DOI: 10.5604/01.3001.0013.8285

PRZEGLĄD STATYSTYCZNY
TOM LXVI – ZESZYT 1 – 2019

Andrzej MŁODAK[1]

Wykorzystanie miernika kompleksowego w ocenie straty informacji na skutek kontroli ujawniania mikrodanych[2]

1. WPROWADZENIE

Jednym ze zjawisk obserwowanych w statystyce jest intensywny popyt na spseudonimizowane mikrodane, które są szczególnie przydatne dla celów naukowo-badawczych. Przez *mikrodane* rozumie się dane jednostkowe zgromadzone w zbiorze, którego zasób informacyjny pochodzi z rejestru administracyjnego lub z badania statystycznego. Opisują one jednostki (zazwyczaj osoby fizyczne lub podmioty gospodarcze), których obsługą w ustalonym przez jego zadania zakresie zajmuje się gestor rejestru lub które były objęte badaniem statystycznym. Pojęcie „spseudonimizowane mikrodane" oznacza zaś zanonimizowane dane jednostkowe, czyli opisujące badane jednostki informacje, z których usunięto bezpośrednie identyfikatory owych jednostek. Owa eliminacja nazywa się *anonimizacją* danych. Przyczyną wspomnianego wzmożonego zapotrzebowania jest fakt, że mikrodane dają większe możliwości informacyjne i analityczne niż zawartość klasycznych publikacji tabelaryczno-graficznych, pozostawiając badaczowi znaczniejszą swobodę sposobu realizacji własnych celów naukowych.

Mimo dokonanej *a priori* anonimizacji, posługiwanie się mikrodanymi w dalszym ciągu może jednak stwarzać ryzyko identyfikacji jednostki lub odtworzenia danych wrażliwych. Ze względu bowiem na – zazwyczaj znaczną – liczbę zmiennych opisujących dane jednostki, może występować bardzo duża liczba możliwych wariantów kombinacji kategorii zmiennych wyrażonych na skali nominalnej lub porządkowej. Istnieje zatem spore ryzyko, że znajdą się wśród nich kombinacje unikatowe, co w konsekwencji pozwoli na identyfikację odpowiadających im jednostek. Ryzyko to może być jeszcze większe, jeśli w bada-

[1] Urząd Statystyczny w Poznaniu, Ośrodek Statystyki Małych Obszarów; Oddział w Kaliszu, pl. J. Kilińskiego 13, 62–800 Kalisz, Polska, e-mail: a.mlodak@stat.gov.pl, Państwowa Wyższa Szkoła Zawodowa im. Prezydenta Stanisława Wojciechowskiego w Kaliszu, Międzywydziałowy Zakład Matematyki i Statystyki, ul. Nowy Świat 4, 62–800 Kalisz, Polska, ORCID: https://orcid.org/0000-0002-6853-9163.

[2] Artykuł stanowi opracowanie wystąpienia wygłoszonego przez autora podczas II Kongresu Statystyki Polskiej z okazji 100-lecia Głównego Urzędu Statystycznego, który odbył się w dniach 10–12 lipca 2018 r. w Warszawie.

niu lub rejestrze gromadzone są informacje mierzone na skali różnicowej (zwanej także przedziałową) lub ilorazowej – a więc ciągłe. Tutaj zakres możliwych wartości jest teoretycznie nieskończony, zatem ta sama wartość rzadko się powtarza, a to zwiększa istotnie zagrożenie identyfikacji jednostki poprzez unikalność takiej wartości, zwłaszcza w powiązaniu z innymi zmiennymi. Ponadto należy wziąć pod uwagę także i to, że potencjalny użytkownik może dysponować również innymi, niezależnymi zasobami danych, które pozwolą mu ową identyfikację ułatwić. Tak więc anonimizacja to tylko pierwszy etap działań zmierzających do skutecznej ochrony wrażliwych informacji statystycznych.

Z uwagi na to, obserwuje się intensywny rozwój kontroli ujawniania danych (ang. *Statistical Disclosure Control*, SDC), czyli metodyki ukrywania lub zniekształcania danych wrażliwych w mikrodanych i tablicach wynikowych, poprzedzającego ich udostępnienie lub publikację. Obejmuje ona szereg zaawansowanych metod (takich jak zaokrąglanie, nakładanie szumu, post-randomizacja, kontrolowane dopasowanie tablic, itp.), opartych na statystyce matematycznej i stosownych rozwiązaniach informatycznych oraz włączonych do metodologii badań statystycznych. Do wiodących w tym zakresie źródeł – ze szczególnym uwzględnieniem mikrodanych – należą m.in. pionierska książka Willenborga, de Waala (1996), opracowania Höningera i innych (2010), Hundepoola i innych (2012), Templa (2017) czy Benschopa i innych (2018).

Zasadnicze cele przeprowadzania SDC są dwa – w dodatku realizowane jednocześnie. Pierwszy z nich stanowi minimalizacja ryzyka ujawnienia wrażliwych kombinacji danych, a tym samym identyfikacji danej jednostki. Przez ryzyko ujawnienia rozumie się tutaj ryzyko bezpośredniego lub pośredniego uzyskania danych wrażliwych objętych ochroną. Ryzyko bezpośrednie wyraża się prawdopodobieństwem jednoznacznego przypisania kombinacji wartości do konkretnej jednostki. W przypadku badań reprezentacyjnych oznacza to, że unikatowa kombinacja w próbie jest jednocześnie taką w populacji (zob. np. Skinner i inni, 1994). Ryzyko pośrednie zaś dotyczy możliwości uzyskania danych wrażliwych poprzez wykorzystanie związków i zależności pomiędzy poszczególnymi informacjami. Drugim celem SDC jest minimalizacja straty informacji na skutek tej ochrony. Pojęcie to oznacza ubytek zasobu informacyjnego zbioru danych statystycznych na skutek ukrycia lub zniekształcenia pewnych zawartych w nim danych spowodowanego zastosowaniem kontroli ujawniania danych. Dodatkowa różnica między ryzykiem a stratą polega na tym, że informacja na temat ryzyka jest poufna i znana tylko osobie zarządzającej danymi oraz przeprowadzającej SDC i przygotowującej dane do ujawnienia. Stanowi ona bowiem ocenę poziomu ochrony informacji. Z kolei wiedza na temat straty informacji winna być ogólnodostępna. Umożliwia ona bowiem użytkownikowi finalnych danych ocenę ich przydatności, a w przypadku ich zastosowania np. w estymacji – uwzględnienie tego aspektu w analizie ogólnej jakości uzyskanych oszacowań.

W prezentowanym artykule skupimy się na ocenie tejże straty. Wszystkie znane obecnie sposoby jej pomiaru mają wszakże jedną wspólną cechę: opierają się na porównaniu zbioru danych oryginalnie uzyskanych z badania statystycznego ze zbiorem powstałym w efekcie zastosowania kontroli ujawniania danych przed ich udostępnieniem lub opublikowaniem. Sposób dokonania tego porównania zależy przede wszystkim od skali pomiarowej, na której wyrażone są rozpatrywane zmienne. W przypadku skali nominalnej opiera się ona na identyczności bądź odmienności dwóch wielkości. Jeśli mamy do czynienia ze skalą porządkową, to bierzemy pod uwagę liczbę kategorii, o którą różnią się obserwacje. Natomiast w przypadku skali różnicowej i ilorazowej strata informacji jest funkcją wartości bezwzględnej różnicy między odpowiednimi wartościami. Istnieją liczne sposoby kompleksowego pomiaru straty informacji oparte na różnorodnych funkcjach dystansu między danymi oryginalnym i poddanymi SDC lub miedzy miarami dyspersji odpowiednich cech (zob. np. Domingo-Ferrer, Mateo-Sanz, Torra, 2001). Jednak nie uwzględniają one na ogół wzajemnych powiązań między zmiennymi pozyskiwanymi w danym badaniu (które zazwyczaj obejmuje *de facto* pewne zjawisko wielowymiarowe). Ich interpretacja także nierzadko nie bywa łatwa (np. w kontekście porównywalności).

