych. Dobrze to widać choćby na przykładzie miar (7)–(9), które w pewnym stopniu ten problem redukują (mimo istotnych odchyleń indywidualnych średnie mogą różnić się od siebie tylko nieznacznie), ale nie dają zbyt dużo informacji o wpływie straty na jakość estymacji.

Inny problem w stosowaniu tych miar dotyczy braków danych. Wiele metod SDC (np. w ramach mikroagregacji) może wypełniać takie luki w mikrodanych informacjami przez siebie generowanymi, co w istocie stanowi pewną formę imputacji. Pominięcie tego rodzaju sytuacji w ogólnym rozrachunku straty może okazać się zbyt kosztowne. Najlepszym rozwiązaniem byłoby zatem przyjęcie takich oryginalnych wartości rozpatrywanych cech, aby odpowiednie jednostkowe straty $d(x_{ij}, x_{ij}^*)$ były możliwie największe (minimalizuje to ryzyko utraty istotnej wiedzy w tym zakresie). Będą one zależeć od skali pomiarowej, na której wyrażone są dane dotyczące rozpatrywanej zmiennej. I tak:

- dla danych wyrażonych na skali nominalnej: oryginalna kategoria x_{ij} powinna być inna niż otrzymana w SDC (x_{ij}^*);
- dla danych wyrażonych na skali porządkowej: oryginalna kategoria x_{ij} powinna być odległa od nałożonej w SDC (x_{ij}^*) o maksymalną możliwą liczbę kategorii w myśl danej kategoryzacji;
- dla danych wyrażonych na skalach różnicowej lub ilorazowej: jeśli wartość otrzymana z SDC (x_{ij}^*) jest bliższa maksymalnej wartości zmiennej dla danych wejściowych, to jako oryginalną wielkość x_{ij} przyjmujemy minimum faktycznych zgromadzonych wartości zmiennej ($\min_{k \in Z_j} x_{kj}$), w przeciwnym razie – maksimum

$(\max_{k \in Z_j} x_{kj})$, gdzie $Z_j \subseteq \{1, 2, \ldots, n\}$ jest zbiorem tych jednostek, dla których otrzymano dane z zakresu zmiennej X_j.

Podobnie postępujemy w przypadku, gdy znamy dane oryginalne. W SDC dla mikrodanych zastosowano metody niezakłóceniowe, wskutek czego dane wrażliwe zostały ukryte. Wtedy musimy przyjąć robocze wartości odpowiednich danych po przeprowadzeniu SDC, które umożliwią otrzymanie możliwie największych indywidualnych strat $d(x_{ij}, x_{ij}^*)$. Oznacza to, że:

- dla danych wyrażonych na skali nominalnej: robocza kategoria x_{ij}^* w danych po przeprowadzeniu SDC powinna być inna niż oryginalna (x_{ij});
- dla danych wyrażonych na skali porządkowej: robocza kategoria x_{ij}^* w danych po przeprowadzeniu SDC powinna być odległa od oryginalnej (x_{ij}) o maksymalną możliwą liczbę kategorii w myśl danej kategoryzacji;
- dla danych wyrażonych na skali różnicowej lub ilorazowej: jeśli wartość oryginalna jest bliższa maksymalnej wartości zmiennej dla danych wejściowych, to jako roboczą wielkość x_{ij}^* po SDC przyjmujemy minimum faktycznych zgromadzonych wartości zmiennej $(\min_{k \in Z_j} x_{kj})$, w przeciwnym razie – maksimum $(\max_{k \in Z_j} x_{kj})$.

Z kolei gdy stosujemy SDC dla tablic, to w przypadku metod opartych na ukrywaniu komórek trzeba do porównań stworzyć tablicę, w której dla brakujących komórek ich wartości zostaną zaimputowane. W przypadku ukrywania komórek strata informacji zostanie wyrażona jako suma kosztów poniesionych wskutek wtórnego ukrycia komórek. Problemem pozostaje to, czy waga każdej komórki w tablicy jest taka sama, czy też komórki o wyższej wartości mają większą wagę. W praktyce ukrycie zbyt dużej liczby komórek o wysokich wartościach może znacznie obniżyć użyteczność publikowanych danych. W zależności od preferencji i potrzeb użytkowników zagadnienie straty informacji może być wyrażone różnie. W ten sposób da się wpłynąć na działanie algorytmu wyboru komórek do wtórnego ukrycia. Hundepool i współpracownicy (2012) wskazują najpopularniejsze kryteria brane pod uwagę przy formułowaniu funkcji kosztu dla ukrywania komórek:

- jednakowa waga dla wszystkich komórek, której celem jest minimalizacja liczby wtórnie ukrytych komórek;
- liczba jednostek w agregacie, który komórka reprezentuje, prowadząca do poszukiwania możliwości ukrycia tylko takich komórek, które łącznie będą reprezentować jak najmniejszą liczbę jednostek;
- wartość komórki, w przypadku której optymalnym rozwiązaniem będzie pozostawienie jak największej liczby komórek o najwyższych wartościach.

