PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1010 — 2004
Ekonometria 13

Marek Walesiak

PROBLEMY DECYZYJNE W PROCESIE KLASYFIKACJI ZBIORU OBIEKTÓW

1. Wstęp

Klasyfikowaniem zajmowano się od zarania dziejów (por. [33]). W starożytności Hindusi dzielili ludzi na 6 klas (oznaczając je nazwami zwierząt) ze względu na płeć, warunki fizyczne i psychiczne. Arystoteles wprowadził nowe klasyfikacje w logice, etyce i polityce. Inne bardziej znane tego typu klasyfikacje to klasyfikacja zwierząt i roślin opracowana przez Linnaeusa w XVIII w. czy klasyfikacja pierwiastków chemicznych opracowana przez Mendelejewa w XIX w.

Na większą skalę zagadnienie klasyfikacji wykorzystano praktycznie, gdy zaczęto stosować formalne procedury klasyfikacji. Pionierem w dziedzinie stosowania metod klasyfikacji był Czekanowski [16], który zastosował własną oryginalną metodę do klasyfikacji 13 czaszek ludzkich. Metoda ta jest zwana diagraficzną metodą Czekanowskiego. W 1951 r. powstała metoda taksonomii wrocławskiej [19]. Podstawowe algorytmy metod klasyfikacji powstały w latach pięćdziesiątych i sześćdziesiątych XX w.

Według najogólniejszej koncepcji klasyfikacja jest zbiorem klas odpowiednio wyróżnionym z klasyfikowanego zbioru obiektów. Oczywiście takie określenie klasyfikacji jest zbyt ogólnikowe. Zawężone sformułowanie zagadnienia klasyfikacji zbioru A o elementach A_i $(i = 1, 2, \ldots, n)$ na klasy P_1, P_2, ..., P_u spełnia warunki:

- zupełności: $\bigcup_{s=1}^{u} P_s = A$;
- rozłączności: $P_s \cap P_{s'} = \varnothing$ $(s, s' = 1, 2, \ldots, u;\ s \neq s')$;
- niepustości: $P_s \neq \varnothing$ $(s = 1, 2, \ldots, u)$.

ISSN 0324-8445
ISSN 1507-3866

Klasyfikacja spełniająca te trzy podstawowe warunki jest najbardziej użyteczna, a w związku z tym najczęściej wykorzystywana w badaniach marketingowych. Nie będą rozważane więc takie metody klasyfikacji, które dają klasy rozmyte lub nierozłączne.

Głównym celem klasyfikacji jest badanie podobieństwa lub odrębności obiektów i ich zbiorów. Celem tym jest więc podział zbioru obiektów na klasy zawierające obiekty podobne ze względu na wartości zmiennych.

W artykule scharakteryzowano problemy decyzyjne wymagające rozstrzygnięcia w procesie klasyfikacji zbioru obiektów. Wyodrębniono osiem etapów procesu klasyfikacji (por. [49, s. 342-343]):

1. Wybór obiektów do klasyfikacji.
2. Wybór zmiennych charakteryzujących poszczególne obiekty.
3. Wybór formuły normalizacji wartości zmiennych.
4. Wybór miary odległości.
5. Wybór metody klasyfikacji.
6. Ustalenie liczby klas.
7. Walidacja wyników klasyfikacji.
8. Opis (interpretacja) i profilowanie klas.

Na podstawie światowej literatury przedmiotu zaprezentowano podejścia służące rozstrzygnięciu pojawiających się problemów decyzyjnych w procesie klasyfikacji zbioru obiektów.

2. Wybór obiektów do klasyfikacji

Należy odpowiedzieć na pytanie, czy badaniem objąć całą populację, czy tylko jej próbkę? Jeśli zdecydowano się na badanie próbkowe (z takimi badaniami mamy zazwyczaj do czynienia w analizach marketingowych), to należy określić elementarną jednostkę badania, wybrać metodę doboru próby i określić jej liczebność.

W każdym badaniu statystycznym, w tym również w niewyczerpującym badaniu wielowymiarowym, można przyjąć jedno z dwóch podejść. Należą do nich podejścia stochastyczne i opisowe. W podejściu stochastycznym zakłada się, że zbiór obserwacji (obiektów) stanowi próbę losową pochodzącą z populacji (populacja może być zbiorem nieskończonym lub skończonym, z reguły o dużej liczebności). W podejściu stochastycznym rozpatrywane zmienne są losowe. Podejście stochastyczne wolno przyjąć przede wszystkim w przypadku badań eksperymentalnych, tzn. gdy istnieje możliwość powtórzenia badania w takich samych warunkach. Wtedy zbiór obserwacji może być traktowany jako próba losowa. W podejściu opisowym zmienne nie są losowe, lecz są zmiennymi w zwykłym sensie. Badaniu nie podlegają wtedy właściwości stochastyczne zbioru obserwacji. Podejście opisowe przyjmuje się z reguły wtedy, gdy dane pochodzą ze sprawozdawczości statystycznej.

Dobór próby powinno się przeprowadzić tak, aby klasy wyodrębnione na jej podstawie odpowiadały strukturze klas populacji.