Dlatego też obecnie zaprezentujemy oryginalne podejście w tym zakresie oparte na taksonomicznym mierniku kompleksowym. Początki koncepcji mierników tego rodzaju (zwanych także wskaźnikami syntetycznymi, kompleksowymi miarami rozwoju lub metacechami) sięgają prac prof. dra hab. Zdzisława Hellwiga z Uniwersytetu Ekonomicznego we Wrocławiu. Stworzył on konstrukcję takiego miernika opartą na taksonomicznym wzorcu rozwojowym (czyli sztucznym idealnym obiekcie, opisanym przez optymalne wartości każdej z cech diagnostycznych) i odległości od tegoż wzorca (zob. np. Hellwig, 1967, 1968). Następnie przeprowadził on szereg badań w zakresie optymalizacji doboru cech diagnostycznych będących podstawą wyznaczania tegoż miernika (zob. np. Hellwig, 1969, 1972a, 1972b). Później wprowadził on także do swych rozważań pojęcie antywzorca (sztucznego obiektu o najbardziej niekorzystnych wartościach cech diagnostycznych) – zob. np. Hellwig (1981). W podobnym czasie praktycznie analogiczną konstrukcję miernika kompleksowego opartą na odległości obiektów od wzorca i antywzorca zaproponowali Hwang, Yoon (1981). Nazwali ją TOPSIS (ang. *The Technique for Order of Preference by Similarity to Ideal Solution*) – i to określenie funkcjonuje powszechnie w literaturze międzynarodowej. Tym niemniej, polscy autorzy (np. Balcerzak, Pietrzak, 2015) niejednokrotnie podkreślają wcześniejsze osiągnięcia prof. Z. Hellwiga w tym zakresie. Wyniki te były w kolejnych latach przedmiotem wielu dalszych badań, między innymi w zakresie efektywnych metod pomiaru odległości od wzorca i antywzorca (zob. np. Walesiak, 2006), normalizacji zmiennych (zob. Zeliaś, 2002; Młodak, 2006b; Pawełek, 2008; Walesiak, 2014a, 2016, 2018; Kukuła, Luty, 2015), specyfiki skal pomiarowych, na których wyrażone są dane (Walesiak, 2014b), czy też szeregów czasowych i danych panelowych (np. Grabiński, 2017). Pojawiły się także odmiany

mierników kompleksowych zastosowane do analizy zbiorów rozmytych (np. Chen, 2000), grupowego podejmowania decyzji (np. Shih i inni, 2007) czy danych przedziałowych (Młodak, 2014).

Cele pracy są zatem dwa: sformułowanie propozycji użycia unormowanego i łatwo interpretowalnego miernika kompleksowego powiązanych cech opartego na wzorcu i antywzorcu rozwojowym w ocenie straty informacji spowodowanej zastosowaniem technik kontroli ujawniania danych oraz sprawdzenie użyteczności tego podejścia w praktyce.

Artykuł składa się z pięciu zasadniczych części. W części drugiej omówimy najistotniejsze charakterystyczne właściwości SDC dla mikrodanych oraz najważniejsze metody przeprowadzania tejże kontroli w takim przypadku. W części trzeciej zajmiemy się kwestią sposobu pomiaru straty informacji i wyznaczaniem wielkości tejże straty, przedstawiając ogólne założenia i typowe rozwiązania w tym zakresie. Następnie (część czwarta) przejdziemy do prezentacji metody konstrukcji miernika kompleksowego (zwanej TOPSIS), by w części piątej omówić efektywny sposób jej wykorzystania w ocenie straty informacji spowodowanej zastosowaniem kontroli ujawniania danych. Podamy tutaj praktyczny przykład wyznaczania takiej straty dla mikrodanych opisujących pewne aspekty statusu badanych osób na rynku pracy. Całość zwieńczą stosowne wnioski.

2. TECHNIKI BEZPIECZNEGO UJAWNIANIA MIKRODANYCH

Mikrodane zawierają cztery kategorie zmiennych:

- identyfikatory – zmienne, które w sposób jednoznaczny identyfikują respondenta (numery PESEL, NIP, REGON, itp.),
- quasi-identyfikatory (zwane także zmiennymi kluczowymi) – zmienne, niekoniecznie identyfikujące bezpośrednio, których połączenie może jednak zidentyfikować respondenta jednoznacznie (np. nazwisko/nazwa, adres, płeć, wiek, miejscowość zamieszkania/siedziby, itp.),
- poufne zmienne wynikowe – zmienne, które zawierają wrażliwe informacje o respondencie (np. dla osoby: dochód osobisty, wyznanie, preferencje polityczne, stan zdrowia, zaś dla podmiotu gospodarczego: liczba zatrudnionych, przychody ze sprzedaży, wypłacone wynagrodzenia, wartość zawartych umów, itp.),
- zmienne wynikowe nie będące poufnymi – zmienne, które nie zaliczają się do żadnej z powyższych kategorii.

Jak wspomnieliśmy na wstępie, podstawową czynnością w zakresie ochrony poufności mikrodanych jest ich ***anonimizacja***. Pojęcie to oznacza usunięcie ze zbioru mikrodanych zmiennych identyfikatorów oraz tych spośród quasi-identyfikatorów, które w danym układzie stały się identyfikatorami w całej lub w znacznej części rekordów rozpatrywanego zbioru. Anonimizacja może, dla przykładu, polegać na usunięciu imion, nazwisk oraz numerów PESEL ze zbioru zgromadzonych w badaniu statystycznym danych dotyczących osób. Nie gwarantuje to wszakże

zabezpieczenia przed możliwością identyfikacji jednostki na podstawie unikalnych kombinacji wartości innych zmiennych i odtworzenia wrażliwych o niej informacji. Na przykład, jeśli w gminie mieszka tylko jedna pracująca kobieta wykonująca zawód agenta ubezpieczeniowego, to posiadanie danych o gminie zamieszkania, płci, statusie na rynku pracy i wykonywanym zawodzie jednoznacznie ją identyfikuje. Anonimizacja stanowi więc wstępny etap kontroli ujawniania danych i nie można jedynie na niej poprzestać.

Konieczne jest zatem stosowanie bardziej zaawansowanych technik bezpiecznego ujawniania danych. Prezentują je np. Hundepool i inni (2006, 2012). Tutaj przytoczymy tylko kilka najważniejszych spośród nich, które są też i najbardziej popularne.

Metody SDC można podzielić na dwie zasadnicze grupy. Pierwszą z nich stanowi ***maskowanie niezakłóceniowe*** (ang. *non-perturbative masking*). Ich zastosowanie prowadzi do tego, że wrażliwe dane stają się – w różny sposób – niewidoczne dla zewnętrznego użytkownika: w finalnym udostępnianym zbiorze informacja jednostkowa albo figuruje w dokładnej postaci, albo jej nie ma wcale. Do metod niezakłóceniowych należą m.in.:

– podpróbkowanie (ang. *subsampling*) – udostępnianie pewnej próbki rekordów spośród figurujących w bazie danych zgromadzonych w trakcie badania statystycznego, dzięki czemu istnieje duża szansa pominięcia unikatowych rekordów; próbka ta może zostać wylosowana zgodnie ze sztuką badania reprezentacyjnego – na przykład, jeśli w danej gminie mieszka tylko jedna osoba w wieku od 35 do 40 lat z wyższym wykształceniem z zakresu metalurgii, to wylosowanie (np. przy użyciu schematu losowania prostego bez zwracania) podpróbki z dużym prawdopodobieństwem ten rekord pominie. Jeśli zaś do tego losowania wprowadzimy ograniczenie wiekowe (np. tylko osoby w wieku 41 i więcej lat) lub co do poziomu albo kierunku wykształcenia (np. wyższe humanistyczne bądź ekonomiczne), to pominie go na pewno,
– przekodowanie (ang. *recoding*) wrażliwych zmiennych – połączenie kilku kategorii w jedną – bardziej zgrubną i o większej liczbie należących do niej jednostek (dla zmiennej kategorialnej, tzn. wyrażonych na skali nominalnej – jak np. płeć czy gmina zamieszkania – lub porządkowej – np. poziom wykształcenia) bądź też zastąpienie zmiennej ciągłej przez jej odpowiednik w postaci dyskretnej; w pierwszym przypadku przykładem może tu być np. zamiana kategorii wieku 40–45 lat i 46–50 lat w kategorię 40–50 lat, w drugim – zastąpienie dokładnej kwoty miesięcznych dochodów do dyspozycji wskazaniem, do którego z ustalonych przedziałów wartości ona należy,
– lokalne ukrywanie danych (ang. *local suppression*) – polega na usuwaniu pewnych wartości niektórych zmiennych kategorialnych dla konkretnych jednostek celem uniknięcia ich identyfikacji; liczba ukrywanych wartości powinna być przy tym możliwie jak najmniejsza.

