DOI: 10.5604/01.3001.0014.1145

PRZEGLĄD STATYSTYCZNY
R. LXIII – ZESZYT 1 – 2016

MAREK WALESIAK[1]

WYBÓR GRUP METOD NORMALIZACJI WARTOŚCI ZMIENNYCH W SKALOWANIU WIELOWYMIAROWYM

1. WPROWADZENIE

Normalizację przeprowadza się dla macierzy danych metrycznych, tzn. gdy zmienne opisujące obiekty badania mierzone są na skali interwałowej lub ilorazowej. Charakterystykę skal pomiaru zawarto m.in. w pracach (Stevens, 1946; Walesiak, 2011, s. 13–16).

Porównanie metod normalizacji wartości zmiennych może być rozpatrywane z punktu widzenia zastosowania konkretnej metody statystycznej analizy wielowymiarowej. W sytuacji, gdy w badaniu będą wykorzystywane metody analizy skupień, porządkowania liniowego i skalowania wielowymiarowego, zmienne muszą być sprowadzone do porównywalności poprzez transformacje normalizacyjne. Dla analizy skupień badania takie przeprowadzili Milligan, Cooper (1988), Schaffer, Green (1996), Walesiak, Dudek (2016), a dla metod porządkowania liniowego Walesiak (2015), Kukuła, Luty (2015)[2]. Inne metody statystycznej analizy wielowymiarowej (analiza regresji, drzewa klasyfikacyjne i regresyjne, *conjoint analysis*, analiza czynnikowa[3], analiza dyskryminacyjna, analiza korelacji kanonicznej, analiza wariancji i kowariancji) nie wymagają uprzedniej transformacji normalizacyjnej.

W artykule zaproponowano procedurę badawczą pozwalającą na wyodrębnienie grup metod normalizacji wartości zmiennych prowadzących do zbliżonych wyników skalowania wielowymiarowego. Propozycja dotyczy problemu wyboru metod normalizacji wartości zmiennych w skalowaniu wielowymiarowym bazującym na macierzy danych metrycznych.

---

[1] Uniwersytet Ekonomiczny we Wrocławiu, Wydział Ekonomii, Zarządzania i Turystyki, Katedra Ekonometrii i Informatyki, ul. Nowowiejska 3, 58-500 Jelenia Góra, Polska, e-mail: marek.walesiak@ue.wroc.pl.

[2] Badanie tutaj przeprowadzone było szersze, ponieważ obejmowało wybór metody (procedury) porządkowania liniowego. Podobne badania wcześniej przeprowadził Grabiński (1984) stosując inne kryteria wyboru.

[3] Metody normalizacyjne będące przekształceniem liniowym (3) nie zmieniają wartości współczynnika korelacji liniowej Pearsona (por. Jajuga, Walesiak, 2000, s. 111), który jest wykorzystywany w analizie czynnikowej.

2. SKALOWANIE WIELOWYMIAROWE NA PODSTAWIE MACIERZY DANYCH METRYCZNYCH – OGÓLNY SCHEMAT POSTĘPOWANIA

Punktem wyjścia skalowania wielowymiarowego jest macierz odległości (niepodobieństw) między obiektami w przestrzeni m-wymiarowej $[\delta_{ik}]$, gdzie $i,k = 1,\ldots,n$ oznacza numer obiektu. Wśród metod wyznaczania macierzy odległości $[\delta_{ik}]$ wyróżnia się (por. np. Borg, Groenen, 2005, s. 111–133; Zaborski, 2001, s. 40–50):

1. Bezpośrednie – np. poprzez porównywanie obiektów parami pod względem ich niepodobieństwa przez poszczególnych respondentów.
2. Pośrednie. Punktem wyjścia jest tutaj macierz danych $[x_{ij}]$ (gdzie: x_{ij} – obserwacja j-tej zmiennej w i-tym obiekcie, $j = 1,\ldots,m$ – numer zmiennej). Obserwacje na zmiennych uzyskujemy stosując szacowanie na skalach pomocniczych (respondenci oceniają tutaj poszczególne obiekty dla każdej zmiennej) lub z wtórnych źródeł danych. Następnie oblicza się odległości między obiektami z wykorzystaniem miar odległości (dla danych metrycznych stosuje się wcześniej normalizację wartości zmiennych).

Skalowanie wielowymiarowe jest metodą reprezentacji macierzy odległości między obiektami w przestrzeni m-wymiarowej $[\delta_{ik}]$ w macierz odległości między obiektami w przestrzeni q-wymiarowej $[d_{ik}]$ $(q < m)$ w celu graficznej prezentacji (wizualizacji) relacji zachodzących między badanymi obiektami oraz określenia (interpretacji) treści q wymiarów. Wymiary q nie są bezpośrednio obserwowalne. Mają one charakter zmiennych ukrytych, które pozwalają na wyjaśnienie podobieństw i różnic między badanymi obiektami. Ze względu na możliwość graficznej prezentacji wyników zazwyczaj q wynosi 2 lub 3.

