Urząd Statystyczny w Poznaniu

Wykorzystanie metody TOPSIS w ocenie efektywności kontroli ujawniania mikrodanych

dr hab. Andrzej Młodak
Urząd Statystyczny w Poznaniu,
Ośrodek Statystyki Małych Obszarów

15.05.2018 Kalisz

Wprowadzenie

- W ostatnich latach rośnie popyt na mikrodane, które są szczególnie przydatne dla celów naukowo – badawczych (dają większe możliwości niż klasyczne publikacje)
- Stwarza to jednak ryzyko identyfikacji jednostki lub odtworzenia danych wrażliwych
- Rozwija się zatem kontrola ujawniania danych (ang. *Statistical Disclosure Control*, SDC), czyli metodologia ukrywania lub zniekształcania danych wrażliwych
- Jednym z celów SDC jest minimalizacja straty informacji na skutek tej ochrony, należy zatem tę stratę ocenić
- Istnieje wiele sposobów pomiaru straty; obecnie zaprezentujemy podejście oparte na taksonomicznej metodzie TOPSIS (ang. *The Technique for Order of Preference by Similarity to Ideal Solution*).

Struktura prezentacji

- Cechy SDC dla mikrodanych
- Metody SDC dla mikrodanych
- Strata informacji i jej wyznaczanie
- Metoda TOPSIS
- Zastosowanie metody TOPSIS w SDC
- Wnioski
- Bibliografia

Cechy SDC dla mikrodanych

- Mikrodane zawierają cztery kategorie zmiennych:
  - identyfikatory – zmienne, które w sposób jednoznaczny identyfikują respondenta (PESEL, NIP, REGON, itp.),
  - quasi–identyfikatory albo zmienne kluczowe – zmienne, niekoniecznie identyfikujące, 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.

Metody SDC dla mikrodanych

- **Anonimizacja** – 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
  - przykład: usunięcie imienia, nazwiska osoby oraz jej numeru PESEL ze zbioru danych zgromadzonych w badaniu osób
  - nie gwarantuje to zabezpieczenia przed możliwością identyfikacji jednostki w oparciu o unikalne kombinacje wartości innych zmiennych i odtworzenia wrażliwych o niej informacji.

Metody SDC dla mikrodanych

- **Metody niezakłóceniowe** (ang. *non–perturbative masking*) prowadzą one do tego, że wrażliwe dane stają się – w różny sposób – niewidoczne dla zewnętrznego użytkownika: w finalnym udostępnianym zbiorze określona informacja jednostkowa albo figuruje w dokładnej postaci, albo jej nie ma wcale.
- Przykłady metod niezakłóceniowych
  - podpróbkowanie (ang. *subsampling*) – udostępnianie pewnej określonej próbki rekordów spośród figurujących w bazie danych zgromadzonych w trakcie badania statystycznego; duża szansa pominięcia unikatowych rekordów

Metody SDC dla mikrodanych

- Przykłady metod niezakłóceniowych:
  - przekodowanie (ang. *recoding*) wrażliwych zmiennych – połączenie kilku kategorii w jedną – bardziej zgrubną i o większej liczbie należących do niej jednostek (dla zmiennej kategorialnej) lub zastąpienie zmiennej ciągłej przez jej odpowiednik w postaci dyskretnej
  - lokalne ukrywanie danych (ang. *local suppression*) – polega na usuwaniu pewnych wartości niektórych zmiennych kategorialnych dla konkretnych jednostek celem uniknięcia ich identyfikacji; liczba ukryć powinna być możliwie jak najmniejsza

Metody SDC dla mikrodanych

- **Metody zakłóceniowe** – zakłócanie wrażliwych wartości zmiennych celem uniemożliwienia dokładnego ich odtworzenia przez nieuprawnionego użytkownika przy jednoczesnej minimalizacji strat informacyjnych.
- Przykłady metod zakłóceniowych:
  - dodawanie szumu (ang. *noise addition*) – nakładnie na oryginalne dane wrażliwe (addytywnie, multyplikatywnie) 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. z rozkładu normalnego lub jednostajnego

Metody SDC dla mikrodanych

- Przykłady metod zakłóceniowych:
  - mikroagregacja (ang. *microaggregation*) – obejmuje ona w istocie pewną rodzinę narzędzi zapewniających ochronę poufności danych w ujęciu makro poprzez zastąpienia wartości indywidualnych odpowiednimi wartościami sumarycznymi dla niewielkich poziomów agregacji 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 $p\%$, $0 < p < 100$); stosowane