Drugą grupą narzędzi SDC jest ***maskowanie zakłóceniowe*** (ang. *perturbative masking*) – zakłócanie wrażliwych wartości zmiennych celem uniemożliwienia dokładnego ich odtworzenia przez nieuprawnionego użytkownika przy jednoczesnej minimalizacji strat informacyjnych. Najbardziej znane przykłady metod zakłóceniowych to:

– dodawanie szumu (ang. *noise addition*) – nakładanie na oryginalne dane wrażliwe specjalnie zdefiniowanych zakłóceń, celem zniekształcenia uniemożliwiającego odtworzenie ich faktycznej postaci, przy minimalizacji negatywnych dla jakości danych dla populacji skutków; szum generowany jest zazwyczaj losowo, np. jako liczba z rozkładu normalnego lub jednostajnego, dodawana do wartości prawdziwej (podejście addytywne) lub będąca jej mnożnikiem (opcja multyplikatywna),
– mikroagregacja (ang. *microaggregation*) – obejmuje ona w istocie pewną rodzinę narzędzi zapewniających ochronę poufności danych w ujęciu makro poprzez zastąpienie wartości indywidualnych odpowiednimi wartościami sumarycznymi (takimi jak np. sumy lub średnie) dla niewielkich poziomów agregacji gdy każdy z nich obejmuje co najmniej k rekordów, a żaden rekord nie dominuje pod danym względem (tzn. jego udział w danej wielkości ogółem dla grupy, do której należy, nie jest większy niż $p\%$, $0 < p < 100$); grupy te można wyodrębniać specjalnie dobranymi metodami analizy skupień, takimi jak np. metoda k-Warda (zob. Mateo-Sanz, Domingo-Ferrer, 1998),
– zaokrąglanie (ang. *rounding*) – mechanizm, w którym oryginalne wartości zastępuje się ich wersjami zaokrąglonymi; wersje te zazwyczaj wybiera się ze zbioru punktów zaokrągleń definiującego zestaw zaokrągleń; najczęściej jako punkty zaokrągleń przyjmuje się wielkości $p_i = b \cdot i$, dla $i = 1, 2, \dots, l$, gdzie liczba naturalna b stanowi podstawę zaokrąglania, zaś l to maksymalny zakres zaokrągleń; dla przykładu[3], niech $b = 10$, $x_{\max}$ oznacza maksymalną wartość badanej zmiennej, a $[x_{\max}/10] = 50$ – wówczas $l = 50$ oraz $p_i = i \cdot 10$, $i = 1, 2, \dots, 50$. Tym samym np. dla wielkości 337,8 zaokrąglenie należeć będzie do przedziału $\langle 34 * 10 - (10/2), 34 * 10 + (10/2) \rangle = \langle 340 - 5, 340 + 5 \rangle = \langle 335, 345 \rangle$, co oznacza, że liczbę tę można zaokrąglić do 340; zaokrąglanie można też czynić losowo, np. 216,5 zaokrąglamy do 215 z prawdopodobieństwem (220–216,5)/5=0,7 lub do 220 z prawdopodobieństwem 1–0,7=0,3,
– metoda postrandomizacyjna (ang. *The Post-Randomization Method*, PRAM) – metoda probabilistyczna, generująca szczególne zakłócenia; wartości zmiennych kategorialnych dla pewnych rekordów zostają tutaj zamienione na inne z wykorzystaniem specyficznego mechanizmu probabilistycznego, a konkretnie – macierzy przejść Markowa, w której prawdopodobieństwa zmiany kategorii w trakcie kontroli ustala się arbitralnie. PRAM łączy w sobie

[3] $[a]$ oznacza część całkowitą liczby rzeczywistej a, czyli największą liczbę całkowitą nie większą od a.

dodawanie szumu, ukrywanie danych oraz przekodowywanie; metoda ta może być stosowana jedynie do zmiennych kategorialnych (zob. np. De Wolf i inni, 1999).

Zaprezentowane powyżej metody cechują się różnorodnymi korzystnymi i niekorzystnymi właściwościami. Pogłębioną ich dyskusję znaleźć można m.in. w książce Hundepoola i innych (2012). Tutaj ograniczymy się do kilku najistotniejszych spostrzeżeń. I tak, podpróbkowanie jest łatwe do przeprowadzenia w każdych warunkach, jednak wymaga dokładnego zaplanowania w zakresie schematu i założeń doboru próbki, aby zminimalizować spowodowane tym obciążenia szacunków dokonywanych na podstawie udostępnionych danych z takiej próbki. Nie zawsze da się też wyeliminować ryzyko wylosowania informacji wrażliwych. Przekodowanie jest skuteczne odnośnie do ochrony wrażliwych informacji, ale może zawężać pole np. estymacji dla małych obszarów (zbyt obszerne poziomy, na których można estymować docelowe zmienne). Z kolei lokalne ukrywanie danych może być optymalnym rozwiązaniem niezakłóceniowym, wymaga jednak często czasochłonnej analizy możliwych wrażliwych kombinacji danych, które dają sposobność odtworzenia informacji chronionych. Metody zakłóceniowe bazują w znacznej mierze na arbitralnym doborze zakłóceń, prawdopodobieństw przejść międzykategorialnych, podstawy zaokrągleń albo narzędzi analizy skupień (co czasem – np. w przypadku mikroagregacji – może prowadzić do nadmiernej redukcji zmienności danej cechy w stosunku do stanu oryginalnego). Z drugiej strony, optymalny dobór parametrów stochastycznych nakładanego szumu czy prawdopodobieństw przejść dla podejścia PRAM w konkretnej sytuacji może ograniczyć powstałą stratę informacji do absolutnego minimum.

3. STRATA INFORMACJI I JEJ WYZNACZANIE

Na skutek stosowania narzędzi kontroli ujawniania danych SDC następuje ubytek zasobu informacyjnego zawartego w zbiorze danych statystycznych poddanych owej kontroli. Ubytek ten nazywa się ***stratą informacji***. Ukrycie lub zniekształcenie faktycznie zebranych informacji skutkuje bowiem zmniejszeniem zakresu dostępnej wiedzy oraz dodatkowym obciążeniem estymacji w przypadku szacunków dokonywanych przez użytkownika na podstawie udostępnionych mikrodanych. Tym samym istotną dlań informacją jest skala wielkości owej straty, umożliwiająca realną ocenę jakości uzyskanych wyników analitycznych.

Pomiar straty informacji opiera się na unormowanych różnicach między odpowiednimi wartościami w zbiorze danych oryginalnych oraz w zbiorze danych zniekształconych z uwzględnieniem skal pomiarowych, na jakich mierzone są poszczególne obserwacje.

Ogólna postać typowej miary straty informacji może być następująca:

$$\lambda = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n} d(x_{ij}, x_{ij}^*)}{mn} \in [0,1],$$

gdzie $d(\cdot,\cdot) \in [0,1]$ jest miarą odległości spełniającą klasyczne warunki zwrotności, symetrii i nierówności trójkąta, x_{ij} oznacza faktyczną wartość j-tej zmiennej dla i-tego obiektu, x_{ij}^* – odpowiednią wartość w zbiorze otrzymanym na skutek przeprowadzenia kontroli ujawniania danych, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, m$. Zerowa wartość miary oznacza brak zmian wprowadzonych przez SDC (co – rzecz jasna – w praktyce nie zachodzi). Im większa wartość λ, tym dokuczliwsza strata informacji. Sytuacja gdy $\lambda = 1$ oznacza całkowitą odmienność obu zbiorów.