W algorytmach skalowania wielowymiarowego stosowane są różne miary dopasowania STRESS (ang. *STandardized REsidual Sum of Squares* – standaryzowana suma kwadratów reszt). W monografii (Borg, Groenen, 2005, s. 250–254) prezentowane są m.in. funkcje: STRESS-1 Kruskala, STRESS-2 Kruskala i Carrolla, współczynnik alienacji Guttmana-Lingoesa, S-STRESS Takane, Younga i De Leeuw. Zagadnieniu oceny wartości miar dopasowania, a co za tym idzie wyborowi liczby wymiarów q skalowania Borg, Groenen (2005) poświęcili podrozdział 3.5 (s. 47–55). Mowa jest tutaj m.in. o wykresie osypiska (ang. *scree test*), prostych normach oceny miar dopasowania. Dodatkowym ważnym kryterium jest interpretowalność wymiarów (osi) skalowania wielowymiarowego (zob. Borg, Groenen, 2005, s. 55).

Dla danego zbioru obiektów $A = \{A_1,\ldots,A_n\}$ oraz odległości (niepodobieństw) δ_{ik} między obiektami A_i oraz A_k w przestrzeni m-wymiarowej poszukuje się takiego odwzorowania zbioru obiektów w zbiór punktów w przestrzeni q-wymiarowej, aby (Borg, Groenen, 2005, s. 39):

$$d_{ik} \approx \hat{d}_{ik} = f(\delta_{ik}), \tag{1}$$

gdzie:

d_{ik} – odległość między obiektami A_i oraz A_k (punktami $\mathbf{x}_i$ oraz $\mathbf{x}_k$) w przestrzeni q-wymiarowej,

$\hat{d}_{ik}$ – funkcja regresji między d_{ik} a δ_{ik}.

Ogólny schemat postępowania w skalowaniu wielowymiarowym zbioru obiektów przeprowadzanych na podstawie danych metrycznych jest następujący:

$$P \rightarrow A \rightarrow X \rightarrow N \rightarrow S \rightarrow I, \tag{2}$$

gdzie:
P – wybór problemu badawczego,
A – wybór obiektów,
X – dobór zmiennych. Zgromadzenie danych i konstrukcja macierzy danych w przestrzeni m-wymiarowej $[x_{ij}]_{nxm}$ ($i = 1,\ldots,n$ – numer obiektu, $j = 1,\ldots,m$ – numer zmiennej),
N – normalizacja wartości zmiennych i konstrukcja macierzy $[z_{ij}]_{nxm}$ (z_{ij} – znormalizowana wartość j-tej zmiennej dla i-tego obiektu),
S – przeprowadzenie skalowania wielowymiarowego: wybór miary odległości (zob. tab. 2) i konstrukcja macierzy odległości w przestrzeni m-wymiarowej $[\delta_{ik}]$, $f : \delta_{ik} \rightarrow d_{ik}$ – odwzorowanie macierzy odległości w przestrzeni m-wymiarowej $[\delta_{ik}]$ w macierz odległości w przestrzeni q-wymiarowej $[d_{ik}]$ ($q < m$). Iteracyjny schemat postępowania w algorytmie smacof przedstawiono w pracy (Borg, Groenen, 2005, s. 204–205), prezentacja macierzy danych w przestrzeni q-wymiarowej $[x_{ij}]_{nxq}$,
I – interpretacja wyników skalowania wielowymiarowego (w tym interpretacja osi).

3. NORMALIZACJA WARTOŚCI ZMIENNYCH[4]

Celem normalizacji wartości zmiennych jest doprowadzenie zmiennych do porównywalności poprzez pozbawienie mian wyników pomiaru oraz ujednolicenie ich rzędów wielkości.

Przegląd metod normalizacji wartości zmiennych przedstawia praca Walesiak (2014). Tabela 1 prezentuje metody normalizacyjne dane przekształceniem liniowym (por. Jajuga, Walesiak, 2000, s. 106–107; Zeliaś, 2002, s. 792):

$$z_{ij} = b_j x_{ij} + a_j = \frac{x_{ij} - A_j}{B_j} = \frac{1}{B_j} x_{ij} - \frac{A_j}{B_j} \quad (b_j > 0), \tag{3}$$

gdzie:
x_{ij} – wartość j-tej zmiennej dla i-tego obiektu,
z_{ij} – znormalizowana wartość j-tej zmiennej dla i-tego obiektu,
A_j – parametr przesunięcia do umownego zera dla j-tej zmiennej,
B_j – parametr skali dla j-tej zmiennej,
$a_j = -A_j/B_j$, $b_j = 1/B_j$ – parametry dla j-tej zmiennej określone w tabeli 1.