W sytuacji występowania silnej asymetrii danych preferowanie zachowania komórek o najwyższej wartości może prowadzić do zbyt dużej nierównowagi dla funkcji kosztu. W tym wypadku zalecana jest transformacja funkcji kosztu. Jednym

z możliwych i często stosowanych podejść w tym zakresie jest transformacja potęgowa (zob. Box i Cox, 1964):

$$y = \begin{cases} x^{\lambda} & \text{dla } \lambda \neq 0, \\ \log(x) & \text{dla } \lambda = 0. \end{cases}$$

W ochronie tablicy, w której wartości komórek są modyfikowane metodami zakłóceniowymi, proponuje się miary straty oparte na odległości między zmienionymi a pierwotnymi wartościami komórek. Dla układu tablica pierwotna – tablica zmieniona strata informacji będzie mierzona sumą odległości pomiędzy wartościami komórek zmienionych i pierwotnych; odległości te są wyznaczane według formuły (4) lub (5) albo (6) bądź podobnej. Braki danych mają wtedy mniejsze znaczenie – podczas wyznaczania wartości w komórkach można je pominąć lub wcześniej dokonać imputacji. Problem może występować jedynie w przypadku braku jakichkolwiek danych dotyczących kategorii wyznaczonej przez daną komórkę. Wtedy trzeba albo zrezygnować z przyjętej konstrukcji tablicy (np. poprzez połączenie pewnych jej kategorii w inną), albo też miarę straty informacji oprzeć na dokonanym w opisany wcześniej sposób pomiarze straty na poziomie mikrodanych odpowiadających tej komórce.

4. Miary wpływu na wariancję szacunków

Druga grupa miar straty informacji jest oparta na wpływie zmian dokonywanych w efekcie użycia metod SDC na zmienność rozpatrywanych wielkości statystycznych. Wpływ ten można ocenić np. za pomocą wieloczynnikowej analizy wariancji ANOVA. Jednak ma ona dość ograniczony zakres stosowania, ponieważ opiera się na podziale jednostek na grupy w myśl określonej klasyfikacji i badaniu zmienności w tych grupach. W SDC podziału na grupy dokonuje się przede wszystkim w przypadku zastosowania mikroagregacji.

Przykład takiego podejścia prezentują Mateo-Sanz i Domingo-Ferrer (1998). Zakładamy zgodnie z nim, że zbiór n jednostek został podzielony na p (gdzie p jest liczbą naturalną, $p < n$) grup $G_1, G_2, \dots, G_p$, z których każda liczy n_l jednostek, n_l to liczba naturalna, $l = 1, 2, \dots, p$, $\sum_{l=1}^{p} n_l = n$. Każda jednostka i opisana jest zatem przez wektor $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})$, $i = 1, 2, \dots, n$. Zgodnie z regułą minimalnej liczby respondentów przyjmuje się też, że $n_l \geq k$, $l = 1, 2, \dots, p$, gdzie liczba naturalna k, $k < n$, oznacza arbitralnie ustaloną minimalną liczbę jednostek, które mogą należeć do każdej grupy.

Wewnątrzgrupowa suma kwadratów SSE (ang. *sum of squared errors of all observations vs respective means* – suma kwadratów błędów dla wszystkich obserwacji względem ich odpowiednich średnich grupowych) jest wtedy dana wzorem

$$SSE = \sum_{l=1}^{p} \sum_{i\in\{1,\,2,\,\ldots,\,n\},\, i\in G_l} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_l)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_l)^T$$

gdzie $\overline{\boldsymbol{x}}_l = \sum_{i\in\{1,\,2,\,\ldots,\,n\},\, i\in G_l} \boldsymbol{x}_i / n_l$ – wektor średnich arytmetycznych badanych zmiennych dla grupy G_l, $l = 1, 2, \ldots, p$.

Międzygrupowa suma kwadratów SSA (ang. *sum of squared errors of all treatment means vs grand mean* – suma kwadratów błędów dla wszystkich średnich grupowych w odniesieniu do średniej globalnej) to

$$SSA = \sum_{l=1}^{p} n_l(\overline{\boldsymbol{x}}_l - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}_l - \overline{\boldsymbol{x}})^{\mathrm{T}},$$

gdzie $\overline{\boldsymbol{x}} = \sum_{i=1}^{n} \boldsymbol{x}_i / n$ – wektor średnich arytmetycznych badanych zmiennych ogółem.