Formuła używanej miary odległości $d(\cdot,\cdot)$ zależy od skali pomiarowej, na której mierzone są dane zmienne. Jeżeli wartości zmiennej X_j mierzone są na skali nominalnej, to

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

Gdy zaś wartości X_j mierzone są na skali porządkowej, wówczas

$$d(x_{ij}, x_{ij}^*) = \frac{r(x_{ij}, x_{ij}^*)}{k_j - 1}, \tag{2}$$

gdzie $r(x_{ij}, x_{ij}^*)$ – liczba kategorii zmiennej X_j, o którą różnią się x_{ij} i x_{ij}^*, k_j – liczba kategorii zmiennej X_j ogółem.

Dla zmiennej ciągłej (czyli takiej, której obserwacje mierzone są na skali różnicowej lub ilorazowej) odległość ta może być np. postaci

$$d(x_{ij}, x_{ij}^*) = \frac{|x_{ij} - x_{ij}^*|}{\max\limits_{k=1,2,\dots,n} |x_{kj} - x_{kj}^*|} \tag{3}$$

lub

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

$i = 1, 2, \dots, n$, $j = 1, 2, \dots, m$.

Pewien problem w tym kontekście może stanowić sytuacja, gdy ocenie poddawana jest strategia informacji powstała na skutek zastosowania metod niezakłóceniowych. Przekodowywanie dokonywane jest jedynie na zmiennych kategorialnych (tj. takich, których wartości wyrażone są na skali nominalnej lub po-

rządkowej). W tym przypadku należy zatem zadbać o to, aby numery kategorii pozostawianych bez zmian w obu wariatach były identyczne. Na przykład, jeżeli zmienna X_j przed poddaniem jej przekodowaniu liczyła 6 kategorii (oznaczonych jako 1, 2, 3, 4, 5, 6), zaś w wyniku przekodowania połączono kategorie 1 i 2 oraz 4 i 5, to nowe kategorie winny mieć odpowiednio numery 1, 3, 4 i 6. Wtedy wzór (2) można zastosować i w tym przypadku. Jeżeli natomiast dane są ukrywane, to gdy wartości zmiennej X_j zostały wyrażone na skali nominalnej, wówczas jeśli obserwacja dla jednostki przypisujemy $x_{ij}^*:=1$. Jeżeli obserwacje zmiennej X_j wyrażają się z kolei na skali porządkowej, wtedy ukrytej wartości tejże zmiennej przyporządkowujemy $x_{ij}^*:=1$, gdy x_{ij} jest bliżej k_j niż 1, a w przeciwnym razie kładziemy $x_{ij}^*:=k_j$. Gdy zaś X_j ma charakter ciągły (a zatem jej dane wyrażone są na skali różnicowej lub ilorazowej), wówczas przyporządkowujemy $x_{ij}^*:=\max\limits_{k=1,2,\dots,n} x_{kj}$, gdy $x_{ij}\leq \operatorname{med}\limits_{k=1,2,\dots,n} x_{kj}$ oraz $x_{ij}^*:=\min\limits_{k=1,2,\dots,n} x_{kj}$, gdy $x_{ij}>\operatorname{med}\limits_{k=1,2,\dots,n} x_{kj}$, $i=1,2,\dots,n$, dla każdego $j\in\{1,2,\dots,m\}$. Pozwala to na należyte wyeksponowanie występujących odmienności.

Przedstawione powyżej rodzaje miar straty informacji nazywa się miarami dla zakłócenia rozkładu zmiennych zawartych w zbiorze danych. Oprócz tego można w tym kontekście stosować mierniki wpływu na wariancję szacunków, w konstrukcji których pod uwagę są brane różnice wariancji dla przeciętnych wartości wyjściowych i otrzymanych w wyniku SDC (jeśli wartości zmiennych wyrażają się na skali różnicowej lub ilorazowej) lub jednoczynnikową analizę wariancji ANOVA dla wybranej zmiennej zależnej względem wybranych niezależnych zmiennych kategorialnych. W przypadku ANOVA miarą straty jest porównanie, jak zmieniają się komponenty współczynnika determinacji R^2 (tzn. wariancja wewnątrzgrupowa i międzygrupowa) na skutek zastosowania narzędzi SDC (zob. np. Shlomo, Skinner, 2010). Szeroki zakres metod oceny straty informacji ukazują np. Domingo-Ferrer, Mateo-Sanz, Torra (2001).

Obecnie w literaturze przedmiotu rozwiązania powyższego rodzaju dość często obarczone są różnymi ułomnościami. Na przykład wskaźniki bazujące na sumach czy średnich arytmetycznych różnic pomiędzy danymi oryginalnymi i przekształconymi w wyniku zastosowania SDC są wrażliwe na przypadki odstające. Incydentalnie bardzo duże różnice tego rodzaju nie muszą natomiast znaleźć swego odzwierciedlenia w jakości estymacji informacji przeprowadzanej na podstawie mikrodanych poddanych SDC. Tym samym, użytkownik może otrzymać przeszacowaną ocenę straty informacji. Z kolei gdy wskaźnik opiera się np. na różnicy między wariancjami czy kowariancjami, wówczas istnieje zagrożenie niekorzystnym wpływem odmienności w zakresie rzędu wielkości i zakresu wartości pomiędzy poszczególnymi danymi na kształty szacunków oceny straty informacji. Niektóre z powyższych operacji nie mogą być też wykonywane na danych kategorialnych. Prezentowana propozycja stara się maksymalnie zniwelować te niedogodności.

4. KONSTRUKCJA MIERNIKA KOMPLEKSOWEGO

Jak wspomniano na wstępie, konstrukcja miernika kompleksowego zróżnicowania obiektów pod kątem danego zjawiska złożonego oparta jest na wzorcu (sztucznym obiekcie idealnym o optymalnych wartościach rozpatrywanych cech) i antywzorcu rozwojowym (obiekcie o wartościach najbardziej niekorzystnych). Jej koncepcja ma swe źródło w pracach Hellwiga (1967, 1968, 1981), który zainicjował badania w zakresie wyznaczania wskaźników syntetycznych, kontynuowane później przez szereg badaczy (zob. np. Lira i inni, 2002; Malina, 2002; Młodak, 2006a, 2014; Kukuła, Luty, 2015 czy Walesiak, 2018). Za sprawą Hwanga, Yoona (1981) – którzy prawdopodobnie osiągnęli swoje wyniki niezależnie od dokonań prof. Z. Hellwiga[4] – podejście to nosi nazwę TOPSIS (ang. *The Technique for Order of Preference by Similarity to Ideal Solution* – technika porządkowania preferencji wedle podobieństwa do idealnego rozwiązania).

Podstawowe założenia konstrukcji mierników tego rodzaju są uniwersalne. Przyjmuje się mianowicie, że każdy obiekt jest opisany przez pewną liczbę zmiennych charakteryzujących rozpatrywane zjawisko złożone. Zmienne te dobiera się według siedmiu kluczowych zasad (zob. np. Śmiłowska, 1997; Młodak, 2006a):

– istotności z punktu widzenia analizowanych zjawisk,
– jednoznaczności i precyzyjności zdefiniowania,
– wyczerpania zakresu zjawiska,
– logiczności wzajemnych powiązań,
– zachowania proporcjonalności reprezentacji zjawisk cząstkowych,
– mierzalności – w sensie możliwości liczbowego wyrażenia poziomu cechy,
– dostępności i kompletności informacji statystycznych (dla wszystkich badanych obiektów).

Zaleca się też aby zmienne miały charakter niwelujący pewne naturalne dysproporcje między obiektami, a zatem posiadały formę wskaźnikową. Na wstępie poddaje się je weryfikacji zmiennościowo-korelacyjnej czyli eliminuje się zmienne o zbyt niskiej zmienności (a zatem nie mające istotnej siły różnicującej) oraz nazbyt skorelowane z innymi (czyli będące nośnikami podobnej informacji jak zawarta w modelu) – zob. np. Śmiłowska (1997), Młodak (2006a). W efekcie uzyskuje się zestaw cech diagnostycznych, które będą podstawą konstrukcji miernika. Załóżmy, że cech diagnostycznych jest m (gdzie m to liczba naturalna). Zatem obiekt i-ty reprezentowany jest przez wektor $\boldsymbol{\gamma}_i = (x_{i1}, x_{i2}, \ldots, x_{im})$, gdzie x_{ij} to wartość j-tej cechy diagnostycznej dla tegoż i-tego obiektu, $i = 1, 2, \ldots, n$. Cechy diagnostyczne mogą być stymulantami (czyli zmiennymi, których wyższa wartość świadczy o lepszej pozycji obiektu z punktu widzenia badanego zjawiska), destymulantami (im wyższe wartości, tym kondycja obiektu pod rozpatrywanym względem jest gorsza) lub nominantami (posiadającymi wartość

[4] Wskazywać na to może brak prac prof. Z. Hellwiga w bibliografii tej pozycji.

optymalną, poniżej której mają charakter stymulanty, a powyżej – destymulanty lub na odwrót). Celem ujednolicenia charakteru cech destymulanty i nominanty zamienia się w stymulanty, na przykład za pomocą przyjęcia odpowiednich wartości ze znakiem przeciwnym.

