Urząd Statystyczny
w Poznaniu

Metoda kompleksowej oceny ryzyka ujawnienia dla mikrodanych

dr hab. Andrzej Młodak

Urząd Statystyczny w Poznaniu,
Ośrodek Statystyki Małych Obszarów

Uniwersytet Kaliski im. Prezydenta Stanisława Wojciechowskiego,
Międzywydziałowy Zakład Matematyki i Statystyki

Wprowadzenie

- Rosnące zapotrzebowanie na informacje statystyczne powoduje dążenie do zwiększania zakresu i szczegółowości danych udostępnianych przez statystykę publiczną i innych gestorów.
- Konieczne jest więc zapewnienie ochrony poufności danych wrażliwych, ale też maksymalizacji użyteczności oferowanych informacji. Realizacja tego celu wymaga stosowania metod kontroli ujawniania danych (ang. *Statistical Disclosure Control, SDC*).
- Jedną z kluczowych kwestii w tym zakresie stanowi zatem ocena ryzyka identyfikacji jednostki (tzw. ryzyka ujawnienia) na różnych etapach procesu SDC, zwłaszcza dla danych wejściowych i dla informacji finalnie przewidzianych do udostępnienia. Szczególnego znaczenia problem ten nabiera w przypadku mikrodanych, a to przede wwszystkim z uwagi na indywidualny poziom informacji oraz ich różnorodny charakter.
- W prezentacji zaproponowana zostanie metoda kompleksowej oceny ryzyka ujawnienia oraz ustalania wkładu poszczególnych zmiennych do tegoż ryzyka w ujęciu globalnym uwzględniająca jego komponenty pochodzące zarówno ze zmiennych kategorialnych, jak i ciągłych.

Struktura prezentacji

1. Istniejące metody oceny ryzyka ujawnienia
2. Główne założenia proponowanego rozwiązania
3. Koncepcja kompleksowej miary ryzyka ujawnienia
4. Wyznaczanie wkładu w ryzyko
5. Zastosowania empiryczne
6. Wnioski
7. Literatura

Istniejące metody oceny ryzyka ujawnienia

Ryzyko ujawnienia i miary dla zmiennych kategorialnych

- Ryzyko ujawnienia oznacza zagrożenie identyfikacją konkretnej jednostki (osoby, firmy) prowadzącą do nieuprawnionego uzyskania dotyczących jej informacji wrażliwych. Ma to szczególne znaczenie w przypadku możliwości identyfikacji pośredniej (a zatem w oparciu o quasi-identyfikatory – zmienne, których powiązanie ze sobą może doprowadzić do identyfikacji jednostki).
- W literaturze można znaleźć liczne metody oceny ryzyka ujawnienia dla mikrodanych. Jednak mają one odrębny charakter dla quasi-identyfikatorów indywidualnych i ciągłych.
- W przypadku zmiennych kategorialnych znane rozwiązania bazują na regułach k-anonimowości lub l-różnorodności. W praktyce oparte są one na częstościach występowania poszczególnych kombinacji wartości takich zmiennych. Na przykład Hundepool i in. (2012) przedstawiają miarę ryzyka indywidualnego (tj. ryzyka identyfikacji podmiotu, któremu odpowiada dany rekord) będącą prawdopodobieństwem prawidłowego powiązania rekordu z ową jednostką w najgorszym możliwym scenariuszu. Natomiast ryzyko globalne liczone jest tutaj jako suma odwróconych częstości kombinacji.

Istniejące metody oceny ryzyka ujawnienia

Miary dla zmiennych kategorialnych i ciągłych

- Inne podejścia w zakresie ryzyka ujawnienia dla zmiennych kategorialnych to m.in.:
  - podejście Benedettiego–Franconi wykorzystujące rozkład Poissona częstości kombinacji
  - metoda SUDA (ang. *Special Uniques Detection Algorithm*) – zob. by Elliot, Manning, and Ford (2002), oparta na minimalnych kombinacjach unikatowych (ang. *Minimal Sample Uniques*, MSU), czyli takich, które są niebezpieczne (ich rzadkość może prowadzić do identyfikacji jednostki), ale żaden ich podzbiór nie jest niebezpieczny.
- Dla ciągłych quasi-identyfikatorów ważny jest miernik określający odsetek obserwacji mieszczących się w przedziale, którego środkiem jest zaburzona wartość, podczas gdy górna granica takiego przedziału odpowiada najgorszemu scenariuszowi, w którym intruz ma pewność, że każdy najbliższy sąsiad tej jednostki jest rzeczywiście nią samą (Templ (2017)). Jednak to podejście może być stosowane tylko w ujęciu porównawczym, tzn. w porównaniu danych przed i po SDC.

Główne założenia proponowanego rozwiązania

Główne założenia konstrukcji miar ryzyka

- Bierzemy pod uwagę tylko zmienne mogące posłużyć do identyfikacji jednostki i ujawnienia jej danych wrażliwych.
- W przypadku zmiennych kategorialnych pierwotna miara ryzyka ujawnień wykorzystuje częstość występowania danej kombinacji kategorii i typowe podejście częstotliwości odwróconej, natomiast dla zmiennych ciągłych – liczbę rekordów należących do otoczenia danej wartości określonego przez przyjęty poziom precyzji. Następnie wyznacza się odsetek zmiennych, dla których częstotliwości te są mniejsze od przyjętego progu.
- Ocena udziału poszczególnych zmiennych w ryzyku całkowitym opiera się także na tym mechanizmie, jednak głównymi narzędziami stosowanymi tutaj są dwa rozwiązania gier kooperacyjnych: wartość Shapleya i wartość solidarnościowa. Zaproponowali je odpowiednio: Shapley (1953) oraz Nowak i Radzik (1994).

Koncepcja kompleksowej miary ryzyka ujawnienia

Podstawy i reguły częstościowe

- Niech $\mathbb{N}$ i $\mathbb{R}$ oznaczają odpowiednio zbiory liczb naturalnych i rzeczywistych. Załóżmy, że analizujemy bazę danych zawierającą n jednostek i m zmiennych $X_1, X_2, \ldots, X_m$ ($n, m \in \mathbb{N}$) będących quasi-identyfikatorami, przy czym – bez straty ogólności – zmienne $X_1, X_2, \ldots, X_h$ są kategorialne a zmienne $X_{h+1}, X_{h+2}, \ldots, X_m$ ($h \in \mathbb{N}, 0 \leq h \leq m$) – ciągłe oraz że nie występują braki odpowiedzi.
- Dla zmiennych kategorialnych ocena ryzyka bazuje zazwyczaj na regule *k-anonimowości*: kombinacja wartości jest bezpieczna gdy występuje w co najmniej k ($k \in \mathbb{N}$) rekordach (najczęściej $k = 3$, $k = 2$ lub $k = 5$).
- Stosuje się też regułę *l-różnorodności* (kombinacja jest bezpieczna gdy dla grupy rekordów, w których występuje jest przynajmniej l, $l \in \mathbb{N}$, różnych wrażliwych wartości o największej częstości występowania) lub regułę *t–bliskości* (klasa rekordów jest bezpieczna gdy rozkład każdej zmiennej w tej klasie jest bliski jej rozkładowi dla populacji (tzn. gdy odległość pomiędzy wrażliwymi atrybutami w tych ujęciach jest mniejsza niż t).