Suma kwadratów ogółem $TSS = SSA + SSE$ (TSS – ang. *total sum of squares*) ma zatem postać

$$TSS = \sum_{l=1}^{p} \sum_{i\in\{1,\,2,\,\ldots,\,n\},\, i\in G_l} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T.$$

Miara straty informacji stanowi wówczas udział wewnątrzgrupowej sumy kwadratów SSE w sumie kwadratów ogółem, czyli

$$\lambda = \frac{SSE}{TSS} = 1 - \frac{SSA}{TSS}. \tag{10}$$

Miara dana wzorem (10) przyjmuje wartości z przedziału $[0, 1]$. Im wyższa wartość, tym poważniejsza strata informacji, ponieważ zastąpienie faktycznych wartości dla jednostek należących do danej grupy przez stosowną średnią arytmetyczną zmniejsza zróżnicowanie wewnątrzgrupowe. Efektywne zastosowanie SDC zmierza do minimalizacji tej redukcji. Miara (10) może być użyta tylko do zmiennych wyrażonych na skali różnicowej lub ilorazowej.

Oceny wpływu metod SDC na zmienność rozpatrywanych informacji można też dokonywać, opierając się na porównaniu wariancji/kowariancji badanych zmien-

nych przed i po SDC. Domingo-Ferrer i współpracownicy (2001) podają następujące przykłady takich miar:

- miary oparte na macierzy kowariancji zmiennych $X_1, X_2, \dots, X_m$ lub tylko na ich elementach diagonalnych:
  - o $\lambda = 2\sum_{l=1}^{m}\sum_{j:1\le j\le l}\left|v_{jl} - v_{jl}^*\right|/(m(m+1))$,
  - o $\lambda = 2\sum_{l=1}^{m}\sum_{j:1\le j\le l}\left(v_{jl} - v_{jl}^*\right)^2/(m(m+1))$,
  - o $\lambda = 2\sum_{l=1}^{m}\sum_{j:1\le j\le l}\left|v_{jl} - v_{jl}^*\right|/(\left|v_{jl}\right|\cdot m(m+1))$,
  - o $\lambda = \sum_{j=1}^{m}\left|v_{jj} - v_{jj}^*\right|/m$,
  - o $\lambda = \sum_{j=1}^{m}\left(v_{jj} - v_{jj}^*\right)^2/m$,
  - o $\lambda = \sum_{j=1}^{m}\left|v_{jj} - v_{jj}^*\right|/(\left|v_{jj}\right|\cdot m)$,

  gdzie v_{jl} i v_{jl}^* to kowariancja zmiennych X_j i X_l przed i po SDC, $j, l = 1, 2, \dots, m$;
- miary oparte na macierzy korelacji zmiennych $X_1, X_2, \dots, X_m$
  - o $\lambda = 2\sum_{l=1}^{m}\sum_{j:1\le j\le l}\left|\rho_{jl} - \rho_{jl}^*\right|/(m(m+1))$,
  - o $\lambda = 2\sum_{l=1}^{m}\sum_{j:1\le j\le l}\left(\rho_{jl} - \rho_{jl}^*\right)^2/(m(m+1))$,
  - o $\lambda = 2\sum_{l=1}^{m}\sum_{j:1\le j\le l}\left|\rho_{jl} - \rho_{jl}^*\right|/(\left|\rho_{jl}\right|\cdot m(m+1))$,

  gdzie ρ_{jl} i ρ_{jl}^* to współczynniki korelacji zmiennych X_j i X_l przed i po SDC, $j, l = 1, 2, \dots, m$.

Można także zaproponować znormalizowaną na $[0, 1]$ miarę straty na wariancji w postaci

$$\lambda = \frac{2}{\pi m}\sum_{j=1}^{m}\operatorname{arctg}\left|v_{jj} - v_{jj}^*\right|.$$

Warto zauważyć, że miary oparte na macierzy korelacji zmiennych – zasadniczo przeznaczone do stosowania dla zmiennych ciągłych o ilorazowej lub różnicowej skali pomiaru – mogą zostać użyte także do danych wyrażonych na skali porządkowej. W tym celu należy oprzeć je na współczynniku korelacji τ-Kendalla. Ponadto współczynniki bazujące na kowariancji i korelacji odnoszą się w znacznej mierze także do związków między określonymi zjawiskami, zatem mogą być zaliczone również do miar wpływu na siłę związku.