Metody SDC dla mikrodanych

- Przykłady metod zakłóceniowych
  - zaokrąglanie (ang. *rounding*) – oryginalne wartości zastępuje się ich wersjami zaokrąglonymi
    - zazwyczaj wybiera się je ze zbioru punktów zaokrągleń definiującego zestaw zaokrągleń, np. przyjęcie jako punktów zaokrągleń $p_i = b \cdot i$, dla $i = 1,2, \ldots, l$, gdzie liczba naturalna b stanowi podstawę zaokrąglania, zaś l to maksymalny zakres zaokrągleń
    - można to czynić losowo, np. 136,5 zaokrąglamy do 135 z prawdopodobieństwem 1–0,65=0,35 lub do 140 z prawdopodobieństwem 0,65.

Metody SDC dla mikrodanych

- Przykłady metod zakłóceniowych
  - metoda postrandomizacyjna (ang. *The Post–RAndomization Method, PRAM*) – metoda probabilistyczna, 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 szumu, ukrywanie danych oraz przekodowywanie; stosowana do zmiennych kategorialnych

Strata informacji i jej wyznaczanie

- Na skutek stosowania SDC następuje ubytek zasobu informacyjnego
- Jego pomiar 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

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

gdzie $d(\cdot,\cdot) \in [0,1]$ jest miarą odległości spełniającą klasyczne warunki zwrotności, symetrii i nierówności trójkąta. Im większa wartość λ , tym dokuczliwsza strata informacji.

Strata informacji i jej wyznaczanie

- Ogólna postać typowej miary straty
  - jeżeli wartości zmiennej X_j mierzone są na skali nominalnej, to

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

  - jeżeli wartości X_j mierzone są na skali porządkowej, wówczas

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

    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 odległość ta może być np. postaci

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

    $i = 1,2,\dots,n,\ j = 1,2,\dots,m$
  - miary straty bywają też oparte na średnich odległościach lub kowariancji zmiennych wyjściowych i zakłócanych

Metoda TOPSIS

- Jest to konstrukcja miernika kompleksowego zróżnicowania zjawiska złożonego oparta na wzorcu i antywzorcu rozwojowym (ang. *Technique for Order Performance by Similarity to Ideal Solution*)
- Każdy obiekt jest opisany przez wartości m cech diagnostycznych: $\boldsymbol{\gamma}_i = (x_{i1}, x_{i2}, \ldots, x_{im})$, $i = 1,2, \ldots, n$.
- Założenia konstrukcji miernika:
  - logiczność wzajemnych powiązań cech diagnostycznych,
  - mierzalność – możliwość liczbowego wyrażenia poziomu cechy diagnostycznej,
  - dostępność i kompletność informacji statystycznych
  - cechy diagnostyczne są stymulantami (jeśli nie były nimi pierwotnie, to zostały na nie przekształcone.

Metoda TOPSIS

- Etapy konstrukcji miernika:
  - normalizacja cech diagnostycznych, np. standaryzacja: $z_{ij} = (x_{ij} - \bar{x}_j)/s_j$, $i = 1,2,\dots,n$, $j = 1,2,\dots,m$
  - zdefiniowanie wzorca i antywzorca rozwojowego
    - wzorzec: $\varphi_j^{(+)} = \max_{i=1,2,\dots,n} z_{ij}$
    - antywzorzec: $\varphi_j^{(-)} = \min_{i=1,2,\dots,n} z_{ij}$, $j = 1,2,\dots,m$
  - wyznaczenie odległości obiektów od wzorca i antywzorca, np. euklidesowej:

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

  - obliczenie wartości miary kompleksowej: $\eta_i = \frac{\delta_i^{(-)}}{\delta_i^{(-)} + \delta_i^{(+)}} \in [0,1]$, $i = 1,2,\dots,n$. Im wyższa, tym sytuacja obiektu lepsza.