Koncepcja kompleksowej miary ryzyka ujawnienia

Komponenty miary ryzyka

- Dla quasi-identyfikatorów kategorialnych jest to odwrotność występowania danej kombinacji w bazie danych.
- Dla zmiennej ciągłej jej wartość x jest bezpieczna jeżeli liczba innych obserwacji należących do przedziału, którego środkiem jest x i o długości wyznaczonej przez arbitralnie ustalony próg precyzji $p \in (0, 1)$ – tzn. do przedziału $[x - px, x + px]$ – przekracza określoną wielkość, np. 3,
- Niech p_j będzie progiem precyzji dla X_j. Definiujemy

$$w_j(i) \stackrel{df}{=} \begin{cases} 1 & \text{gdy } \eta(x_{ij}, p_j) < k, \\ 0 & \text{gdy } \eta(x_{ij}, p_j) \geq k, \end{cases}$$

gdzie $\eta(x_{ij}, p_j)$ jest liczbą obserwacji innych niż i-ta należących do przedziału $[(1 - p_j)x_{ij}, (1 + p_j)x_{ij}]$.

- Rekord $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \in \mathbb{R}^m$ jest niebezpieczny gdy kombinacja $(x_{i1}, x_{i2}, \ldots, x_{ih})$ jest niebezpieczna lub $w_j(i) = 1$ dla przynajmniej jednego $j \in \{h + 1, h + 2, \ldots, m\}$ (tzn. $\sum_{j=h+1}^{m} w_j(i) > 0$), $i = 1, 2, \ldots, n$.

Koncepcja kompleksowej miary ryzyka ujawnienia

Miara ryzyka

- Miara ryzyka dla i-tego rekordu (w badaniu pełnym):

$$r_i \stackrel{df}{=} \frac{1}{2}\left(\frac{1}{f_i} + \frac{\sum_{j=h+1}^{m} w_j(i)}{m-h}\right),$$

  gdzie f_i to częstość kombinacji $(x_{i1}, x_{i2}, \ldots, x_{ih})$ w całej bazie, $i = 1, 2, \ldots, n$. $r_i \in [0, 1]$ (im większa wartość, tym większe ryzyko)

- Ryzyko globalne:

$$r \stackrel{df}{=} \frac{1}{n}\sum_{i=1}^{n} r_i \in [0, 1].$$

Koncepcja kompleksowej miary ryzyka ujawnienia

Miara ryzyka

- Miara ryzyka dla i-tego rekordu w przypadku badania reprezentacyjnego z wagami ω_i:

$$w_j^*(i) \stackrel{df}{=} \begin{cases} 1 & \text{gdy } \eta(x_{ij}, p_j) < k \text{ or } [\omega_i \eta(x_{ij}, p_j)] < k, \\ 0 & \text{w przeciwnym razie,} \end{cases}$$

gdzie $[a]$ to część całkowita $a \in \mathbb{R}$,

$$r_i^* \stackrel{df}{=} \frac{1}{2}\left(\frac{f_i}{f_i + \hat{F}_i(f_i - 1)} + \frac{\sum_{j=h+1}^{m} w_j^*(i)}{m-h}\right) \in [0,1],$$

gdzie F_i to częstość kombinacji $(x_{i1}, x_{i2}, \ldots, x_{ih})$ w populacji a f_i – w próbie

$$r^* \stackrel{df}{=} \frac{1}{n}\sum_{i=1}^{n} r_i^*.$$

Wyznaczanie wkładu w ryzyko

Narzędzia gier kooperacyjnych

- Niech $n \in \mathbb{N}, n \geq 2$. n-osobowa *gra kooperacyjna*) to para (N, v), gdzie $N = \{1, 2, \ldots, n\}$ jest zbiorem graczy a $v : 2^N \to \mathbb{R}$ with $v(\emptyset) = 0$ nazywa się *funkcją charakterystyczną*. $v(S)$ to *wartość* koalicji $S \subseteq N$. Niech $|S| = s$.
- Dwóch graczy $i, j \in N$, $i \neq j$, nazywa się *symetrycznymi* w v jeżeli $v(S \cup \{i\}) = v(S \cup \{j\})$ dla każdego $S \subseteq N \backslash \{i, j\}$. Gracz $i \in N$ jest *niemy* gdy $v(S \cup \{i\}) = v(S) + v(\{i\})$ dla każdego $S \subseteq N \backslash \{i\}$. Gracz i nazywa się *graczem zerowym* jeśli $v(S \cup \{i\}) = v(S)$ dla każdego $S \subseteq N \backslash \{i\}$. Gracz jest $\mathcal{A}$-*zerowy* gdy $\mathcal{A}_v(S) = 0$ dla każdego $S \subseteq N$, takiego, że $i \in S$, gdzie $\mathcal{A}_v(S) \stackrel{df}{=} \frac{1}{s} \sum_{k \in N : k \in S} (v(S) - v(S \backslash \{k\}))$.
- Niech G_N będzie zbiorem wszystkich n-osobowych gier kooperacyjnych v.
- Gra *zerowa* $\underline{\mathbf{o}} \in G_N$ jest grą trywialną, w której $\underline{\mathbf{o}}(S) = 0$ dla każdego $S \subseteq N$.
- Suma gier of $v, w \in G_N$, $v + w \in G_N$, to gra $(v + w)(S) = v(S) + w(S)$ dla wszystkich $S \subseteq N$.

Wyznaczanie wkładu w ryzyko

Narzędzia gier kooperacyjnych

- Iloczyn gier, $v \cdot w \in G_N$, określa się jako $(v \cdot w)(S) = v(S) \cdot w(S)$ dla wszystkich $S \subseteq N$ for all $S \subseteq N$.
- Jeżeli $v \in G_N$ i $a \in \mathbb{R}, a \neq 0$, to definiujemy grę $(a \cdot v) \in G_N$ jako $(a \cdot v)(S) = a \cdot v(S)$ dla wszystkich $S \subseteq N$.
- Niech $T \subseteq N$, $T \neq \emptyset$. Grę *jednomyślności* $u_T \in G_N$ definiuje się jako $u_T(S) = 1$ gdy $T \subseteq S$ i $u_T(S) = 0$ w przeciwnym razie, dla każdego $S \subseteq N$. Gdy $v(K) \leq v(S)$ dla każdych $K \subseteq S \subseteq N$, wtedy gra v jest nazywana *monotoniczną*. Jeżeli ρ jest permutacją zbioru N, wtedy określamy grę $\rho v \in G_N$ jako $\rho v(\rho(S)) = v(S)$ dla każdego $S \subseteq N$.
- A *rozwiązaniem* (*wartością*) $\varphi(v) = (\varphi_1(v), \varphi_2(v), \ldots, \varphi_n(v))$ na G_N jest przekształcenie wektorowe $\varphi : G_N \to \mathbb{R}^n$, które dla każdej gry $v \in G_N$ jednoznacznie wyznacza rozkład dobra ogółem dostępnego dla graczy $1, 2, \ldots, n$ poprzez ich uczestnictwo w grze v. Zatem $\varphi_i(v)$ reprezentuje *wypłatę* gracza i w grze v na N.