5. Miary wpływu na siłę związku

Dla praktycznych skutków zastosowania SDC bardzo istotne jest zachowanie kierunków i siły związków między badanymi zjawiskami, które odzwierciedlają zebrane dane, a przynajmniej możliwie najmniejszy ubytek w tym zakresie. Tylko wtedy bowiem posługiwanie się takimi danymi ma sens i istnieje szansa na adekwatność

wniosków sformułowanych na podstawie analizy udostępnionych danych w stosunku do faktycznych realiów. Stratę informacji pod tym względem można oceniać rozmaicie. Najbardziej oczywistym narzędziem służącym do tej oceny wydaje się współczynnik korelacji, np. współczynnik korelacji liniowej Pearsona. Jednak ma on zastosowanie tylko do danych wyrażonych na skali różnicowej lub ilorazowej, a w przypadku występowania zależności innego typu niż liniowa może nie być dostatecznie użyteczny. Dlatego warto w tym kontekście rozważyć także alternatywne użycie współczynnika korelacji τ-Kendalla.

Dokładny kształt miar straty informacji opartych na współczynniku korelacji może być taki sam, jak przedstawiono w poprzedniej części artykułu. Oznacza to, że rzeczone miary mogą stanowić funkcję wartości bezwzględnych lub kwadratów różnic między współczynnikami korelacji odpowiednich zmiennych przed i po SDC. Inna możliwość to analiza macierzy odwrotnych do macierzy korelacji: wejściowej $\mathbf{R} = \left[\rho_{jl}\right]$ i po dokonaniu SDC $\mathbf{R}^* = \left[\rho_{jl}^*\right]$. Macierze odwrotne to odpowiednio $\mathbf{R}^{-1} = \left[\rho_{jl}^{(-1)}\right]$ i $(\mathbf{R}^*)^{-1} = \left[\rho_{jl}^{*(-1)}\right]$, $j, l = 1, 2, \ldots, m$. Diagonalne elementy każdej z takich macierzy $\rho_{jj}^{(-1)}$ i $\rho_{jj}^{*(-1)}$, $j = 1, 2, \ldots, m$, odpowiednio, należą do przedziału $[1, \infty)$ i pokazują siłę związku informacyjnego odpowiedniej zmiennej z pozostałymi, z uwzględnieniem także powiązań nieuchwytnych formalnie. Tym samym suma wartości bezwzględnych różnic między tymi elementami może być dobrym miernikiem straty informacji:

$$\lambda = \sum_{j=1}^{m} \left| \rho_{jj}^{(-1)} - \rho_{jj}^{*(-1)} \right|. \tag{11}$$

Można też oczywiście rozważać taki miernik w postaci znormalizowanej, opartej na sferze jednostkowej i odległości euklidesowej:

$$\lambda = \frac{1}{2} \sqrt{\sum_{j=1}^{m} \left(\frac{\rho_{jj}^{(-1)}}{\sqrt{\sum_{l=1}^{m} \left(\rho_{ll}^{(-1)}\right)^2}} - \frac{\rho_{jj}^{*(-1)}}{\sqrt{\sum_{l=1}^{m} \left(\rho_{ll}^{*(-1)}\right)^2}} \right)^2} \in [0, 1]. \tag{12}$$

W każdym z tych przypadków wyższa wartość miernika świadczy o większej stracie informacyjnej dotyczącej związków między obserwowanymi zjawiskami.

Inne podejście, szczególnie przydatne w przypadku zmiennych kategorialnych, polega na konstrukcji tablic kontyngencji dotyczących porównywanych zmiennych.

Na podstawie tych tablic wykonuje się test niezależności pomiędzy odpowiednimi zmiennymi. Test niezależności dla tablicy dwuwymiarowej tego rodzaju jest oparty na współczynniku zgodności chi-kwadrat pomiędzy wartościami obserwowanymi a teoretycznymi. Alternatywnie można tu skorzystać z testu ilorazu wiarygodności chi-kwadrat lub testu Mantela-Haenszla. Miarę związku ocenia współczynnik V Cramera, ewentualnie współczynnik φ czy współczynnik kontyngencji Pearsona. Ocena straty opiera się na względnej, procentowej różnicy pomiędzy wartościami danego współczynnika obliczonymi dla tablicy źródłowej oraz dla tablicy opracowanej na podstawie danych, dla których zastosowano ochronę poufności. W przypadku tablic wielowymiarowych zależności warunkowe oraz wartości teoretyczne można wyrazić za pomocą modeli log-liniowych.

Kolejne możliwości w tym kierunku rodzi porównywanie parametrów strukturalnych (tzn. parametrów funkcji regresji) oraz parametrów struktury stochastycznej (czyli cech rozkładu czynnika losowego) wraz z ocenami ich jakości w odpowiednich modelach ekonometrycznych. Wówczas można uzyskać obraz potencjalnego zniekształcenia informacyjnego wnioskowania o współzależności zjawisk, powstałego na skutek zastosowania SDC.