Zastosowanie metody TOPSIS w SDC

- Załóżmy, że mamy dwa zbiory danych badania określonego zjawiska: wyjściowy $\mathbf{X} = [x_{ij}]$ i po poddaniu go SDC $\mathbf{X}^* = [x_{ij}^*]$, $i = 1,2,\dots,n$, $j = 1,2,\dots,m$
- Wyznaczamy odległości $d(x_{ij}, x_{ij}^*)$ między odpowiednimi obserwacjami według formuły zależnej od skali pomiarowej zmiennej X_j, $i = 1,2,\dots,n$, $j = 1,2,\dots,m$
- Zmienne X_j, takie, że $y_{ij} = d(x_{ij}, x_{ij}^*)$, $i = 1,2,\dots,n$, $j = 1,2,\dots,m$, są „stymulantami” straty (im większa wartość tym większa strata)
- Ponieważ odległość $d(\cdot,\cdot)$ jest znormalizowana, zatem normalizacja zmiennych $Y_1, Y_2, \dots, Y_m$ nie jest konieczna
- Na podstawie macierzy Y konstruujemy przy użyciu metody TOPSIS miernik kompleksowy.

Zastosowanie metody TOPSIS w SDC

- Przykład
  - rozpatrzmy zbiór mikrodanych dotyczących 100 osób, pochodzący z pewnego badania rynku pracy
  - zmienne zawarte w bazie
    - **ID** - identyfikator (kolejny numer rekordu)
    - **płeć** (skala nominalna: M – mężczyzna, K – kobieta)
    - **wykształcenie** (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)

Zastosowanie metody TOPSIS w SDC

- Przykład
  - zmienne zawarte w bazie
    - **status na rynku pracy** (skala nominalna: 1 – pracujący, 2 – bezrobotny, 3 –bierny zawodowo)
    - **odległość od miejsca zamieszkania do głównego miejsca pracy w km** (skala ilorazowa)
    - **przychód miesięczny w zł** (skala ilorazowa)
  - początek bazy

ID	PLEC	WYKSZTALCENIE	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

Zastosowanie metody TOPSIS w SDC

- Przykład
  - zastosowane metody:
    - dla zmiennych kategorialnych (wyrażonych na skali nominalnej lub porządkowej) – PRAM z prawdopodobieństwami zmiany kategorii:
      - 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), ograniczono też możliwe zmiany do trzech najbliższych kategorii,
      - status na rynku pracy: 1 – 0,7, 2 – 0,7, 3 – 0,8

Zastosowanie metody TOPSIS w SDC

- Przykład
  - dla zmiennych ciągłych (wyrażonych tutaj na skali ilorazowej) mikroagregacja z minimalną liczebnością grup wynoszącą 3
  - SDC przeprowadzono w programie μ–Argus
  - liczba wrażliwych kombinacji dla wykształcenia