Wyznaczanie wkładu w ryzyko

Własności rozwiązań gier kooperacyjnych

- W naszym studium ważne będą następujące własności rozwiązań gier kooperacyjnych (zwane także *aksjomatami*):
  - *Efektywność*: Rozwiązanie φ jest efektywne gdy $\sum_{i=1}^{n} \varphi_i(v) = v(N)$ dla każdego $v \in G_N$,
  - *Symetria*: Rozwiązanie φ jest symetryczne gdy $\varphi_{\rho(i)}(\rho v) = \varphi_i(v)$ dla każdej permutacji ρ zbioru N i każdego $v \in G_N$, $i = 1, 2, \ldots, n$,
  - *Równoprawność*: Jeżeli gracze $i, j \in N$ są symetryczni wówczas $\varphi_i(v) = \varphi_j(v)$ dla każdego $v \in G_N$,
  - *Liniowość*: Dla każdego $v, w \in G_N$ i $a, b \in \mathbb{R}$ $\varphi(av + bw) = a\varphi(v) + b\varphi(w)$,
  - *Monotoniczność*: Jeżeli gra $v \in G_N$ jest monotoniczna, to $\varphi_i(v) \geq 0$ dla każdego $i \in N$,
  - *Aksjomat gracza niemego*: Jeżeli gracz $i \in N$ jest niemy, to $\varphi_i(v) = v(\{i\})$ dla każdej gry $v \in G_N$,
  - *Aksjomat gracza $\mathcal{A}$-zerowego*: Jeśli gracz $i \in N$ jest $\mathcal{A}$-zerowy, wtedy $\varphi_i(v) = 0$, dla każdego $v \in G_N$.
- Własność równoprawności jest słabsza niż symetria.

Wyznaczanie wkładu w ryzyko

Rozwiązania gier kooperacyjnych

- Najpopularniejszym i szeroko stosowanym rozwiązaniem gry kooperacyjnej jest *Wartość Shapleya*, wprowadzona przez Shapleya (1953). Dla gracza $i \in N$ gry $v \in G_N$ jest ona definiowana jako

$$Sh_i(v) = \sum_{S \subseteq N} \frac{s!(n-s-1)!}{n!} \left(v(S \cup \{i\}) - v(S)\right).$$

- Pierwszą aksjomatyzację (twierdzenie o zestawie własności, które jednoznacznie wskazują dane rozwiązanie) dla tego rozwiązania wykazał sam twórca (Shapley (1953)). Wedle niej $Sh(v)$ jest jedynym rozwiązaniem na G_N spełniającym warunki efektywności, symetrii, liniowości i gracza niemego. Young (1985) pokazał, że symetrię można zastąpić słabszą równoprawnością.
- W kolejnych dekadach powstał cały szereg różnorodnych aksjomatyzacji tego rozwiązania.

Wyznaczanie wkładu w ryzyko

Rozwiązania gier kooperacyjnych

- Radzik and Driessen (2013) rozpatrywali specjalną rodzinę rozwiązań uogólniającą ideę wartości Shapleya. Chodzi o rozwiązania spełniające warunki efektywności, symetrii i liniowości (w skrócie: wartości ESL):

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{s!(n-s-1)!}{n!} (q_{s+1} v(S \cup \{i\}) - q_s v(S)),$$

  gdzie $q_s \in \mathbb{R}$, $s = 1, 2, \ldots, n$ z $q_0 = 0$ i $q_n = 1$.

- Rozwiązanie klasy ESL spełnia aksjomat monotoniczności wtedy i tylko wtedy gdy $q_n = 1$ oraz $0 \leq q_s \leq 1$ dla każdego $s = 1, 2, \ldots, n-1$.
- Dla wartości Shapleya jest $q_s = 1$ dla każdego $s = 1, 2, \ldots, n-1$. Jeżeli

$$q_n = 1 \quad \text{i} \quad q_s = \frac{1}{s+1}, \quad s = 1, 2, \ldots, n-1,$$

  to takie rozwiązanie klasy ESL nazywa się *wartością solidarnościową*. Wartość solidarnościowa jest jedynym rozwiązaniem na G_N posiadającym własności efektywności, symetrii, liniowości i gracza $\mathcal{A}$-zerowego. Oczywiście spełnia ono także aksjomat monotoniczności.

Wyznaczanie wkładu w ryzyko

Wkład do ryzyka ujawnienia

- Rozważmy zbiór zmiennych $M = \{X_1, X_2, \ldots, X_m\}$ (dla uproszczenia: $M = \{1, 2, \ldots, m\}$) rozumiany tutaj jako zbiór "graczy". Gra kooperacyjna $v_i \in G_M$ jest zdefiniowana jako

$$v_i(S) \stackrel{df}{=} \begin{cases} 1 & \text{gdy } \{x_{ij} : j \in S \cap \{1, 2, \ldots, h\}\} \text{ jest niebezpieczna lub } \max\limits_{j \in S \cap \{h+1, h+2, \ldots, m\}} w_j^*(i) = 1, \\ 0 & \text{w przeciwnym razie,} \end{cases}$$

  dla każdego $S \subseteq M$, $i = 1, 2, \ldots, n$.

- Równoważnie:

$$v_i(S) = \max \left\{ 1 - \prod_{T \in \mathcal{Q}_i} (1 - u_T(S)); 1 - \prod_{j \in S \cap \{h+1, h+2, \ldots, m\}} (1 - w_j^*(i)) \right\},$$

  gdzie $\mathcal{Q}_i$ jest zbiorem minimalnych niebezpiecznych podzbiorów zbioru $\{x_{i1}, x_{i2}, \ldots, x_{ih}\}$, tzn. takich, że każda $K \in T$ jest niebezpieczna ale żaden właściwy podzbiór K nie jest niebezpieczny, zaś u_T jest grą jednomyślności na M, $S \subseteq M$, $i = 1, 2, \ldots, n$.