[4] Punkt ten opracowano na podstawie artykułu Walesiak (2014).

Tabela 1.

Metody normalizacji wartości zmiennych

Typ	Nazwa metody	Parametr		Skale pomiaru zmiennych	
		b_j	a_j	przed normalizacją	po normalizacji
n1	Standaryzacja	$1/s_j$	$-\bar{x}_j/s_j$	ilorazowa lub interwałowa	interwałowa
n2	Standaryzacja pozycyjna	$1/mad_j$	$-med_j/mad_j$	ilorazowa lub interwałowa	interwałowa
n3	Unitaryzacja	$1/r_j$	$-\bar{x}_j/r_j$	ilorazowa lub interwałowa	interwałowa
n3a	Unitaryzacja pozycyjna	$1/r_j$	$-med_j/r_j$	ilorazowa lub interwałowa	interwałowa
n4	Unitaryzacja zerowana	$1/r_j$	$-\min_i\{x_{ij}\}/r_j$	ilorazowa lub interwałowa	interwałowa
n5	Normalizacja w przedziale [–1; 1]	$\frac{1}{\max_i\lvert x_{ij}-\bar{x}_j\rvert}$	$\frac{-\bar{x}_j}{\max_i\lvert x_{ij}-\bar{x}_j\rvert}$	ilorazowa lub interwałowa	interwałowa
n5a	Normalizacja pozycyjna w przedziale [–1; 1]	$\frac{1}{\max_i\lvert x_{ij}-med_j\rvert}$	$\frac{-med_j}{\max_i\lvert x_{ij}-med_j\rvert}$	ilorazowa lub interwałowa	interwałowa
n6	Przekształcenia ilorazowe	$1/s_j$	0	ilorazowa	ilorazowa
n6a		$1/mad_j$	0	ilorazowa	ilorazowa
n7		$1/r_j$	0	ilorazowa	ilorazowa
n8		$1/\max_i\{x_{ij}\}$	0	ilorazowa	ilorazowa
n9		$1/\bar{x}_j$	0	ilorazowa	ilorazowa
n9a		$1/med_j$	0	ilorazowa	ilorazowa
n10		$1/\sum_{i=1}^n x_{ij}$	0	ilorazowa	ilorazowa
n11		$1/\sqrt{\sum_{i=1}^n x_{ij}^2}$	0	ilorazowa	ilorazowa
n12	Normalizacja	$\frac{1}{\sqrt{\sum_{i=1}^n (x_{ij}-\bar{x}_j)^2}}$	$\frac{-\bar{x}_j}{\sqrt{\sum_{i=1}^n (x_{ij}-\bar{x}_j)^2}}$	ilorazowa lub interwałowa	interwałowa
n12a	Normalizacja pozycyjna	$\frac{1}{\sqrt{\sum_{i=1}^n (x_{ij}-med_j)^2}}$	$\frac{-med_j}{\sqrt{\sum_{i=1}^n (x_{ij}-med_j)^2}}$	ilorazowa lub interwałowa	interwałowa
n13	Normalizacja z zerem usytuowanym centralnie	$\frac{1}{r_j/2}$	$-\frac{m_j}{r_j/2}$	ilorazowa lub interwałowa	interwałowa

$\bar{x}_j$ – średnia dla j-tej zmiennej, s_j – odchylenie standardowe dla j-tej zmiennej, r_j – rozstęp dla j-tej zmiennej, $m_j = \frac{\max_i\{x_{ij}\} + \min_i\{x_{ij}\}}{2}$ – środek rozstępu (ang. *mid-range*), $med_j = \underset{i}{med}(x_{ij})$ – mediana dla j-tej zmiennej, $mad_j = \underset{i}{mad}(x_{ij})$ – medianowe odchylenie bezwzględne dla j-tej zmiennej.

Źródło: Walesiak (2014, s. 364–365).

W artykule normalizację wartości zmiennych przeprowadzono w pakiecie `clusterSim` (zob. Walesiak, Dudek, 2015) programu R (R Development Core Team, 2015) z wykorzystaniem funkcji `data.Normalization`.