6. Przykład zastosowania miar straty informacji

Załóżmy, że w wyniku pewnego badania zgromadzono dane o 25 pracujących osobach. Dane te dotyczą następujących zmiennych:

- *STC* – stan cywilny prawny: 1 – kawaler/panna, 2 – żonaty/zamężna, 3 – wdowiec/wdowa, 4 – rozwiedziony/rozwiedziona, 9 – nieustalony;
- *WYN* – wynagrodzenie miesięczne brutto w złotych;
- *STAZ* – staż pracy w latach;
- *ODL* – odległość od miejsca zamieszkania do miejsca pracy w kilometrach.

Są to informacje bardzo wrażliwe, a zatem konieczna jest ich ochrona. W tabl. 1 ukazano oryginalną postać danych oraz ich kształt po zastosowaniu jednej z metod SDC, a mianowicie wymiany rang[4]. Wybór tego właśnie narzędzia motywowany był możliwością jego stosowania do zmiennych zarówno kategorialnych, jak i ciągłych, a oba te typy są reprezentowane w analizowanym zbiorze.

[4] Jest to zakłóceniowa metoda SDC, którą można stosować do danych wyrażonych zarówno na skali porządkowej, jak i na skalach silniejszych. Polega ona na uporządkowaniu wartości zmiennej X w kolejności rosnącej. Następnie każda zrangowana w ten sposób wartość zmiennej X jest zamieniana z inną wartością, losowo wybraną spośród tych wartości, których rangi zawierają się w pewnym ograniczonym przedziale – np. spośród tych, których rangi nie różnią się od rangi danej wartości więcej niż o p% całkowitej liczby rekordów, gdzie $p \in (0, 100)$ jest ustalonym parametrem. Zob. np. Hundepool i współpracownicy (2012).

Tabl. 1. Dane oryginalne i dane po wymianie rangowej

ID	*STC_O*	*WYN_O*	*STAZ_O*	*ODL_O*	*STC_R*	*WYN_R*	*STAZ_R*	*ODL_R*
1	2	2852,34	5	2	2	2258,33	8	3
2	2	3927,55	7	5	2	3263,22	3	3
3	1	2258,33	4	8	1	2852,34	8	10
4	2	2594,17	8	1	2	2074,88	4	2
5	3	3263,22	10	4	4	3927,55	11	6
6	1	2965,84	11	3	1	3552,11	10	4
7	2	3552,11	3	7	1	2965,84	7	10
8	1	3147,53	19	2	1	3475,12	17	1
9	1	2074,88	9	9	1	2594,17	13	12
10	1	3475,12	6	12	1	3147,53	4	9
11	2	4021,44	21	3	2	4410,82	16	5
12	4	2384,65	15	10	4	2204,59	10	8
13	2	5213,51	17	2	2	4215,39	19	1
14	3	4100,89	8	3	4	4315,27	12	5
15	4	2003,77	4	5	2	2551,28	6	3
16	4	2551,28	12	15	3	2003,77	8	15
17	1	3571,19	16	11	1	3128,73	21	8
18	2	3128,73	23	1	4	3571,19	17	2
19	1	4215,39	28	8	2	5213,51	28	11
20	3	4521,76	13	3	3	4056,83	9	2
21	2	4410,82	10	1	2	4021,44	15	2
22	2	6017,94	25	6	2	6017,94	18	4
23	2	4056,83	17	10	2	4521,76	23	7
24	4	2204,59	8	4	3	2384,65	5	3
25	3	4315,27	18	2	3	4100,89	25	1

Uwaga. xxxx_O – wartości oryginalne zmiennej *xxxx* przed wymianą rangową, xxxx_R – wartości zmiennej *xxxx* po wymianie rangowej.

Źródło: opracowanie własne z wykorzystaniem programu μ-Argus 5.1 (dane fikcyjne).

Zastosujemy najpierw miarę zakłócenia rozkładu (1). W przypadku zmiennej *STC* – jako zmiennej wyrażonej na skali nominalnej – zastosowano odległość cząstkową (2). Pozostałe zmienne: *WYN*, *STAZ* i *ODL* to zmienne ciągłe. Wykorzystano tutaj trzy warianty pomiaru odległości: normalizację z użyciem wartości maksymalnej odchyleń bezwzględnych (modułów) (4), normalizację bazującą na maksimum kwadratów odchyleń (5) i znormalizowaną funkcję arcus tangens odchyleń (6). W tabl. 2 uwidoczniono – dla celów poglądowych – podstawowe statystyki opisowe sum odchyleń wartości po przeprowadzeniu SDC od oryginalnych dla tych trzech zmiennych i każdego wariantu według rekordów oraz wartości miernika straty informacji w każdym z tych trzech przypadków (przy czym miernik uwzględnia także odchylenia dla zmiennej *STC*).