Wyznaczanie wkładu w ryzyko

Wkład do ryzyka ujawnienia

- Dla każdego $i \in N$ gra v_i na M jest monotoniczna. Stąd wartość Shapleya i wartość solidarnościowa dla i są zawsze nieujemne, efektywne i symetryczne – a więc mogą dobrze odzwierciedlać wkład poszczególnych zmiennych do ryzyka ujawnienia.
- Niech φ będzie wartością Shapleya lub solidarnościową na G_M. Wtedy $\varphi_j(v_i)$ będzie wypłatą – a więc tutaj wkładem do ryzyka indywidualnego dla i-tego rekordu – zmiennej X_j, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Całkowity wkład X_j do ryzyka ogółem można zatem wyznaczyć jako

$$v \stackrel{df}{=} \frac{1}{n_0} \sum_{i \in N_0} v_i.$$

- Stąd, mocą liniowości

$$\varphi_j(v) = \frac{1}{n_0} \sum_{i \in N_0} \varphi_j(v_i) \tag{1}$$

dla każdego $j = 1, 2, \ldots, m$, gdzie $N_0 = \{i \in \{1, 2, \ldots, n\} : v_i(M) \neq 0\}$ and $n_0 = |N_0|$.

Zastosowania empiryczne

Eksperyment symulacyjny

- Wygenerowano zbiór $n = 500$ jednostek charakteryzowanych przez $m = 7$ zmiennych, z których $h = 4$ są kategorialne, a $m - h = 3$ – ciągłe.
- Zmienne kategorialne:
  - X_1 – przyjmuje wartości 1 lub 2 z jednakowym prawdopodobieństwem,
  - X_2 – przyjmuje wartości ze zbioru $\{1, 2, \ldots, 7\}$ z jednakowym prawdopodobieństwem,
  - X_3 – liczby losowe z rozkładu wielomianowego z prawdopodobieństwami 0,052, 0,014, 0,183, 0,049, 0,217, 0,314, 0,171,
  - X_4 – liczby losowe z rozkładu wielomianowego z prawdopodobieństwami 0,028, 0,042, 0,086, 0,038, 0,591, 0,215.
- Zmienne ciągłe:
  - X_5 – liczby losowe z rozkładu jednostajnego na $[2000, 10000]$,
  - X_6 – liczby losowe z rozkładu normalnego z wartością oczekiwaną 50 i odchyleniem standardowym równym 10,
  - X_7 – liczby losowe z rozkładu Fishera-Snedecora z 5 i 20 stopniami swobody.

Zastosowania empiryczne

Eksperyment symulacyjny

- Każda kategoria każdej zmiennej kategorialnej występuje o co najmniej 5 rekordów. Pomimo tego, ryzyko indywidualne waha się od 0,0909 do 1,0000 (I kwartyl 0,1667, mediana – 0,3333, średnia arytmetyczna – 0,4120 i trzeci kwartyl – 0,5000).
- 100 rekordów (20,0%) narusza 2-anonimowość, 178 (35,6%) – 3-anonimowość oraz 317 (63,4%) – 5-anonimowość. Ryzyko globalne wynosi 41,2% a oczekiwana liczba reidentyfikacji – 206).

Tab. 1: Podstawowe dane o zmiennych kategorialnych

Zmienna	Liczba kategorii	Średnia liczebność	Liczebność najmniejsza (>0)
X_1	2	250.000	234
X_2	7	71.429	66
X_3	7	71.429	5
X_4	6	83.333	9

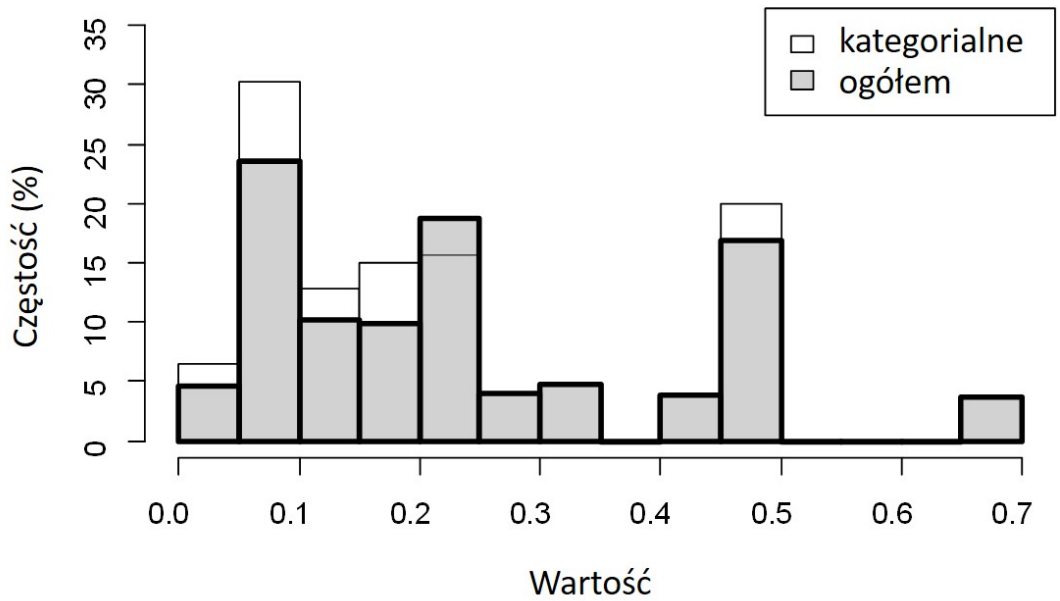
Źródło: Obliczenia własne z wykorzystaniem pakietu sdcMicro *środowiska R.*

Zastosowania empiryczne

Eksperyment symulacyjny

- Przyjęto: $k = 3$ dla k-anonimowości oraz bezpieczeństwa dla zmiennych ciągłych; precyzja dla zmiennych ciągłych 0,01 dla X_5, 0,03 dla X_6 oraz 0.02 dla X_7.

Histogram częstości dla komponentu ryzyka indywidualnego – zmienne kategorialne oraz ryzyka indywidualnego ogółem