Dla uzyskania jednolitości mian (a czasem i zakresu wartości) cechy diagnostyczne poddaje się normalizacji, takiej jak np. standaryzacja: $z_{ij} = (x_{ij} - \bar{x}_j)/s_j$, przy czym z_{ij} to wartość znormalizowanej j-tej cechy diagnostycznej dla i-tego obiektu $i = 1, 2, \dots, n$, $j = 1, 2, \dots, m$. Formuł normalizacyjnych jest wiele, niektóre z nich wykorzystują także punkty osobliwe w przestrzeni wielowymiarowej (zob. Zeliaś, 2002; Młodak, 2006b).

Kluczową czynność metody TOPSIS stanowi zdefiniowanie – na podstawie zestymulowanych i znormalizowanych cech diagnostycznych – wzorca i antywzorca rozwojowego. Wzorzec określa się najczęściej jako sztuczny obiekt, opisany przez maksymalne wartości cech diagnostycznych:

$$\varphi_j^{(+)} = \max_{i=1,2,\dots,n} z_{ij},$$

zaś antywzorzec – jako obiekt wyznaczony przez ich wartości minimalne:

$$\varphi_j^{(-)} = \min_{i=1,2,\dots,n} z_{ij},$$

$j = 1, 2, \dots, m$. W razie potrzeby wzorzec bądź antywzorzec można też ustalić arbitralnie na podstawie pewnych przyjętych norm lub zaleceń (np. standardy ustalone odpowiednimi przepisami bądź wytycznymi międzynarodowymi). Następnie oblicza się odległości poszczególnych obiektów od wzorca ($\delta_i^{(+)}$) i antywzorca ($\delta_i^{(-)}$). Czyni się to np. przy użyciu formuły euklidesowej, w wyniku czego otrzymujemy

$$\delta_i^{(+)} = \sqrt{\sum_{j=1}^{m} \left(z_{ij} - \varphi_j^{(+)}\right)^2} \text{ oraz } \delta_i^{(-)} = \sqrt{\sum_{j=1}^{m} \left(z_{ij} - \varphi_j^{(-)}\right)^2}.$$

$i = 1, 2, \dots, n$. Miarę kompleksową obliczamy jako iloraz odległości obiektów od antywzorca i sumy odległości od obu tych szczególnych obiektów, czyli stosując formułę:

$$\eta_i = \frac{\delta_i^{(-)}}{\delta_i^{(-)} + \delta_i^{(+)}} \in [0,1], i = 1,2, \dots, n. \tag{4}$$

Zerowa wartość miernika oznacza najgorszą możliwą sytuację obiektu, tzn. identyczność z antywzorcem. Wartość 1 osiąga miernik w sytuacji, gdy dany obiekt jest najlepiej rozwinięty pod badanym względem, czyli gdy okazuje się tożsamy ze wzorcem. Im wyższa wartość miernika, tym sytuacja obiektu lepsza.

5. ZASTOSOWANIE MIERNIKA KOMPLEKSOWEGO W SDC

Miernik kompleksowy można zastosować w kontroli ujawniania danych właśnie do oceny straty informacji powstałej na skutek tejże kontroli, szczególnie w kontekście miar dla zakłócenia rozkładu. Załóżmy zatem, że mamy dwa zbiory danych badania rozpatrywanego zjawiska: wyjściowy $\mathbf{X} = [x_{ij}]$ i po poddaniu go SDC $\mathbf{X}^* = [x_{ij}^*]$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Dla każdej zmiennej X_j wyznaczamy odległości $d(x_{ij}, x_{ij}^*)$ między odpowiednimi obserwacjami według formuły zależnej od skali pomiarowej owej zmiennej (a zatem na podstawie wzorów (1), (2) lub (3)), $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Następnie definiujemy cechy diagnostyczne $Y_1, Y_2, \ldots, Y_m$ w postaci $y_{ij} = d(x_{ij}, x_{ij}^*)$, gdzie y_{ij} oznacza wartość cechy diagnostycznej Y_j dla i-tej jednostki, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$.

Ponieważ badanie dotyczy z reguły zjawiska złożonego, a zmienne obejmują ściśle powiązane z przedmiotem badania i ze sobą kwestie, więc można uznać, że istnieją w tym przypadku formalne przesłanki do zastosowania metody taksonomicznej. W kontekście opisanej w części 4 konstrukcji miernika kompleksowego metodą opartą na idei Z. Hellwiga, a zwaną TOPSIS, mamy jednak w tym przypadku sytuację dość szczególną. Polega ona mianowicie na tym, że trzeba tutaj zrezygnować z etapu weryfikacji zmiennych wejściowych. Rzecz leży bowiem w tym, że do rzetelnej oceny straty informacji potrzebne są dane dotyczące zmian w zasobie informacyjnym dla wszystkich obserwacji. Po drugie, z uwagi na założone *a priori* unormowanie odległości $d(\cdot,\cdot)$ na przedziale $[0,1]$ jakakolwiek inna normalizacja jest zbędna. Po trzecie wreszcie, w analizie straty informacji to owa strata jest zjawiskiem złożonym, zaś zmienne $Y_1, Y_2, \ldots, Y_m$ – jej „stymulantami" (im większa wartość każdej z tych zmiennych tym większa strata informacji).

Na podstawie macierzy $\mathbf{Y} = [y_{ij}]$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, konstruujemy miernik kompleksowy według zasad opisanych w części 4 i formuły (4). Użyteczność tego miernika zilustrujemy na konkretnym przykładzie.

Rozpatrzmy mianowicie zbiór mikrodanych dotyczących 100 osób, pochodzący z pewnego badania rynku pracy. W zbiorze tym zawarte są następujące zmienne:

- **ID** – identyfikator (kolejny numer rekordu),
- **płeć** (ozn. PLEC, skala nominalna: M – mężczyzna, K – kobieta),
- **wykształcenie** (ozn. WYKSZTALCENIE, skala porządkowa: 1 – wyższe ze stopniem naukowym co najmniej doktora, 2 – wyższe z tytułem magistra, lekarza lub równorzędnym, 3 – wyższe z tytułem inżyniera, licencjata, dyplomowanego ekonomisty, 4 – dyplom ukończenia kolegium, 5 – policealne z maturą, pomaturalne, 6 – policealne bez matury, 7 – średnie zawodowe z maturą, 8 – średnie zawodowe bez matury, 9 – średnie ogólnokształcące z maturą, 10 – średnie ogólnokształcące bez matury, 11 – zasadnicze zawodowe, 12 – gimnazjalne, 13 – podstawowe, 14 – podstawowe nieukończone i bez wykształcenia),

– **status na rynku pracy** (ozn. STATUSRP, skala nominalna: 1 – pracujący, 2 – bezrobotny, 3 – bierny zawodowo),
– **odległość od miejsca zamieszkania do głównego miejsca pracy w km** (ozn. ODLEGLOSC, skala ilorazowa),
– **przychód miesięczny w zł** (ozn. PRZYCHOD, skala ilorazowa).

Na rysunku 1 przedstawiono fragment rozpatrywanej bazy danych.