4. PROCEDURA BADAWCZA POZWALAJĄCA NA WYODRĘBNIENIE GRUP METOD NORMALIZACJI PROWADZĄCYCH DO ZBLIŻONYCH WYNIKÓW SKALOWANIA WIELOWYMIAROWEGO

Procedura badawcza pozwalająca na wyodrębnienie grup metod normalizacji wartości zmiennych prowadzących do zbliżonych wyników skalowania wielowymiarowego obejmuje następujące kroki:

1. Przeprowadza się, zgodnie z ogólnym schematem postępowania w skalowaniu wielowymiarowym, następujące etapy $P \rightarrow A \rightarrow X \rightarrow N$. Do normalizacji wartości zmiennych wykorzystuje się wszystkie dopuszczalne metody ujęte w tabeli 1 (dla zmiennych ilorazowych dostępnych jest 18 metod normalizacyjnych, a dla zmiennych przedziałowych – 10 metod normalizacyjnych).
2. Dla wszystkich macierzy danych po normalizacji wartości zmiennych oblicza się odległości między obiektami (zob. tabela 2) i zestawia w macierze odległości $\left[\delta_{ik}^{r}\right]$ (r – numer metody normalizacyjnej). Dla zmiennych ilorazowych otrzymuje się 18 macierzy odległości, a dla zmiennych przedziałowych – 10 macierzy odległości. Miara odległości Canberra nie zależy od parametru skali B_j (Pawełek, 2008, s. 94). Zatem dopuszczalne metody normalizacyjne n6–n11 nie zmieniają wartości tej odległości (zob. tab. 2).

Tabela 2.

Miary odległości dla danych metrycznych (interwałowych, ilorazowych)

Nazwa	Odległość δ_{ik}	Rozstęp	Dozwolone normalizacje
Minkowski ($p \geq 1$)	$\sqrt[p]{\sum_{j=1}^{m}\left\|z_{ij}-z_{kj}\right\|^{p}}$	$[0;\infty)$	n1–n13
– Manhattan (miejska) ($p = 1$)	$\sum_{j=1}^{m}\left\|z_{ij}-z_{kj}\right\|$	$[0;\infty)$	n1–n13
– Euklidesa ($p = 2$)	$\sqrt{\sum_{j=1}^{m}\left(z_{ij}-z_{kj}\right)^{2}}$	$[0;\infty)$	n1–n13
– Czebyszewa (maximum) ($p \rightarrow \infty$)	$\max_{j}\left\|z_{ij}-z_{kj}\right\|$	$[0;\infty)$	n1–n13
GDM1 (Walesiak, 2002; Jajuga, Walesiak, Bąk, 2003)	$\frac{1}{2}-\frac{\sum_{j=1}^{m}(z_{ij}-z_{kj})(z_{kj}-z_{ij})+\sum_{j=1}^{m}\sum_{\substack{l=1\\ l\neq i,k}}^{n}(z_{ij}-z_{lj})(z_{kj}-z_{lj})}{2\left[\sum_{j=1}^{m}\sum_{l=1}^{n}(z_{ij}-z_{lj})^{2}\cdot\sum_{j=1}^{m}\sum_{l=1}^{n}(z_{kj}-z_{lj})^{2}\right]^{\frac{1}{2}}}$	$[0;1]$	n1–n13

Tabela 2. (cd.)

Nazwa	Odległość δ_{ik}	Rozstęp	Dozwolone normalizacje
Bray-Curtis (Bray, Curtis, 1957)*	$\frac{\sum_{j=1}^{m}\lvert z_{ij}-z_{kj}\rvert}{\sum_{j=1}^{m}(z_{ij}+z_{kj})}$	[0;1]	n6–n11
Canberra (Lance, Williams, 1966)	$\sum_{j=1}^{m}\frac{\lvert z_{ij}-z_{kj}\rvert}{(z_{ij}+z_{kj})}=\sum_{j=1}^{m}\frac{\lvert x_{ij}-x_{kj}\rvert}{(x_{ij}+x_{kj})}$	[0;1]	n6–n11

$i,k,l = 1,\ldots,n$ – numery obiektów, $j = 1,\ldots,m$ – numer zmiennej, m – liczba zmiennych, z_{ij} (z_{kj} , z_{lj}) – znormalizowana wartość j-tej zmiennej dla i-tego (k-tego, l-tego) obiektu.

* Zob. również pracę (Cormack, 1971, s. 367).

Źródło: opracowanie własne.

3. Dla każdej macierzy odległości $\left[\delta_{ik}^{r}\right]$ przeprowadza się skalowanie wielowymiarowe dla ustalonej liczby wymiarów q otrzymując macierz odległości między obiektami w przestrzeni q-wymiarowej $\left[d_{ik}^{r}\right]$.
4. Otrzymane, dla dopuszczalnych metod normalizacyjnych, macierze odległości $\left[d_{ik}^{r}\right]$ porównuje się z wykorzystaniem odległości miejskiej d_{rs}:

$$d_{rs}=\sum_{\substack{i,k=1\\ i<k}}^{\frac{n(n-1)}{2}}\left|\frac{d_{ik}^{r}}{\max\limits_{i,k}\{d_{ik}^{r}\}}-\frac{d_{ik}^{s}}{\max\limits_{i,k}\{d_{ik}^{s}\}}\right|, \tag{4}$$

gdzie: r, s – numery metod normalizacyjnych.