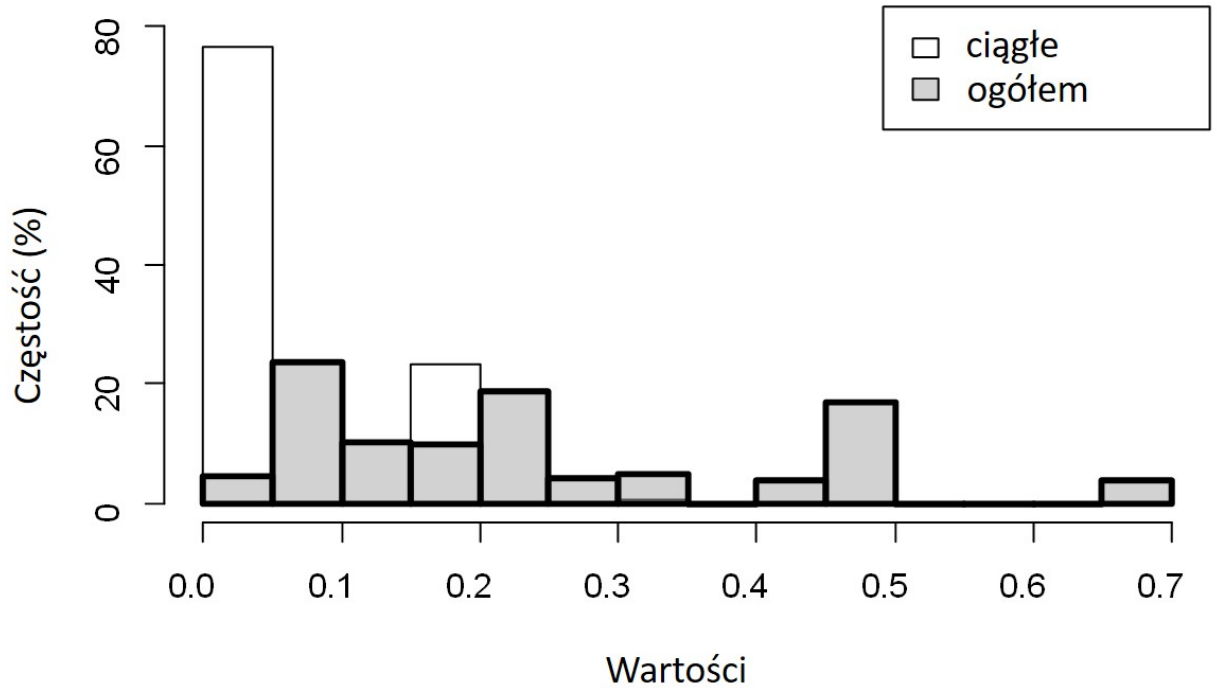


Źródło: Obliczenia własne z wykorzystaniem oryginalnego skryptu napisanego w środowisku R.

Zastosowania empiryczne

Eksperyment symulacyjny

Histogram częstości dla komponentu ryzyka indywidualnego – zmienne ciągłe oraz ryzyka indywidualnego ogółem

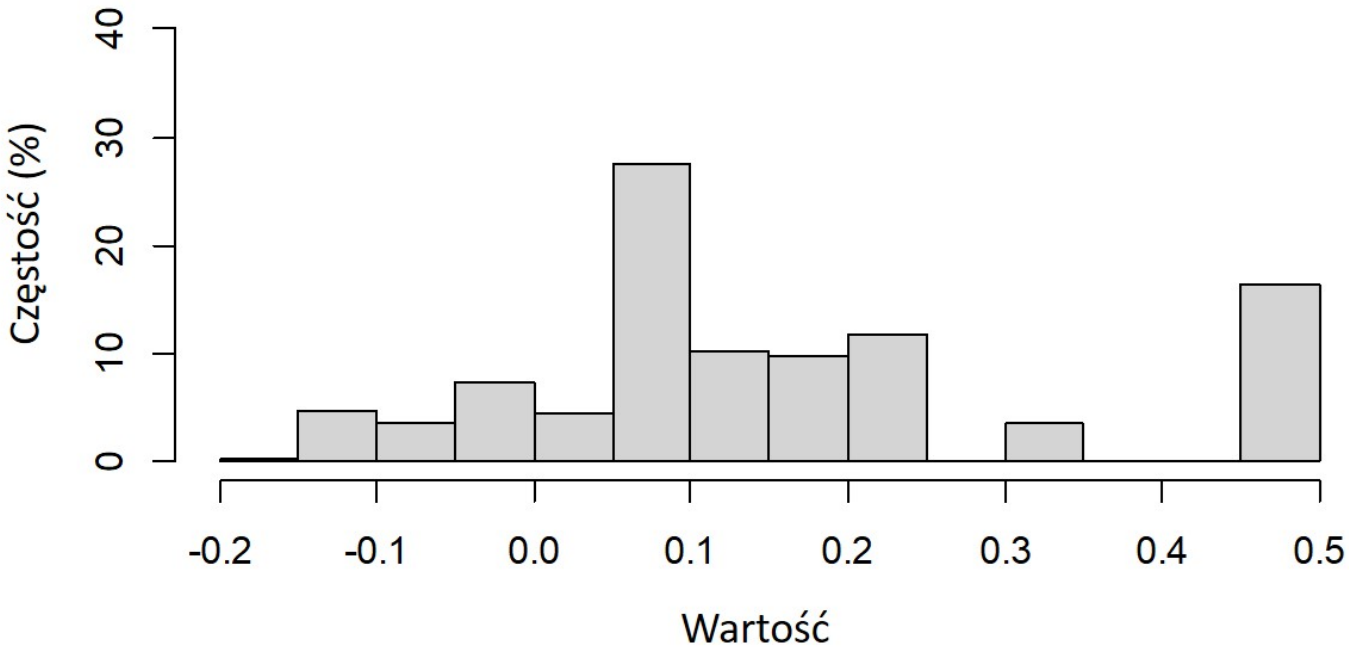


Źródło: Obliczenia własne z wykorzystaniem oryginalnego skryptu napisanego w środowisku R.

Zastosowania empiryczne

Eksperyment symulacyjny

Histogram częstości różnicy pomiędzy komponentami ryzyka pochodzącymi od zmiennych kategorialnych i ciągłych.