Rysunek 1. Fragment bazy danych dotyczących osób na rynku pracy

ID	PLEC	WYKSZTALCE-NIE	STATUSRP	ODLEGLOSC	PRZYCHOD
1	M	6	3	2,0	2998,57
2	M	6	3	4,0	3905,51
3	K	5	2	3,0	1500,89
4	K	11	1	1,5	2682,93
5	K	9	1	5,0	2633,73
6	M	1	2	1,0	3243,32
7	K	3	2	0,5	3894,26
8	M	6	3	3,0	3600,04
9	K	6	2	10,0	1132,27
10	M	4	2	9,5	4207,18
11	K	3	1	7,0	3770,02
12	K	9	1	3,0	2569,40
13	M	4	2	4,0	7390,28
14	M	7	1	2,5	4430,25
15	K	5	2	3,5	1852,06
16	K	5	2	7,0	2032,44
17	M	1	2	10,0	4021,18
18	M	1	3	15,0	2663,83
19	K	9	3	11,5	3170,60
20	K	9	2	7,0	4196,99

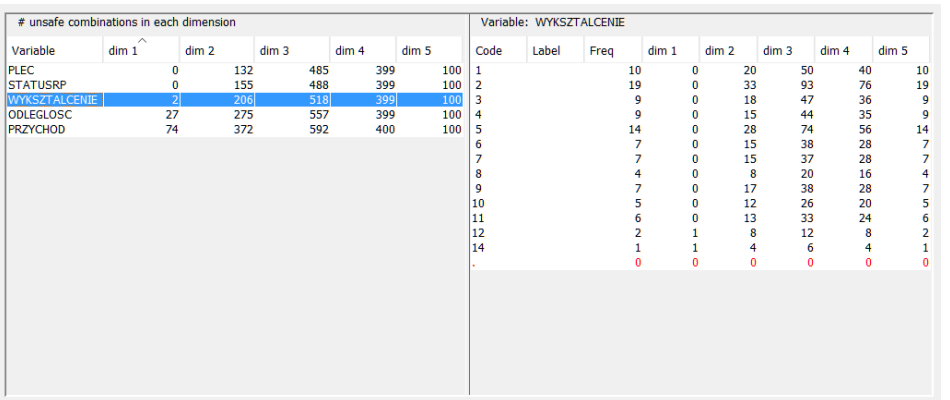
Ź r ó d ł o: opracowanie własne. Dane są fikcyjne.

Dla tak przygotowanego zbioru danych przeprowadzono kontrolę ujawniania danych. W przypadku zmiennych kategorialnych (tzn. wyrażonych na skali nominalnej lub porządkowej) zastosowano podejście postrandomizacyjne (PRAM) z prawdopodobieństwami zmiany kategorii postaci:
– płeć: M – 0,8, K – 0,8,
– wykształcenie: 1 – 0,8, 2 – 0,7, 3 – 0,6, 4 – 0,6, 5 – 0,6, 6 – 0,6, 7 – 0,7, 8 – 0,7, 9 – 0,8, 10 – 0,8, 11 – 0,8, 12 – 0,5, 14 – 0,5 (kategoria 13 nie wystąpiła), w tym przypadku celem ochrony przed nadmiernym zniekształceniem informacji ograniczono też możliwe zmiany do trzech najbliższych kategorii,
– status na rynku pracy: 1 – 0,7, 2 – 0,7, 3 – 0,8.

Dla zmiennych ciągłych (wyrażonych tutaj na skali ilorazowej) dokonano natomiast mikroagregacji z minimalną liczebnością grup wynoszącą 3. SDC przeprowadzono w programie μ-Argus, opracowanym właśnie do tego celu w Centralnym Biurze Statystycznym Holandii (Centraal Bureau voor de Statistiek, ang. Statistics Netherlands). Jest on – wraz ze stosowną dokumentacją i podręczni-

kiem – dostępny bezpłatnie pod adresem http://neon.vb.cbs.nl/casc/mu.htm. Na rysunku 2 uwidoczniono wyświetlone przez tenże program potencjalne możliwości ujawnienia danych wrażliwych.

Rysunek 2. Kombinacje danych o rynku pracy mogące prowadzić do ujawnienia informacji wrażliwych



unsafe combinations in each dimension

Variable	dim 1	dim 2	dim 3	dim 4	dim 5
PLEC	0	132	485	399	100
STATUSRP	0	155	488	399	100
WYKSZTALCENIE	2	206	518	399	100
ODLEGLOSC	27	275	557	399	100
PRZYCHOD	74	372	592	400	100

Variable: WYKSZTALCENIE

Code	Label	Freq	dim 1	dim 2	dim 3	dim 4	dim 5
1		10	0	20	50	40	10
2		19	0	33	93	76	19
3		9	0	18	47	36	9
4		9	0	15	44	35	9
5		14	0	28	74	56	14
6		7	0	15	38	28	7
7		7	0	15	37	28	7
8		4	0	8	20	16	4
9		7	0	17	38	28	7
10		5	0	12	26	20	5
11		6	0	13	33	24	6
12		2	1	8	12	8	2
14		1	1	4	6	4	1
.		0	0	0	0	0	0

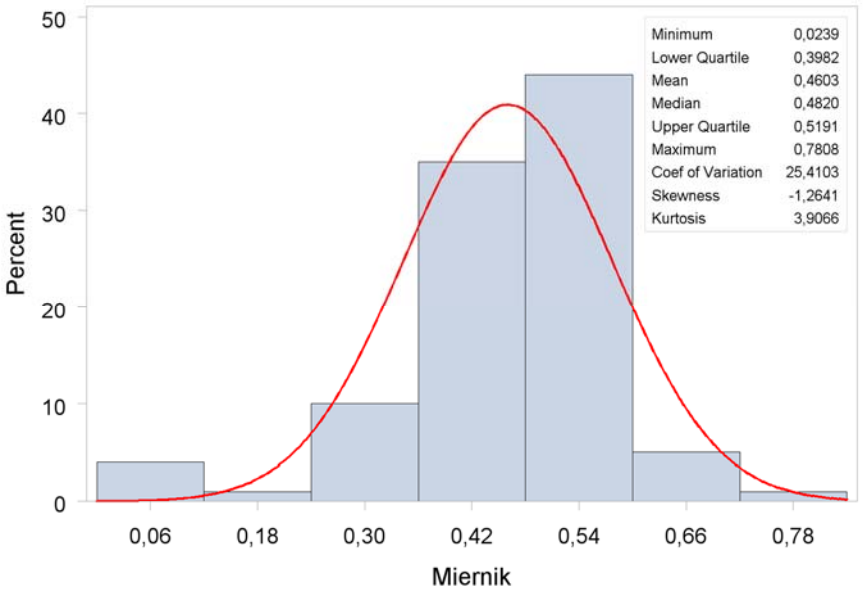
Ź r ó d ł o: opracowanie własne przy użyciu programu μ-Argus.

Po lewej stronie dla każdej zmiennej ukazano liczbę 1, 2, 3, 4 i 5-wymiarowych kombinacji wartości uznanych za niebezpieczne, obejmujących realizacje owej zmiennej (*dim1*, *dim2*, *dim3*, *dim4* i *dim5*). Kombinację uznaje się za niebezpieczną, jeśli występuje ona w zbiorze nie więcej niż k razy, gdzie k jest pewną arbitralnie ustaloną liczbą naturalną. W rozpatrywanym przypadku – zgodnie ze znaną i utrwaloną w naszym kraju praktyką – przyjęto $k = 2$. Po prawej stronie rys. 2 dla poszczególnych wartości zmiennej WYKSZTAŁCENIE uwidoczniono liczbę kombinacji wymiarów 1, 2, 3, 4 i 5 niebezpiecznych dla każdej kategorii (*Code*) oraz liczbę osób należących do poszczególnych kategorii pod tym względem (*Freq*)[5]. Widzimy zatem, że np. dla kategorii 12 i 14 już same te wartości są niebezpieczne, gdyż liczba rekordów, w których występują wynosi odpowiednio 2 i 1. W bazie jest też sporo niebezpiecznych kombinacji innych wymiarów.

Ocenę straty informacji opartą na mierniku kompleksowym przeprowadzono przy pomocy programu SAS (w tym jego środowiska IML). Wzorzec w tej sytuacji to (1, 1, 1, 1, 1) (maksymalne unormowane odchylenia od wartości prawdziwych), antywzorzec – (0, 0, 0, 0, 0,000828), a zatem najmniejsza możliwa strata, czyli praktyczna identyczność z danymi faktycznymi. Wizualizację rozkładu otrzymanych wartości miernika dla poszczególnych rekordów wraz z naniesioną dopasowaną krzywą normalną ukazano na rysunku 3. Podano na nim również wartości podstawowych statystyk opisowych dla tego rozkładu.