W celu sprowadzenia macierzy odległości $\left[d_{ik}^{r}\right]$ do porównywalności we wzorze (4) podzielono odległości w każdej macierzy odległości przez wartość maksymalną. Po tej operacji odległości w każdej macierzy odległości zawarte będą w przedziale [0;1]. Im mniejszą wartość przyjmuje miara d_{rs} o postaci (4), tym większe jest podobieństwo wyników skalowania wielowymiarowego dla metod normalizacyjnych o numerach r oraz s.

5. Na podstawie macierzy odległości $[d_{rs}]$ przeprowadza się analizę skupień, która pozwala wyodrębnić grupy metod normalizacji wartości zmiennych prowadzących do zbliżonych wyników skalowania wielowymiarowego. Można zastosować tutaj jedną z wielu metod klasyfikacji (zob. np. Everitt i in., 2011; Gordon, 1999). W artykule zastosowano hierarchiczną metodę aglomeracyjną najdalszego sąsiada.
Dla miary odległości GDM1 oraz odległości Braya-Curtisa można sformułować spostrzeżenia odnośnie metod normalizacyjnych ujęte w tabeli 3.

Tabela 3.

Grupy metod normalizacyjnych prowadzących do identycznych odległości w macierzy odległości wyznaczonej za pomocą miary GDM1 oraz odległości Braya-Curtisa

Grupy metod	Metody normalizacyjne	
	odległość GDM1	odległość Braya-Curtisa
A	n1, n6, n12	–
B	n2, n6a	–
C	n3, n3a, n4, n7, n13	–
D	n9, n10	n9, n10

Źródło: opracowanie własne.

Identyczne macierze odległości dla grup metod A, B, C i D wynikają z tego, że miara GDM1 nie zależy od parametru przesunięcia A_j stosowanego w metodach normalizacyjnych. Ponadto przemnożenie wartości znormalizowanych przez stałą nie zmienia odległości GDM1 i Braya-Curtisa (Walesiak, 2015):

– dla metody n13 stała równa się 2:

$$z_{ij} = \frac{x_{ij}}{r_j / 2} - \frac{m_j}{r_j / 2} = 2 \cdot \left(\frac{x_{ij}}{r_j} - \frac{m_j}{r_j} \right), \tag{5}$$

– dla metody n12 stała równa się $\sqrt{\frac{1}{n-1}}$:

$$z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j \right)^2}} - \frac{\bar{x}_j}{\sqrt{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j \right)^2}} = \sqrt{\frac{1}{n-1}} \cdot \left(\frac{x_{ij}}{s_j} - \frac{\bar{x}_j}{s_j} \right), \tag{6}$$

– dla metody n10 stała równa się 1/*n*:

$$z_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} = \frac{1}{n} \cdot \left(\frac{x_{ij}}{\bar{x}_j} \right). \tag{7}$$

W pracy (Pawełek, 2008, s. 94) wykazano, że wartości miar odległości Minkowskiego (miejska, euklidesowa, Czebyszewa) nie zależą od parametru przesunięcia A_j stosowanego w metodach normalizacyjnych. Zatem identyczne macierze odległości otrzymuje się dla grup metod normalizacyjnych ujętych w tabeli 4.

Tabela 4.

Grupy metod normalizacyjnych prowadzących do identycznych macierzy odległości dla odległości Minkowskiego

Grupy metod	Metody normalizacyjne	
	D1	D2
A	n1, n6	n1, n6, n12*
B	n2, n6a	n2, n6a
C	n3, n3a, n4, n7	n3, n3a, n4, n7, n13*
D	–	n9, n10*

D2 – po podzieleniu odległości w każdej macierzy odległości przez wartość maksymalną.

* – dla tej metody normalizacji macierz odległości jest przemnożona przez stałą (zob. wzory (5)–(7)).

Źródło: opracowanie własne.