unsafe combinations in each dimension

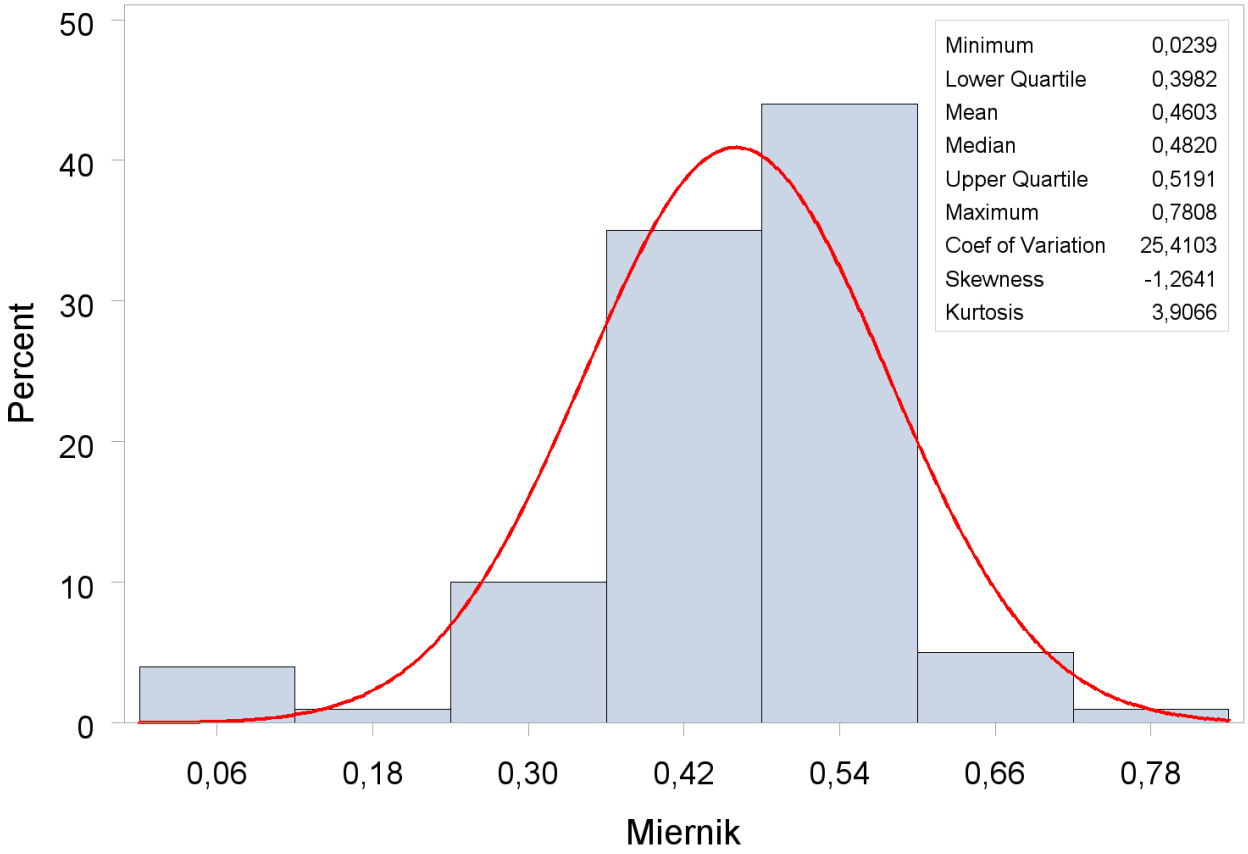
Variable	dim 1	dim 2	dim 3
PLEC	0	132	292
STATUSRP	0	155	295
WYKSZTALCENIE	2	206	487
ODLEGLOSC	27	275	487
PRZYCHOD	74	372	500

Variable: WYKSZTALCENIE

Code	Label	Freq	dim 1	dim 2	dim 3
1		10	0	20	48
2		19	0	33	90
3		9	0	18	44
4		9	0	15	43
5		14	0	28	70
6		7	0	15	35
7		7	0	15	35
8		4	0	8	19
9		7	0	17	35
10		5	0	12	25
11		6	0	13	28
12		2	1	8	10
14		1	1	4	5
.		0	0	0	0

Zastosowanie metody TOPSIS w SDC

- Przykład
  - wyniki obliczeń miernika (w programie SAS IML)



Zastosowanie metody TOPSIS w SDC

- Przykład
  - Testy normalności

Test	Statystyka		p–value	
Shapiro-Wilka	W	0.886775	Pr < W	<0.0001
Kołmogorowa-Smirnowa	D	0.153543	Pr > D	<0.0100
Cramera-von Misesa	W-Sq	0.488842	Pr > W-Sq	<0.0050
Andersona-Darlinga	A-Sq	3.056867	Pr > A-Sq	<0.0050

  - wyniki obliczeń można interpretować jako procentowe straty informacji w poszczególnych rekordach na skutek zastosowania SDC – strat ogółem wynosi ok. 48%
  - straty te są zróżnicowane, a ich rozkład jest lewostronnie asymetryczny

Wnioski

- Miernik kompleksowy uzyskany metodą TOPSIS stanowi wartościowe narzędzie oceny straty informacji. Jego zalety to:
  - unikanie bezpośredniego wiązania odmiennych informacji dostarczanych przez zmienne (jakie zachodzi np. w przypadku sumowania bezwzględnych różnic lub ich kwadratów)
  - łatwa interpretacja, dająca możliwość oceny stopnia straty informacji na poszczególnych rekordach, jej rozkładu oraz oczekiwanej sumarycznej wielkości
  - możliwość konfrontacji realiów z oczekiwaniami (poprzez dobór wzorca)
- Oczekiwana strata informacji winna być też podawana do wiadomości użytkownika danych

Bibliografia

- 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.
- Hsu-Shih Shih, Huan-Jyh Shyur, E. Stanley Lee, *An extension of TOPSIS for group decision making*, Mathematical and Computer Modelling, 45 (2007), str. 801–813.

Urząd Statystyczny w Poznaniu

Dziękuję za uwagę