[5] Kategorie w pliku nie posiadały etykiet, stąd odnośna kolumna (*Label*) jest pusta.

Rysunek 3. Rozkład wyników obliczeń miernika TOPSIS straty informacji



Ź r ó d ł o: opracowanie własne przy użyciu programu SAS (w tym narzędzi środowiska IML).

Wyniki obliczeń można interpretować jako procentowe straty informacji w poszczególnych rekordach powstałe na skutek zastosowania SDC. Widzimy więc, że wahały się one od nieco ponad 2% do ponad 78%. Przeciętna strata informacji dla rekordu wynikła z SDC wynosi ok. 48%. Trzeba też zauważyć, że ponad połowa obserwacji wykazuje stratę mniejszą niż 50%, a w 75% rekordów strata nie przekracza 52%. Te wysokie – bądź co bądź – poziomy wynikają w znacznej mierze ze stosunkowo niedużej liczebności rozpatrywanego zbioru. Im większa liczba rekordów i im większa przewaga tejże liczby nad liczbą zmiennych, tym strata informacji spowodowana zastosowaniem SDC jest na ogół mniejsza. Rozkład wartości miernika jest dość wyraźnie lewostronnie asymetryczny i lepto-kurtyczny. Jego zmienność okazuje się względnie umiarkowana. Nic zatem dziwnego, że wartości testów normalności każą odrzucić hipotezę o normalnym kształcie tego rozkładu już na poziomie istotności 0,01 a nawet niższym (zob. tabela 1).

Tabela 1. WARTOŚCI TESTÓW NORMALNOŚCI DLA ROZKŁADU WARTOŚCI MIERNIKA TOPSIS

Test	Statystyka		Wartość p	
Shapiro–Wilka	W	0,8868	Pr < W	<0,0001
Kołmogorowa–Smirnowa	D	0,1535	Pr > D	<0,0100
Cramera–von Misesa	W–Sq	0,4888	Pr > W–Sq	<0,0050
Andersona–Darlinga	A–Sq	3,0569	Pr > A–Sq	<0,0050

Ź r ó d ł o: opracowanie własne przy użyciu programu SAS (w tym jego środowiska IML).

6. WNIOSKI

Z powyższych rozważań wynika jasno, że miernik kompleksowy uzyskany metodą zwaną TOPSIS stanowi wartościowe narzędzie oceny straty informacji. Główne zalety techniki opartej o konstrukcję wskaźnika syntetycznego (miernika kompleksowego) w ocenie straty informacji spowodowanej zastosowaniem kontroli ujawniania danych przed ich upublicznieniem są dwie. Pierwsza z nich polega na tym, że miernik kompleksowy dostarcza jednolitej oceny dla całego zbioru danych, bez względu na skalę pomiarową, na jakiej prowadzone są ich obserwacje. Zazwyczaj z uwagi na tę specyfikę strata musiała być szacowana odrębnie dla danych kategorialnych i odrębnie dla danych wyrażonych na skali różnicowej bądź ilorazowej (por. Domingo-Ferrer i inni, 2001).

Drugi walor przedstawionego podejścia to unikanie bezpośredniego wiązania odmiennych informacji dostarczanych przez zmienne (jakie zachodzi np. w przypadku sumowania bezwzględnych różnic lub ich kwadratów, jeśli zmienne mają realizację na skali różnicowej czy ilorazowej), a tym samym wrażliwości na naturalne różnice w skali i zakresach wartości poszczególnych zmiennych.

Oprócz tego użycie techniki opartej o konstrukcję miernika kompleksowego w ocenie straty informacji spowodowanej zastosowaniem metod SDC okazało się bardzo użyteczne w praktyce. Wyniki są łatwo interpretowalne, co daje możliwość oceny stopnia straty informacji na poszczególnych rekordach, kształtu i parametrów jej rozkładu oraz oczekiwanej sumarycznej wielkości. Dzięki pewnej swobodzie w doborze wzorca i antywzorca ukazana metoda umożliwia konfrontację realiów z oczekiwaniami.

Rozpatrywane podejście w ukazanej formie da się też użyć do szacowania straty informacji opartego na wpływie na wariancję szacunków. Mianowicie w przypadku zmiennych, których wartości wyrażają się na skali różnicowej lub ilorazowej wystarczy bowiem jako cechy diagnostyczne przyjąć np. unormowane na [0,1] odległości między średnimi arytmetycznymi, rangami bądź kowariancjami odpowiednich zmiennych przed i po zastosowaniu SDC. Wykorzystując komponenty odległości od wzorca i antywzorca można też ustalać rozmiar straty informacji na poszczególnych zmiennych.

Na zakończenie warto przypomnieć i podkreślić, że oczekiwana strata informacji powstałej na skutek zastosowania kontroli ujawniania danych stanowi jedną z kluczowych cech jakościowych udostępnianych użytkownikom mikrodanych statystycznych, w związku z czym winna być każdorazowo podawana im do wiadomości. Dzięki temu użytkownik ma możność nie tylko poznania skali ingerencji motywowanej koniecznością ochrony poufności, ale i uwzględniania tejże straty w ocenie jakości uzyskanych przez siebie na podstawie owych mikrodanych oszacowań.

LITERATURA

Balcerzak A. P., Pietrzak M. B., (2015), Wpływ efektywności instytucji na jakość życia w Unii Europejskiej. Badanie panelowe dla lat 2004–2010, *Przegląd Statystyczny*, 62 (1), 71–91.

Benschop T., Machingauta C., Welch M., (2018), *Statistical Disclosure Control for Microdata: A Practice Guide*, World Bank, Development Data Group (WB-DECDG), dostępny pod adresem: https://media.readthedocs.org/pdf/sdcpractice/latest/sdcpractice.pdf.

Chen C.-T., (2000), Extensions of the TOPSIS for Group Decision-Making under Fuzzy Environment, *Fuzzy Sets and Systems*, 114 (1), 1–9.

De Wolf P.-P., Gouweleeuw J. M., Kooiman P., Willenborg L. C. R. J., (1999), Reflections on PRAM. w: Domingo-Ferrer J., (red.), *Statistical Data Protection*, Office for Official Publications of the European Communities, Luxembourg, 337–349.

Domingo-Ferrer J., Mateo-Sanz J. M., Torra V., (2001), Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk, w: *Pre-proceedings of ETK-NTTS (Exchange of Technology and Know-how – New Techniques and Technologies for Statistics)*, 2, 807–826, http://neon.vb.cbs.nl/casc/NTTSJosep.pdf.

Grabiński T., (2017), Uproszczona metoda delimitacji wektorowej, *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie*, 5 (965), 69–86.

Hellwig Z., (1967), Procedure of Evaluating High Level Manpower Data and Typology of Countries by Means of the Taxonomic Method, w: *Study III of the UNESCO Statistical Office; Towards a System of Quantitative Indicators of Components of Human Resources Indicators Development*, UNESCO, Paris.

Hellwig Z., (1968), Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr, *Przegląd Statystyczny*, 15 (4), 307–327.

Hellwig Z., (1969), On the Problem of Weighting in International Comparisons, w: *Study VII of the UNESCO Statistical Office; Towards a System of Quantitative Indicators of Components of Human Resources Indicators Development*, UNESCO, Paris.

Hellwig Z., (1972a), Approximative Methods of Selection of an Optimal Set of Predictors, w: *Study XVI of the UNESCO Statistical Office; Towards a System of Quantitative Indicators of Components of Human Resources Indicators Development*, UNESCO, Paris.

Hellwig Z., (1972b), On Optimal Choice of Predictors, w: Gostkowski Z., (red.), *Towards a System of Human Resources Indicators for Less Developed Countries*, UNESCO – Ossolineum, Paris – Wrocław, 69–90.

Hellwig Z., (1981), Wielowymiarowa analiza porównawcza i jej zastosowanie w badaniach wielocechowych obiektów gospodarczych, w: Welfe W., (red.), *Metody i modele ekonomiczno-matematyczne w doskonaleniu zarządzania gospodarką socjalistyczną*, PWE, Warszawa, 46–68.