5. WYNIKI BADANIA EMPIRYCZNEGO

W badaniu empirycznym wykorzystane zostaną dane statystyczne z artykułu (Gryszel, Walesiak, 2014) dotyczące poziomu atrakcyjności turystycznej 29 powiatów Dolnego Śląska. Ocenę poziomu atrakcyjności turystycznej powiatów Dolnego Śląska przeprowadzono z wykorzystaniem 16 zmiennych metrycznych (mierzonych na skali ilorazowej):

x1 – miejsca noclegowe w obiektach na 1 km^2 powierzchni powiatu,
x2 – liczba noclegów turystów rezydentów (Polaków) przypadających dziennie na 1 tys. mieszkańców powiatu,
x3 – liczba noclegów turystów zagranicznych przypadających dziennie na 1 tys. mieszkańców powiatu,
x4 – emisja zanieczyszczeń gazowych w tonach na 1 km^2 powierzchni powiatu,
x5 – liczba przestępstw o charakterze kryminalnym oraz przestępstw przeciwko życiu i zdrowiu na 1 tys. mieszkańców powiatu,
x6 – liczba przestępstw przeciwko mieniu na 1 tys. mieszkańców powiatu,
x7 – liczba obiektów zabytkowych na 100 km^2 powierzchni powiatu,
x8 – lesistość powiatu w %,
x9 – udział obszarów prawnie chronionych w powierzchni powiatu w %,
x10 – liczba imprez oraz wydarzeń kulturalnych i turystycznych w powiecie,
x11 – liczba pomników przyrody w przeliczeniu na 1 km^2 powierzchni powiatu,
x12 – liczba podmiotów gospodarki turystycznej na 1 tys. mieszkańców powiatu (osoby fizyczne i prawne),
x13 – wydatki gmin i powiatów na turystykę, kulturę i ochronę dziedzictwa narodowego oraz kulturę fizyczną na 1 mieszkańca powiatu w zł,
x14 – widzowie w kinach na 1 tys. mieszkańców powiatu,
x15 – zwiedzający muzea na 1 tys. mieszkańców powiatu,

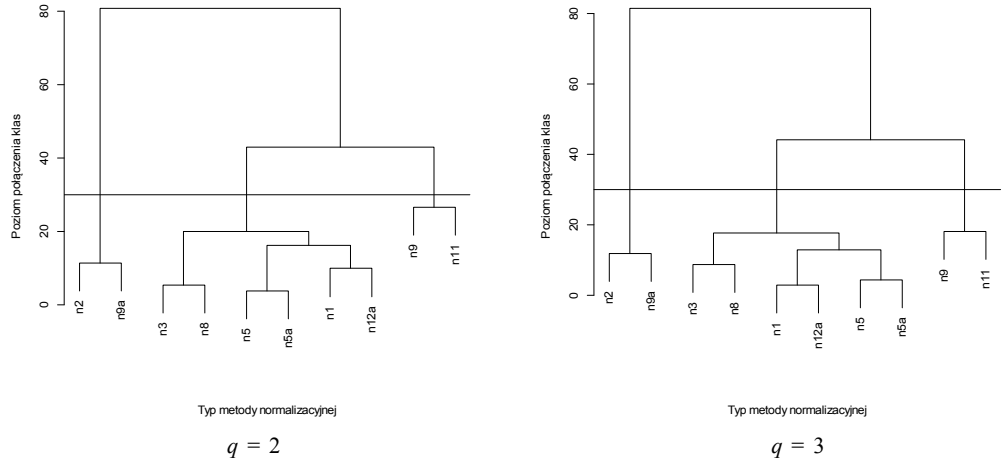
x16 – liczba pozwoleń na budowę (hotele i budynki zakwaterowania, budynki handlowo-usługowe, budynki transportu i łączności, obiekty inżynierii lądowej i wodnej) wydanych w powiecie w latach 2011–2012 na 1 km^2 powierzchni powiatu.

Dane statystyczne pochodzą z roku 2012 z Banku Danych Lokalnych (BDL), jedynie dane dla zmiennej x7 pochodzą od wojewódzkiego konserwatora zabytków.

W artykule zastosowano skrypt programu R przygotowany zgodnie z procedurą badawczą z sekcji 4, która pozwala na wyodrębnienie grup metod normalizacji wartości zmiennych prowadzących do zbliżonych wyników skalowania wielowymiarowego.

Pomiar zmiennych na skali ilorazowej dopuszcza wszystkie metody normalizacyjne (badaniem objęto zatem 18 metod). Z uwagi na to, że grupy metod normalizacyjnych A, B, C i D dają dla skalowania wielowymiarowego identyczne wyniki dalszej analizie poddano pierwsze metody ze wskazanych grup (n1, n2, n3, n9) oraz pozostałe metody (n5, n5a, n8, n9a, n11, n12a).

Skalowanie wielowymiarowe 29 powiatów Dolnego Śląska ze względu na poziom atrakcyjności turystycznej przeprowadzono z wykorzystaniem funkcji `smacofSym` pakietu `smacof` (Mair i in., 2015). Jako miarę odległości zastosowano odległość Euklidesa. Grupy metod normalizacji wartości zmiennych prowadzących do zbliżonych wyników skalowania wielowymiarowego prezentuje dendrogram na rysunku 1.