Tabl. 2. Sumy miar odległości dla zmiennych ciągłych oraz wartości miernika straty informacji

Statystyki opisowe dla sum odchyleń	Wariant pomiaru odchyleń dla zmiennych ciągłych		
	normalizacja max modułów	normalizacja max kwadratów	arcus tangens
Minimum	0,9423	0,3005	1,6145
Pierwszy kwartyl	1,4250	0,7083	2,2042
Mediana	1,5614	1,0423	2,3934
Średnia arytmetyczna	1,5990	1,0582	2,3456
Trzeci kwartyl	1,8332	1,2140	2,5479
Maksimum	2,3229	2,0000	2,6887
Współczynnik zmienności w %	22,8123	43,5248	11,8531
Współczynnik asymetrii	0,0833	0,4113	–1,1069
Sumaryczny miernik straty informacji	0,4797	0,3446	0,6664

Źródło: opracowanie własne na podstawie danych z tabl. 1.

Zauważmy, że rezultaty uzyskane za pomocą wariantu opartego na funkcji arcus tangens różnią się od pozostałych. Otrzymane w ten sposób wyniki są wyższe niż w przypadku normalizacji opartej na maksymalnym odchyleniu bezwzględnym czy na maksymalnym kwadracie odchyleń. Wynika to z faktu, że zastosowanie funkcji cyklometrycznej nie spłaszcza obrazu skali najbardziej istotnych odchyleń, co ma miejsce w przypadku opcji (4) i (5). Przekłada się to także na kształt rozkładu: sposób z wykorzystaniem funkcji arcus tangens daje wyniki cząstkowe o rozkładzie wyraźnie lewostronnie asymetrycznym, podczas gdy dla innych rozkład jest lekko prawostronnie asymetryczny. Wartość współczynnika zmienności w wariancie z funkcją cyklometryczną okazuje się dość wyraźnie niższa niż w innych przypadkach. Przyczyna tego stanu rzeczy tkwi w różnorodności analizowanych zmiennych, zmienna *WYN* przyjmuje bowiem wartości znacznie wyższego rzędu niż *STAZ* i *ODL*, w konsekwencji czego odchylenia są dla niej – co do wartości bezwzględnej – na ogół wyższe. Tym samym, zważywszy na malejący wzrost wartości funkcji arcus tangens dla coraz większych argumentów, można dojść do wniosku, że zróżnicowanie odległości mierzonych przy jej użyciu jest mniejsze, a końcowa zmienność – niższa.

Z uwagi na naturę wymiany rang wariancje zmiennych nie ulegają zmianie. Jej zastosowanie nie wywiera więc wpływu na wariancję szacunków, przynajmniej pod tym względem. Możliwe jest zaś oddziaływanie na siłę związku między zmiennymi. Aby to ocenić, zastosujemy podejście oparte na odwróconej macierzy korelacji oraz miernikach (11) i (12). Macierze korelacji τ-Kendalla dla zmiennych przed kontrolą i po kontroli dla zmiennych ciągłych mają odpowiednio postaci uwidocznione w tabl. 3.

Tabl. 3. Macierze korelacji τ-Kendalla przed zastosowaniem i po zastosowaniu wymiany rang

Zmienne	*WYN_O*	*STAZ_O*	*ODL_O*
Przed wymianą rang			
WYN_O	1,0000	0,3435	–0,1481
STAZ_O	0,3435	1,0000	–0,1079
ODL_O	–0,1481	–0,1079	1,0000
Po wymianie rang			
WYN_R	1,0000	0,4849	–0,1550
STAZ_R	0,4849	1,0000	–0,0661
ODL_R	–0,1550	–0,0661	1,0000

Źródło: opracowanie własne z wykorzystaniem danych z tabl. 1 oraz programu SAS Enterprise Guide 4.3.

Tablica 4 obrazuje odpowiednie macierze do nich odwrotne.

Tabl. 4. Macierze odwrotne do macierzy korelacji τ-Kendalla przed zastosowaniem i po zastosowaniu wymiany rang

	WYN_O	*STAZ_O*	*ODL_O*
Przed wymianą rang			
WYN_O	1,1500	–0,3810	0,1292
STAZ_O	–0,3810	1,1380	0,0663
ODL_O	0,1292	0,0663	1,0263
Po wymianie rang			
WYN_R	1,3338	–0,6358	0,1647
STAZ_R	–0,6358	1,3075	–0,0121
ODL_R	0,1647	–0,0121	1,0247

Źródło: opracowanie własne z wykorzystaniem danych z tabl. 1.

Tym samym wartość współczynnika straty informacji w postaci „surowej" wynosi 0,3549, a w postaci znormalizowanej – 0,0318. Oznacza to, że zastosowanie wymiany rangowej spowodowało stratę 3,18% informacji o związkach pomiędzy badanymi zmiennymi. Ubytek ten można uznać za niewielki.