Höninger J., Pattloch D., Voshage R., (2010), *On-Site Access to Micro Data: Preserving the Treasure, Preventing Disclosure*, State Statistical Institute Berlin-Brandenburg, Research Data Centre.

Hundepool A., Domingo-Ferrer J., Franconi L., Giessing S., Lenz R., Longhurst J., Schulte Nordholt E., Seri G., de Wolf P.-P., (2006), *Handbook on Statistical Disclosure Control*, Version 1.0 CENEX SDC – a CENtre of EXcellence for Statistical Disclosure Control, Eurostat, Luxembourg, https://ec.europa.eu/eurostat/cros/system/files/CENEX-SDC_handbook.pdf.

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

Hwang C. L., Yoon K., (1981), *Multiple Attribute Decision Making: Methods and Applications*, Springer-Verlag, New York.

Kukuła K., Luty L., (2015), Propozycja procedury wspomagającej wybór metody porządkowania liniowego, *Przegląd Statystyczny*, 62 (2), 219–231.

Lira J., Wagner W., Wysocki F., (2002), Mediana w zagadnieniach porządkowania obiektów wielocechowych, w: Paradysz J., (red.), *Statystyka regionalna w służbie samorządu lokalnego i biznesu*, Internetowa Oficyna Wydawnicza Centrum Statystyki Regionalnej, Akademia Ekonomiczna w Poznaniu, Poznań, 87–99.

Malina A., (2002) Wielokryterialna taksonomia w analizie porównawczej struktur gospodarczych Polski, w: Zeliaś A., (red.), *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków, 305-312.

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

Młodak A., (2006a), *Analiza taksonomiczna w statystyce regionalnej*, Centrum Doradztwa i Informacji DIFIN, Warszawa.

Młodak A., (2006b), Multilateral Normalisations of Diagnostic Features, *Statistics in Transition*, 7 (5), 1125–1139.

Młodak A., (2014), On the Construction of an Aggregated Measure of the Development of Interval Data, *Computational Statistics*, 29, 895–929.

Pawełek B., (2008), *Metody normalizacji zmiennych w badaniach porównawczych złożonych zjawisk ekonomicznych*, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie. Seria specjalna. Monografie, nr 187, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.

Shih H. S., Shyur H. J., Lee E. S., (2007), An Extension of TOPSIS for Group Decision Making, *Mathematical and Computer Modelling*, 45 (7–8), 801–813.

Shlomo N., Skinner C., (2010), Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata, *The Annals of Applied Statistics*, 4 (3), 1291–1310.

Skinner C., Marsh C., Openshaw S., Wymer C., (1994), Disclosure Control for Census Microdata, *Journal of Official Statistics*, 10, 31–51.

Śmiłowska T., (1997) *Statystyczna analiza poziomu życia ludności Polski w ujęciu przestrzennym*, Studia i Prace. Z Prac Zakładu Badań Statystyczno-Ekonomicznych Głównego Urzędu Statystycznego i Polskiej Akademii Nauk, Zeszyt 247, Warszawa.

Templ M., (2017), *Statistical Disclosure Control for Microdata Using Methods and Applications in R*, Springer International Publishing AG, Cham, Szwajcaria.

Walesiak M., (2006), *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław.

Walesiak M., (2014a), Przegląd formuł normalizacji wartości zmiennych oraz ich własności w statystycznej analizie wielowymiarowej, *Przegląd Statystyczny*, 61 (4), 364–372.

Walesiak M., (2014b), Wzmacnianie skali pomiaru dla danych porządkowych w statystycznej analizie wielowymiarowej, w: Jajuga K., Walesiak M., (red.), *Taksonomia 22. Klasyfikacja i analiza danych – teoria i zastosowania*. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, 327, 60–68.

Walesiak M., (2016), Wybór grup metod normalizacji wartości zmiennych w skalowaniu wielowymiarowym, *Przegląd Statystyczny*, 63 (1), 7–18.

Walesiak M., (2018), The Choice of Normalization Method and Rankings of the Set of Objects Based on Composite Indicator Values, *Statistics in Transition – New Series*, 19 (4), 693–710.

Willenborg L., de Waal T., (1996), *Statistical Disclosure Control in Practice*, Lecture Notes in Statistics, Springer Verlag, New York, Inc.

Zeliaś A., (2002), Some Notes on the Selection of Normalization of Diagnostic Variables, *Statistics in Transition*, 5 (5), 787–802.

WYKORZYSTANIE MIERNIKA KOMPLEKSOWEGO W OCENIE STRATY INFORMACJI NA SKUTEK KONTROLI UJAWNIANIA MIKRODANYCH

Streszczenie

Praca zawiera propozycję oryginalnej metody oceny straty informacji powstałej na skutek zastosowania kontroli ujawniania danych (ang. Statistical Disclosure Control, SDC) dokonywanej podczas przygotowywania danych wynikowych do publikacji i do udostępniania ich zainteresowanym użytkownikom. Narzędzia SDC umożliwiają ochronę danych wrażliwych przed ujawnieniem – tak bezpośrednim, jak i pośrednim. Artykuł koncentruje się na przypadku spseudonimizowanych mikrodanych, czyli wykorzystywanych do badań naukowych danych jednostkowych pozbawionych zasadniczych cech identyfikacyjnych. SDC polega tu zazwyczaj na ukrywaniu, zamienianiu czy zakłócaniu oryginalnych danych. Tego rodzaju ingerencja wiąże się jednak ze stratą pewnych informacji. Stosowane tradycyjnie metody pomiaru owej straty są nierzadko wrażliwe na odmienności wynikające ze skali i zakresu wartości zmiennych oraz nie mogą być zastosowane do danych wyrażonych na skali porządkowej. Wiele z nich słabo uwzględnia też powiązania między zmiennymi, co bywa istotne w różnego rodzaju analizach. Stąd celem artykułu jest przedstawienie propozycji użycia – mającej swe źródło w pracach Zdzisława Hellwiga – metody konstrukcji unormowanego i łatwo interpretowalnego miernika kompleksowego (zwanego także wskaźnikiem syntetycznym) powiązanych cech opartego na wzorcu i antywzorcu rozwojowym w ocenie straty informacji spowodowanej zastosowaniem wybranych technik SDC oraz zbadanie jej praktycznej użyteczności. Miernik został tutaj skonstruowany na podstawie odległości między danymi wyjściowymi a danymi po zastosowaniu SDC z uwzględnieniem skal pomiarowych.

Słowa kluczowe: kontrola ujawniania danych, mikrodane, strata informacji, miernik kompleksowy, miara odległości

USING THE COMPLEX MEASURE IN AN ASSESSMENT OF THE INFORMATION LOSS DUE TO THE MICRODATA DISCLOSURE CONTROL

Abstract

The paper contains a proposal of original method of assessment of information loss resulted from an application of the Statistical Disclosure Control (SDC) conducted during preparation of the resulting data to the publication and disclosure to interested users. The SDC tools enable protection of sensitive data from their disclosure – both direct and indirect. The article focuses on pseudon-

imised microdata, i.e. individual data without fundamental identifiers, used for scientific purposes. This control is usually to suppress, swapping or disturbing of original data. However, such intervention is connected with the loss of some information. Optimization of choice of relevant SDC method requires then a minimization of such loss (and risk of disclosure of protected data). Traditionally used methods of measurement of such loss are not rarely sensitive to dissimilarities resulting from scale and scope of values of variables and cannot be used for ordinal data. Many of them weakly take also connections between variables into account, what can be important in various analyses. Hence, this paper is aimed at presentation of a proposal (having the source in papers by Zdzisław Hellwig) concerning use of a method of normalized and easy interpretable complex measure (called also the synthetic indicator) for connected features based on benchmark and anti–benchmark of development to the assessment of information loss resulted from an application of some SDC techniques and at studying its practical utility. The measure is here constructed on the basis of distances between original data and data after application of the SDC taking measurement scales into account.

Keywords: Statistical Disclosure Control, microdata, information loss, complex measure, distance measure

JEL Codes: C80, C19, C63