$q = 2$ $q = 3$

Rysunek 1. Dendrogram podobieństwa metod normalizacji w skalowaniu wielowymiarowym 29 powiatów Dolnego Śląska ze względu na poziom atrakcyjności turystycznej
Źródło: opracowanie własne z wykorzystaniem programu R.

Na podstawie dendrogramu wyróżniono trzy grupy metod normalizacyjnych (zarówno w dwóch, jak w trzech wymiarach) prowadzących do zbliżonych wyników skalowania wielowymiarowego w sensie macierzy odległości $\left[d_{ik}^{r}\right]$ oraz rozmieszczenia

obiektów w przestrzeni q-wymiarowej (w nawiasach przedstawiono metody normalizacyjne dające identyczne wyniki skalowania wielowymiarowego):
grupa 1 (3 metody): (n2, n6a), n9a,
grupa 2 (12 metod): (n1, n6, n12), (n3, n3a, n4, n7, n13), n5, n5a, n8, n12a,
grupa 3 (3 metody): (n9, n10), n11.

Do wyboru liczby klas można wykorzystać tutaj indeksy oceny jakości klasyfikacji przedstawione w pakietach `NbClust` (Charrad i in., 2014; Charrad i in., 2015) oraz `clusterSim` (Walesiak, Dudek, 2015).

W analizowanym przypadku istotne różnice między wynikami skalowania wielowymiarowego pojawiają się dla metod normalizacji wartości zmiennych z różnych grup. W dotychczasowej praktyce, nie uwzględniając zaproponowanej procedury badawczej, dokonując wyboru metody normalizacji wartości zmiennych w skalowaniu wielowymiarowym dla danych metrycznych mieliśmy do wyboru 18 propozycji (tabela 1). Rozważania ujęte w tab. 3 i 4 zmniejszają tę liczbę do 10 metod normalizacji. Wybór nadal staje się arbitralny i trudny do uzasadnienia. Zaproponowane podejście nie rozwiązuje całkowicie problemu, ale pozwala wyodrębnić grupy metod normalizacji prowadzące do zbliżonych wyników skalowania wielowymiarowego. W analizowanym przykładzie mamy już do wyboru *de facto* 3 metody normalizacji (metody normalizacji znajdujące się w tych samych grupach dają identyczne lub zbliżone wyniki skalowania wielowymiarowego). Zatem przedstawiona propozycja pozwala ograniczyć problem wyboru metody normalizacyjnej.

6. PODSUMOWANIE

W artykule przedstawiono propozycję procedury badawczej pozwalającą na wyodrębnienie grup metod normalizacji wartości zmiennych prowadzących do zbliżonych wyników skalowania wielowymiarowego. Propozycja pozwala ograniczyć problem wyboru formuły normalizacji wartości zmiennych w skalowaniu wielowymiarowym. Istotne różnice między wynikami skalowania wielowymiarowego pojawiają się dla formuł normalizacyjnych z różnych grup.

Wskazano dla miar odległości GDM1 oraz Braya-Curtisa metody normalizacyjne dające identyczne odległości w macierzy odległości. Analogiczne spostrzeżenia sformułowano dla miar odległości Minkowskiego (miejska, euklidesowa, Czebyszewa).

Wyniki badawcze zobrazowano przykładem empirycznym dotyczącym zastosowania funkcji `smacofSym` pakietu `smacof` w celu przeprowadzenia skalowania wielowymiarowego 29 powiatów Dolnego Śląska ze względu na poziom atrakcyjności turystycznej z wykorzystaniem 18 formuł normalizacyjnych.

LITERATURA

Borg I., Groenen P. J. F., (2005), *Modern Multidimensional Scaling. Theory and Applications*, 2nd Edition, Springer Science+Business Media, New York.

Bray J. R., Curtis J. T., (1957), An Ordination of the Upland Forest Communities of Southern Wisconsin, *Ecological Monographs,* 27 (4), 325–349.

Charrad M., Ghazzali N., Boiteau V., Niknafs A., (2014), NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set, *Journal of Statistical Software*, 61 (6), 2–36.

Charrad M., Ghazzali N., Boiteau V., Niknafs A., (2015), NbClust Package for Determining the Best Number of Clusters. R Package Version 3.0, URL http://CRAN.R-project.org/package=NbClust.

Cormack R. R., (1971), A Review of Classification, *Journal of the Royal Statistical Society*, Series A, 134 (3), 321–367.

Everitt B. S., Landau S., Leese M., Stahl D., (2011), *Cluster Analysis*, John Wiley & Sons, Chichester.

Gordon A. D., (1999), *Classification*, 2nd Edition, Chapman & Hall/CRC, London.

Grabiński T., (1984), *Wielowymiarowa analiza porównawcza w badaniach dynamiki zjawisk ekonomicznych*, Zeszyty Naukowe Akademii Ekonomicznej w Krakowie, Seria specjalna: Monografie nr 61.