7. Podsumowanie

Metodologia kontroli ujawniania danych oferuje wiele różnorodnych narzędzi prowadzących do możliwie najlepszej ochrony danych wrażliwych (czyli wyeliminowania lub absolutnego zminimalizowania ryzyka ich ujawnienia) przy jednoczesnej minimalizacji straty informacji spowodowanej tego rodzaju działaniem. Warto przy tym zauważyć, że zastosowanie takich narzędzi nie tylko skutkuje poniesieniem straty informacji, lecz także zostawia ślad na agregacji danych w przypadku badań pełnych lub na estymacji – i jej jakości – wybranych parametrów populacji w przypadku

badań reprezentacyjnych. Trzeba mieć to na względzie, gdy dane zabezpieczone wspomnianymi metodami udostępniane są do celów naukowych. Zarówno dla gestora, jak i dla użytkownika danych ważne jest, aby pomimo zastosowania metod niezakłóceniowych lub zakłóceniowych na mikrodanych możliwe było uzyskanie agregatów lub ocen parametrów populacji (lub w bardziej szczegółowych przekrojach) tożsamych z tymi, które mogłyby zostać otrzymane na podstawie oryginalnych danych jednostkowych bądź niewiele się od nich różniących. Z jednej strony użytkownik danych chce jak najlepiej poznać badane zjawisko, z drugiej – gestor danych udostępniający określone informacje będzie kojarzony z każdymi wynikami opracowywanymi i publikowanymi przez osoby ze środowiska naukowego. Dlatego informacja o oczekiwanej stracie informacji na skutek zastosowania SDC i metodzie jej wyznaczenia powinna być łatwo dostępna dla użytkownika.

W artykule przedstawiono trzy najważniejsze grupy mierników straty informacji oraz ich mocne i słabe strony. Do tych pierwszych należy przede wszystkim – zwłaszcza w przypadku miar zakłócenia rozkładu – możność dostosowania cząstkowego sposobu pomiaru straty do skali pomiarowej, na której wyrażona jest zmienna. Główną niedogodnością łączącą wszystkie te miary jest problem normalizacji, która zapewniłaby przejrzystą interpretację uzyskanych ocen. Znane z literatury podejścia nie dają wyników należących do przedziału $[0, 1]$, a wykorzystywanie w tym celu ilorazu danego odchylenia i statystyki opisowej odchyleń (np. sumy albo wartości maksymalnej) prowadzi do niewrażliwości lub znikomej wrażliwości na zmiany największych odchyleń. Dlatego w pracy zaproponowano alternatywne rozwiązania, oparte na ograniczonej funkcji cyklometrycznej, jaką jest arcus tangens. W znacznej mierze redukują one wyżej wskazane ułomności. Ponadto można przypuszczać, że wrażliwość funkcji cyklometrycznej na zmiany odchyleń będzie zachowana również w większej skali. Do słabszych stron tego podejścia należy – zrozumiałe w przypadku funkcji ograniczonych na prostej rzeczywistej – znaczne zbliżenie do granicy dla dostatecznie dużych argumentów, co w rozpatrywanych realiach może prowadzić do pewnego przeszacowania straty informacji. Receptą na to mogłoby być np. odpowiednie przeskalowanie funkcji arcus tangens, czyli przyjęcie jako odległości między odpowiednimi wartościami przed i po SDC (y i y^*) wartości $\operatorname{arctg}(c|y - y^*|)$, gdzie $c \in (0, 1)$ jest pewną stałą „przesuwającą" znormalizowane wartości do obszaru, gdzie będą mogły być potencjalnie bardziej zróżnicowane. Wyzwaniem badawczym jest tu jednak takie wskazanie tej stałej c, aby z jednej strony nie dawało ono użytkownikom danych powodu do posądzeń o manipulację, a z drugiej – by oryginalne zróżnicowanie dystansów zostało w jak największym stopniu zachowane. Oprócz tego funkcja arcus tangens jest bardzo zależna od poziomu zróżnicowania poszczególnych zmiennych, dlatego wskazane byłoby najpierw znormalizowanie ich w jednolity sposób.

W przypadku oceny wpływu SDC na siłę związku zaproponowano metodę opartą na diagonalnych elementach odwróconej macierzy korelacji. Umożliwia ona traktowanie zbioru danych jak nierozerwanej całości i uwzględnienie również takich związków, które nie zawsze dadzą się formalnie wyrazić. To jest niewątpliwie jej duża zaleta. Natomiast za wadę można uznać nie zawsze łatwą praktyczną interpretację owych elementów diagonalnych, mimo że finalny wskaźnik daje czytelną informację.