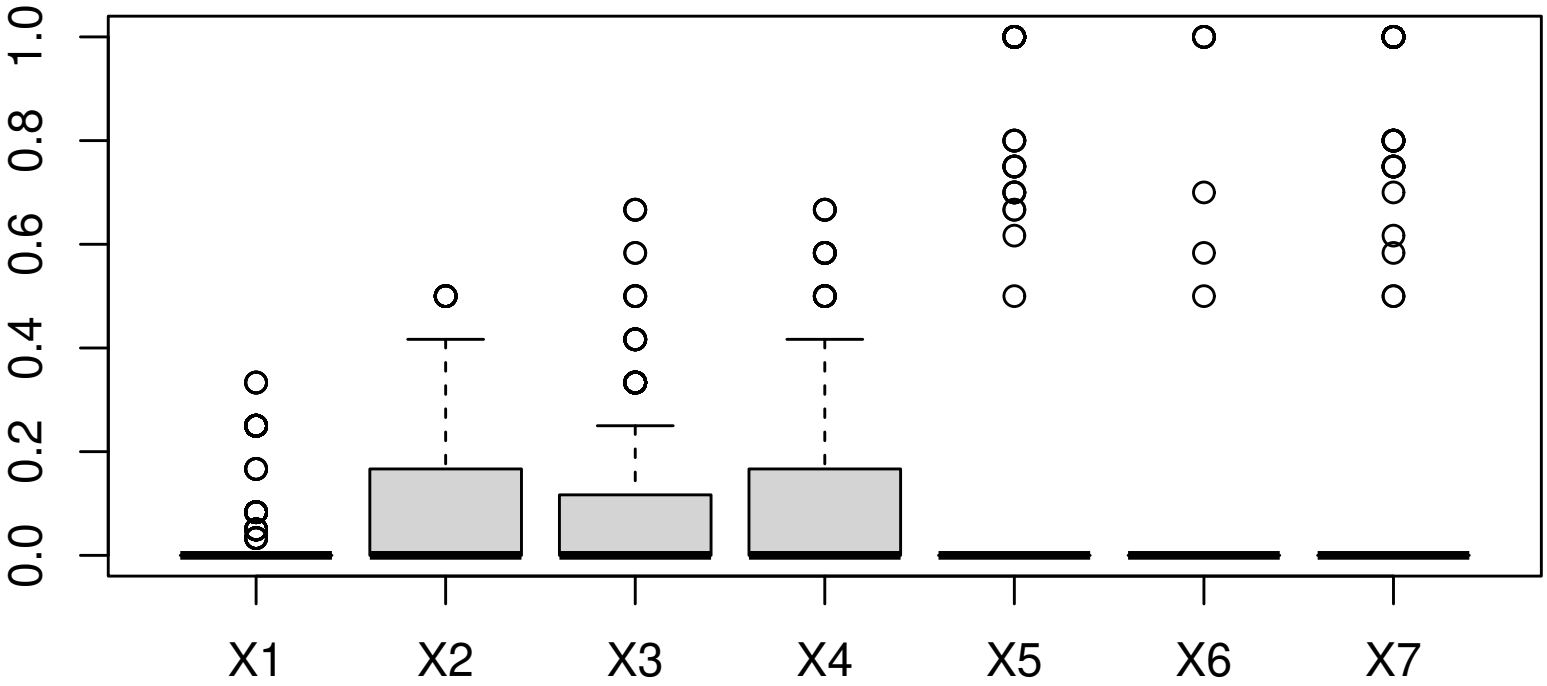


Źródło: Obliczenia własne z wykorzystaniem oryginalnego skryptu napisanego w środowisku R.

Zastosowania empiryczne

Eksperyment symulacyjny

Rozkład wartości Shapleya dla poszczególnych zmiennych wśród jednostek

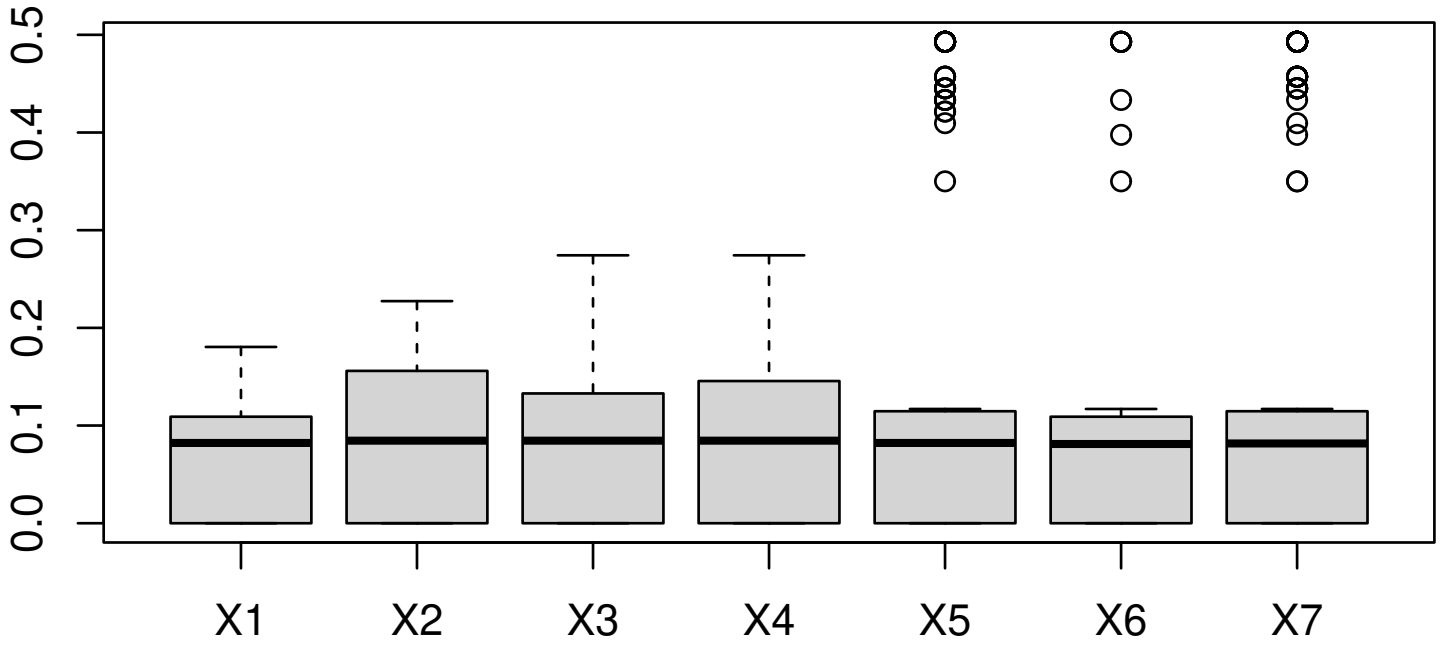


Źródło: Obliczenia własne z wykorzystaniem oryginalnego skryptu napisanego w środowisku R.

Zastosowania empiryczne

Eksperyment symulacyjny

Rozkład wartości solidarnościowej dla poszczególnych zmiennych wśród jednostek



Źródło: Obliczenia własne z wykorzystaniem oryginalnego skryptu napisanego w środowisku R.

Zastosowania empiryczne

Eksperyment symulacyjny

Wartości: Shapleya i solidarnościowa dla analizowanych zmiennych

Zmienna	Wartość Shapleya	Wartość solidarnościowa
X_1	0.0697	0.1133
X_2	0.1698	0.1357
X_3	0.1627	0.1342
X_4	0.1794	0.1380
X_5	0.1745	0.1747
X_6	0.0301	0.1126
X_7	0.2139	0.1915

Źródło: Obliczenia własne z wykorzystaniem oryginalnego skryptu napisanego w środowisku R.