Gryszel P., Walesiak M., (2014), Zastosowanie uogólnionej miary odległości GDM w ocenie atrakcyjności turystycznej powiatów Dolnego Śląska, *Folia Turistica*, 31, 127–147.

Jajuga K., Walesiak M., (2000), Standardisation of Data Set under Different Measurement Scales, w: Decker R., Gaul W., (red.), *Classification and Information Processing at the Turn of the Millennium*, 105–112. Springer-Verlag, Berlin, Heidelberg.

Jajuga K., Walesiak M., Bąk A., (2003), On the General Distance Measure, w: Schwaiger M., Opitz O., (red.), *Exploratory Data Analysis in Empirical Research*, 104-109, Springer-Verlag, Berlin, Heidelberg.

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

Lance G. N., Williams W. T., (1966), Computer Programs for Hierarchical Polythetic Classification ("Similarity Analyses"), *The Computer Journal*, 9 (1), 60–64.

Mair P., De Leeuw J., Borg I., Groenen P. J. F., (2015), smacof: Multidimensional Scaling. R Package Version 1.7-0, URL http://CRAN.R-project.org/package=smacof.

Milligan G. W., Cooper M. C., (1988), A Study of Standardization of Variables in Cluster Analysis, *Journal of Classification*, 5 (2), 181–204.

Pawełek B., (2008), *Metody normalizacji zmiennych w badaniach porównawczych złożonych zjawisk ekonomicznych*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.

R Development Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, URL http://www.R-project.org.

Schaffer C. M., Green P. E., (1996), An Empirical Comparison of Variable Standardization Methods in Cluster Analysis, *Multivariate Behavioral Research*, 31 (2), 149–167.

Stevens S. S., (1946), On the Theory of Scales of Measurement, *Science*, 103 (2684), 677–680.

Walesiak M., (2002), *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław.

Walesiak M., (2011), *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R*, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.

Walesiak M., (2014), Przegląd formuł normalizacji wartości zmiennych oraz ich własności w statystycznej analizie wielowymiarowej, *Przegląd Statystyczny*, 61 (4), 363–372.

Walesiak M., (2015), The Results of Linear Ordering of the Set of Objects via Synthetic Measures and the Choice of Normalization Formula, *Statistics in Transition – new series*, w recenzji.

Walesiak M., Dudek A., (2015), clusterSim: Searching for Optimal Clustering Procedure for a Data Set. R package version 0.44-2, URL http://CRAN.R-project.org/package=clusterSim.

Walesiak M., Dudek A., (2016), The Choice of Variable Normalization Method in Cluster Analysis with clusterSim Package and R Environment, w przygotowaniu.
Zaborski A., (2001), *Skalowanie wielowymiarowe w badaniach marketingowych*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław.
Zeliaś A., (2002), Some Notes on the Selection of Normalisation of Diagnostic Variables, *Statistics in Transition*, 5 (5), 787–802.

WYBÓR GRUP METOD NORMALIZACJI WARTOŚCI ZMIENNYCH W SKALOWANIU WIELOWYMIAROWYM

Streszczenie

W skalowaniu wielowymiarowym przeprowadzanym na podstawie macierzy danych metrycznych (przedziałowych, ilorazowych) jednym z etapów jest wybór metody normalizacji wartości zmiennych. W badaniu zastosowano funkcję `data.Normalization` pakietu `clusterSim` programu R. Funkcja ta zawiera 18 różnych metod normalizacyjnych.

W artykule zaproponowano procedurę badawczą pozwalającą na wyodrębnienie grup metod normalizacji wartości zmiennych prowadzących do zbliżonych wyników skalowania wielowymiarowego. Propozycja pozwala ograniczyć problem wyboru metody normalizacji wartości zmiennych w skalowaniu wielowymiarowym. Wyniki zilustrowano przykładem empirycznym.

Słowa kluczowe: normalizacja zmiennych, skalowanie wielowymiarowe, miary odległości, program R, pakiet `clusterSim`

THE CHOICE OF GROUPS OF VARIABLE NORMALIZATION METHODS IN MULTIDIMENSIONAL SCALING

Abstract

In multidimensional scaling carried out on the basis of metric data matrix (interval, ratio) one of the stages is the choice of the variable normalization method. The R package `clusterSim` with `data.Normalization` function has been developed for that purpose. It provides 18 data normalization methods.

In this paper the proposal of procedure which allows to isolate groups of normalization methods that lead to similar multidimensional scaling results were presented. The proposal can reduce the problem of choosing the normalization method in multidimensional scaling. The results are illustrated via empirical example.

Keywords: normalization of variables, multidimensional scaling, distance measures, R program, `clusterSim` package