Warto w tym miejscu wspomnieć jeszcze o wpływie na jakość estymacji straty informacji spowodowanej SDC. Skutkiem zastosowania metod niezakłóceniowych (czyli prowadzących albo do ukrycia danych, albo zachowania ich w oryginalnej postaci) do danych jednostkowych może być zmniejszenie szczegółowości (np. poprzez zastąpienie dokładnych wartości zmiennej przedziałami, w jakich one się mieszczą, przejście na słabszą skalę pomiarową czy łączenie kategorii zmiennej) będzie zmniejszenie szczegółowości możliwych do uzyskania agregatów, jak również węższy zakres estymacji punktowej bądź przedziałowej określonych parametrów. Strata informacji na estymacji dla całej populacji będzie wówczas zależeć także od sposobu ustalenia stosownych wag uogólniających lub – w przypadku badania reprezentacyjnego – metody odpowiedniego skalibrowania wag z losowania albo narzędzi imputacyjnych. Z kolei metody zakłóceniowe (gdy wrażliwe wartości zostają jedynie odpowiednio zniekształcone) mają w założeniu pozwalać na uzyskanie agregatów bądź szacunków parametrów populacji niewiele różniących się od estymat bazujących na danych oryginalnych. Jednak w bardziej szczegółowych przekrojach ta bliskość może nie być zachowana, a zapotrzebowanie dotyczy właśnie coraz bardziej szczegółowych informacji statystycznych. Warto też, aby użytkownik – w takim zakresie, w jakim nie stwarza to ryzyka ujawnienia danych wrażliwych – był poinformowany o parametrach zastosowanych metod zakłóceniowych (np. rozkładu, z którego losowane są liczby dodawane do wielkości oryginalnych, tworzące szum). Oczywiście istnieje ryzyko, że nie każdy użytkownik właściwie zrozumie istotę procedur SDC i może oskarżać statystykę o manipulowanie danymi. Zagrożenie mylną interpretacją lub nadinterpretacją określonych informacji można zminimalizować, formułując przekazywany użytkownikowi komunikat w możliwie najbardziej zrozumiały, zwięzły i nieskomplikowany sposób.

Bez względu na zastosowaną metodę kontroli ujawniania mikrodanych konieczne jest każdorazowe dokładne zbadanie jej wpływu na jakość estymacji – na precyzję estymatora, jego obciążenie oraz dokładność. Dla przykładu w książce Biemera i współpracowników (2017) opisano wpływ metod ograniczających ryzyko ujawnienia – m.in. dodawania szumu oraz wymiany rang – na całkowity błąd estymacji. Niemniej kwestia wypracowania efektywnych i w miarę uniwersalnych metod oceny wpływu straty informacji spowodowanej przeprowadzeniem SDC na jakość estyma-

cji danych populacyjnych, w tym zachowania ich określonych własności w szerszej skali, wymaga dalszych pogłębionych studiów i analiz.

Podziękowania

Autor dziękuje Karolinie Warno z Departamentu Programowania i Koordynacji Badań Głównego Urzędu Statystycznego oraz dr. Janowi Kubackiemu z Ośrodka Statystyki Matematycznej Urzędu Statystycznego w Łodzi za cenne sugestie i opinie, które istotnie przyczyniły się do podniesienia jakości tej pracy.

Bibliografia

Antal, L. (2016). *Statistical Disclosure Control for Frequency Tables* [Rozprawa doktorska, University of Manchester]. Pobrane z: https://www.research.manchester.ac.uk/portal/files/54587025/FULL_TEXT.PDF.

Biemer, P. P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, N. C., West, B. T. (2017). *Total Survey Error in Practice*. Hoboken: John Wiley & Sons.

Box, G. E., Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, (26), 211–252.

Domingo-Ferrer, J., Mateo-Sanz, J. M., Torra, V. (2001). *Comparing SDC methods for microdata on the basis of information loss and disclosure risk*. Pre-proceedings of ETK-NTTS (Exchange of Technology and Know-how – New Techniques and Technologies for Statistics), (2), 807–826. Pobrane z: http://neon.vb.cbs.nl/casc/NTTSJosep.pdf.

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

Mateo-Sanz, J. M., Domingo-Ferrer, J. (1998). A Comparative Study of Microaggregation Methods. *Qüestiió*, *22*(3), 511–526. Pobrane z: https://upcommons.upc.edu/bitstream/handle/2099/4090/article.pdf.

Młodak, A. (2019). Wykorzystanie miernika kompleksowego w ocenie straty informacji na skutek kontroli ujawniania mikrodanych. *Przegląd Statystyczny*, *66*(1), 7–26.

Shlomo, N., Young, C. (2006). *Information loss measures for frequency tables*. Monographs of official statistics, Work session on statistical data confidentiality, Geneva, 9–11 November 2005 (s. 277–289). Luxembourg: Office for Official Publications of the European Communities.