Zastosowania empiryczne

Analiza na danych rzeczywistych

- Mikrodane z Badania Osób Dorosłych (BOD) w Polsce w ramach Bilansu Kapitału Ludzkiego w 2019 r.(https://www.parp.gov.pl/component/site/site/bilans-kapitalu-ludzkiego).
- 2533 rekordy i 47 zmiennych (+ 3 identyfikatory). Zmienne analizowane:
  - Woj – województwo (2–Dolnośląskie, 4–Kujawsko-pomorskie, 6–Lubelskie, 8–Lubuskie, 10–Łódzkie, 12–Małopolskie, 14–Mazowieckie, 16–Opolskie, 18–Podkarpackie, 20–Podlaskie, 22–Pomorskie, 24–Śląskie, 26–Świętokrzyskie, 28–Warmińsko-Mazurskie, 30–Wielkopolskie, 32–Zachodniopomorskie),
  - Makr – makroregion (1–Centralny, 2–Północno-zachodni, 3–Północny, 4–Południowo-zachodni, 5–Południowy, 6–Województwo mazowieckie, 7–Wschodni),
  - Mzam – miejsce zamieszkania (1–wieś, 2–miasto do 9999 mieszkańców, 3–miasto 10000–19999 mieszkańców, 4–miasto 20000–49999 mieszkańców, 5–miasto 50000–99999 mieszkańców, 6–miasto 100000–199999 mieszkańców, 7–miasto 200000–499999 mieszkańców, 8–miasto 500000 lub więcej mieszkańców, bez Warszawy, 9–Warszawa),
  - Wiek – wiek w latach,
  - Plec – płeć (0–mężczyzna, 1–kobieta),
  - Sytrp – sytuacja na rynku pracy (0–pracujący, 1–bezrobotny, 2–nieaktywny),
  - Wykszt – poziom wykształcenia (0–niepełne podstawowe, 1–podstawowe, 2–gimnazjalne, 3–zasadnicze zawodowe, 4–średnie ogólne (LO), 5–średnie zawodowe (liceum zawodowe/profilowane), 6 – średnie zawodowe (technikum), 7–średnie zawodowe (szkoła policealna), 8– średnie zawodowe (inna szkoła, nie wyższa), 9–wyższe licencjackie, 10–wyższe inżynierskie, 11–wyższe magisterskie, 12–wyższe magistersko-inżynierskie, 13– wyższe (studia podyplomowe), 14–wyższe (MBA), 15–wyższe (doktorat)),
  - Wldz – prowadzenie własnej działalności gospodarczej (1–tak, 0–nie).

Zastosowania empiryczne

Analiza na danych rzeczywistych

Podstawowe dane o zmiennych kategorialnych z BOD

Zmienna	Liczba Kategorii	Średnia liczebność	Liczebność najmniejsza (>0)
Woj	13	30.769	1
Makr	7	57.143	4
Mzam	8	50.000	19
Plec	2	200.000	195
Sytrp	3	133.333	14
Wykszt	15	526.667	1
Wldz	2	200.000	51

Źródło: Obliczenia własne z wykorzystaniem pakietu sdcMicro *środowiska R.*

Zastosowania empiryczne

Analiza na danych rzeczywistych

Podstawowe statystyki opisowe dla ryzyka ogółem i jego komponentów w BOD

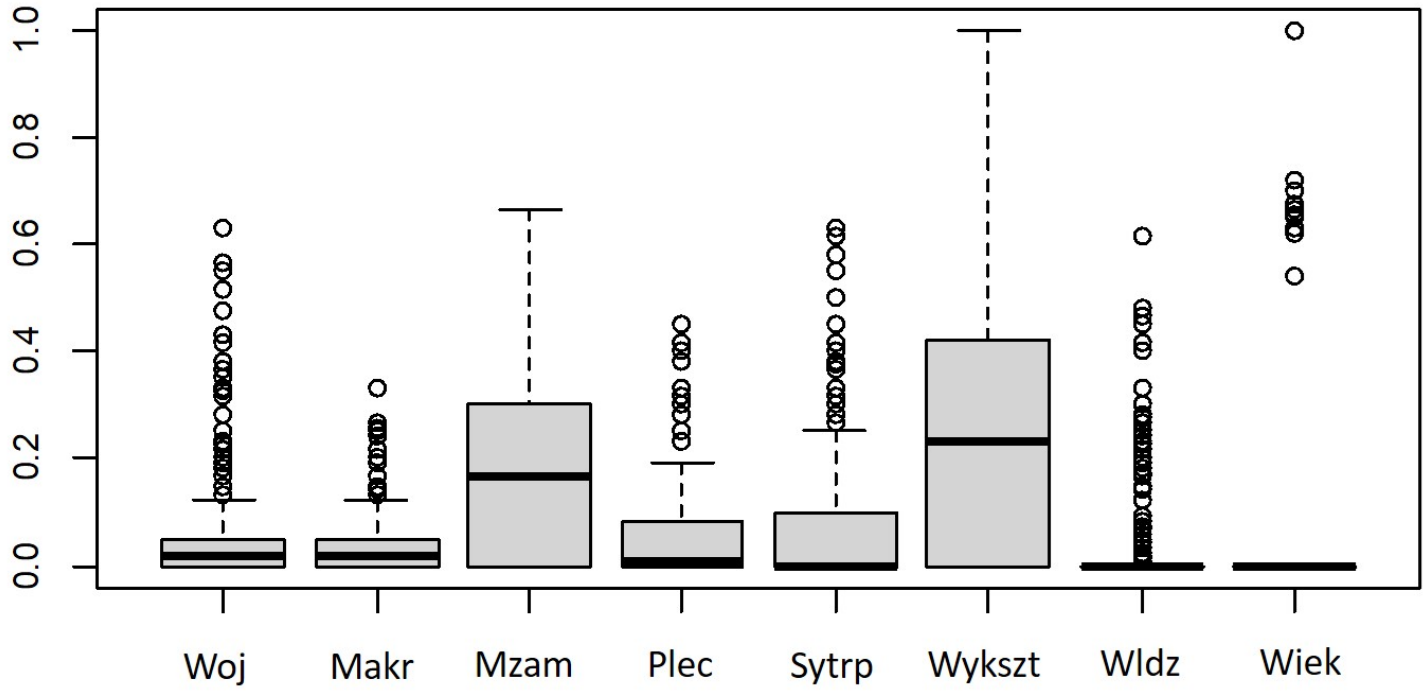
Rodzaj ryzyka	Minimum	1 kwartyl	Mediana	Średnia	3 kwartyl	Maksimum
Wagi dla próby						
Tylko zmienne kategorialne	0.03737	0.19373	0.50000	0.34918	0.50000	0.50000
Tylko zmienne ciągłe	0.00000	0.00000	0.00000	0.05625	0.00000	0.50000
Ogółem	0.03737	0.20825	0.50000	0.40543	0.50000	1.00000
Wagi dla populacji						
Tylko zmienne kategorialne	0.000004	0.000032	0.500000	0.251271	0.500000	0.500000
Tylko zmienne ciągłe	0.000000	0.000000	0.000000	0.021250	0.000000	0.500000
Ogółem	0.000004	0.000053	0.500000	0.272521	0.500000	1.000000

Źródło: Obliczenia własne z wykorzystaniem oryginalnego skryptu napisanego w środowisku R.

Zastosowania empiryczne

Analiza na danych rzeczywistych

Rozkład wartości Shapleya dla poszczególnych zmiennych wśród jednostek – BOD

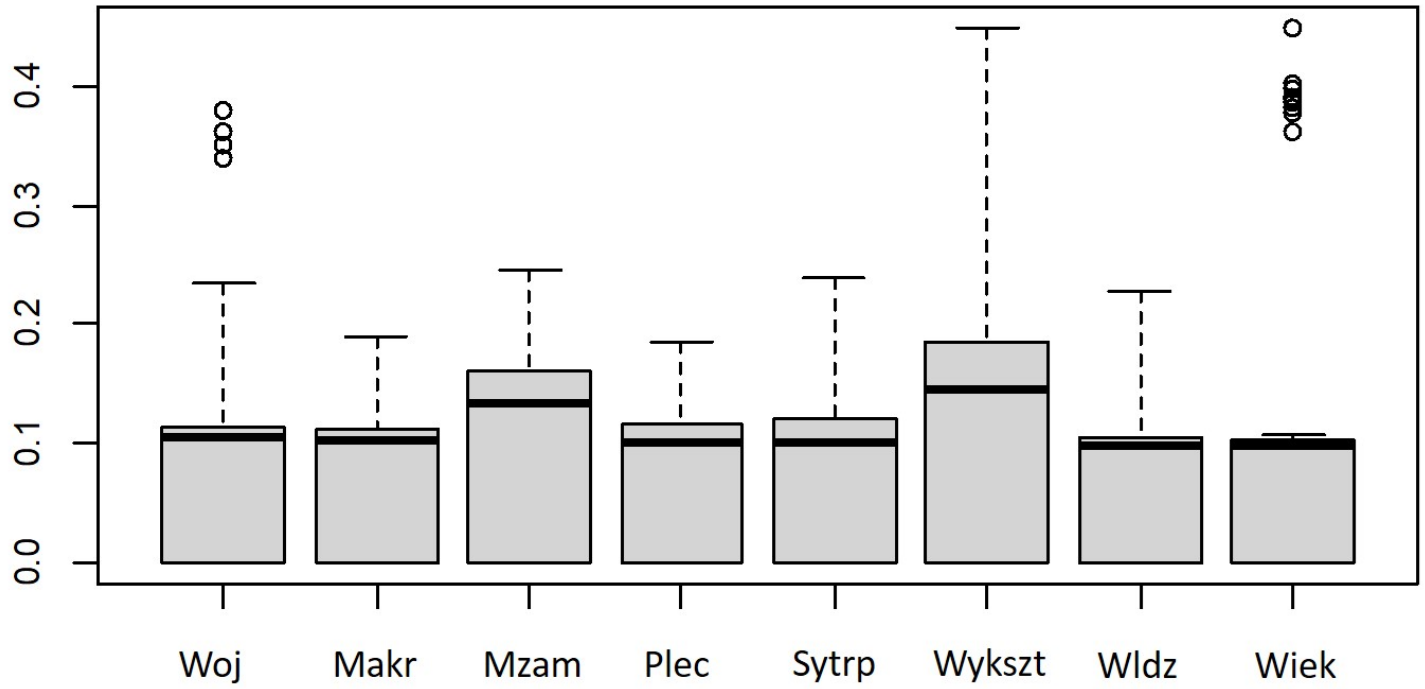


Źródło: Obliczenia własne z wykorzystaniem oryginalnego skryptu napisanego w środowisku R.

Zastosowania empiryczne

Analiza na danych rzeczywistych

Rozkład wartości solidarnościowej dla poszczególnych zmiennych wśród jednostek – BOD



Źródło: Obliczenia własne z wykorzystaniem oryginalnego skryptu napisanego w środowisku R.

Zastosowania empiryczne

Analiza na danych rzeczywistych

Wartości: Shapleya i solidarnościowa dla analizowanych zmiennych

Variable	Shapley value	Solidarity value
Woj	0.0727	0.1181
Makr	0.0497	0.1096
Mzam	0.2520	0.1486
Plec	0.0843	0.1124
Sytrp	0.1003	0.1172
Wykszt	0.3487	0.1722
Wldz_act	0.0482	0.1065
Wiek	0.0441	0.1155

Źródło: Obliczenia własne z wykorzystaniem oryginalnego skryptu napisanego w środowisku R.

Wnioski

- Proponowana miara może być użyteczna w ocenie potrzeb dotyczących ochrony poufności danych oraz efektywności stosowania narzędzi SDC.
- Jej istotną zaletą jest łączenie wpływu zarówno zmiennych kategorialnych, jak i ciągłych na ogólne bezpieczeństwo rekordu. Dzięki oparciu się w obu przypadkach na częstości miara jest spójna i łatwo interpretowalna oraz dekomponowalna. Może także być podstawą do wyznaczania specjalnych poziomów ryzyka, np. ryzyka hierarchicznego.
- Wykorzystanie rozwiązań gier kooperacyjnych celem oceny wpływu poszczególnych zmiennych na ryzyko ujawnienia zwiększa możliwości efektywnego wsparcia procesu SDC poprzez wskazanie zmiennych, na które należy zwrócić szczególną uwagę w kontekście niwelacji ryzyka ujawnienia. Wartość Shapleya daje tutaj bardziej zróżnicowane rezultaty niż wartość solidarnościowa.
- Jako nieco kontrowersyjne może być postrzegane łączenie ujawniania tożsamości z ujawnieniem atrybutu. Poza tym wyznaczanie wartości Shapleya i solidarnościowej jest bardzo czasochłonne. Być może jednak stosowny algorytm można by udoskonalić wykorzystując niektóre własności gier monotonicznych.

Literatura

- Elliot, M. J., Manning, A. M., & Ford, R. W. (2002). A computational algorithm for handling the special uniques problem. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05), 493–509.
- Hundepool, A., Domingo–Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & de Wolf, P. (2012). *Statistical Disclosure Control*. John Wiley & Sons, Ltd.
- Młodak, A., Pietrzak, M., Klimanek, T., Józefowski, T. & Lańduch, P. (2023), *Poufność a użyteczność informacji statystycznych. Dylematy ochrony udostępnianych danych*. Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań. https://doi.org/10.18559/978-83-8211-168-2.
- Nowak, A. S., & Radzik, T. (1994). A solidarity value for *n*-person transferable utility games. *International Journal of Game Theory*, 23, 43–48.
- Radzik, T.,& Driessen, T. S. H. (2013). On a family of values of TU-games generalizing the Shapley value. *Mathematical Social Sciences*, 65, 105–111.
- Shapley, L. S. (1953). A Value for *n*–person Game. *Annals of Mathematical Studies*, 28, 307–317.
- Templ, M. (2017). *Statistical Disclosure Control for Microdata. Methods and Applications in R*. Springer International Publishing AG, Cham, Switzerlan
- Young, H. P. (1985). Monotonic Solutions of Cooperative Games. *International Journal of Game Theory*, 14, 65–72.

Urząd Statystyczny
w Poznaniu

Dziękuję za uwagę!