3. Wybór zmiennych charakteryzujących poszczególne obiekty

Dobór zmiennych w statystycznej analizie wielowymiarowej jest jednym z najważniejszych, a zarazem najtrudniejszych zagadnień. Od jakości zestawu zmiennych bowiem zależy wiarygodność ostatecznych wyników i trafność podejmowanych na ich podstawie decyzji w zagadnieniach marketingowych. Do rozwiązania zagadnienia doboru zmiennych służą zasadniczo dwa ujęcia: dobór merytoryczny w ścisłym tego słowa znaczeniu, dobór merytoryczno-formalny.

Punktem wyjścia obu ujęć jest skonstruowanie, na podstawie merytorycznej znajomości zagadnienia, potencjalnej (wstępnej) listy zmiennych. Zadanie to jest szczególnie odpowiedzialne. Badaczowi nie wolno z jednej strony opuścić żadnej zmiennej mającej istotne znaczenie przy wyjaśnianiu przedmiotu badania, a z drugiej strony z listy należy usunąć zmienne, które słabo, pośrednio lub pozornie wyjaśniają ten przedmiot. Mniejszym złem – jak z tego wynika – jest wprowadzenie zmiennych nieistotnych w początkowej fazie niż opuszczenie zmiennych istotnych. Zmienne nieistotne mogą być w dalszej fazie usunięte, natomiast pominięcie tych drugich wypacza w dużym stopniu wyniki badań (por. [7, s. 31]).

Przy konstrukcji potencjalnej listy zmiennych trzeba mieć na względzie następujące wymagania: postulat ekonomiczności badań (koszt uzyskania informacji o zmiennych), dostępność danych statystycznych, wiarygodność danych statystycznych. Tak ustalona wstępna lista zmiennych jest – jak pisze Cieślak [12, s. 110] – „wypadkową znajomości przedmiotu badania oraz tradycji statystycznej (zbiera się dane statystyczne dotyczące dobrze już znanych zjawisk)".

Merytoryczny dobór zmiennych jest działaniem w głównej mierze subiektywnym. Redukcji potencjalnej listy zmiennych dokonuje się na podstawie własnej znajomości przedmiotu badania, wykorzystując współpracę przedstawicieli odpowiednich dyscyplin naukowych (ekspertów) oraz opierając się na szeroko pojętej teorii ekonomii.

Podejście merytoryczno-formalne polega na tym, że ze wstępnej listy zmiennych (wybranych na podstawie analizy merytorycznej) usuwa się najpierw zmienne, które charakteryzują się małą zawartością informacyjną (tradycyjnie mierzy się ją zmiennością). Następnie do tak zredukowanej liczby zmiennych stosuje się formalny algorytm wyboru zmiennych. W zagadnieniu klasyfikacji zbioru obiektów celem zastosowania tych algorytmów jest wybór takiego zestawu zmiennych, w którym zmienne są wzajemnie niezależne oraz są zależne od zmiennych nie wchodzących do wybranego zestawu (postulat niepowielania informacji).

W podsumowaniu zagadnienia traktującego o doborze zmiennych należy jednoznacznie stwierdzić, że dobór merytoryczny w ścisłym tego słowa znaczeniu

powinien stanowić punkt wyjścia wszystkich badań. Zwracają na to uwagę m.in. Aldenderfer i Blashfield [3, s. 20]. W sytuacjach, gdy na podstawie posiadanej wiedzy merytorycznej o badanym zjawisku nie jesteśmy w stanie wybrać zmiennych lub jest ich zbyt dużo, możemy skorzystać z metod formalnych. Przy doborze zmiennych nie można polegać tylko na wskazaniach metod.

4. Wybór formuły normalizacji wartości zmiennych

Celem normalizacji zmiennych jest pozbawienie mian wyników pomiaru oraz ujednolicenie ich rzędów wielkości. Normalizację przeprowadza się, gdy zmienne opisujące obiekty badania mierzone są na skali przedziałowej i (lub) ilorazowej. Z uwagi na to, że jedynymi dopuszczalnymi przekształceniami na skali przedziałowej i ilorazowej są przekształcenia liniowe, formuły normalizacyjne można wyrazić ogólnym wzorem:

$$z_{ij} = bx_{ij} + a,\ b > 0, \tag{1}$$

gdzie: x_{ij} (z_{ij}) – wartość (znormalizowana wartość) j-tej zmiennej zaobserwowana w i-tym obiekcie.

Szczególnymi przypadkami tego wzoru są następujące formuły (por. np. [1; 9, s. 297-308; 29, s. 35-38; 37; 41; 50; 52, s. 38-39; 69, s. 19]):

A. Standaryzacja:

- klasyczna: $z_{ij} = s_j^{-1}x_{ij} - \bar{x}_j s_j^{-1}$,
- Webera[1]: $z_{ij} = \left(x_{ij} - Me_j\right)/1{,}4826 \cdot MAD_j$.

B. Unitaryzacja: $z_{ij} = r_j^{-1}x_{ij} - \bar{x}_j r_j^{-1}$.

C. Unitaryzacja zerowana: $z_{ij} = \left[x_{ij} - \min_i\{x_{ij}\}\right]/r_j$.

D. Normalizacja[2] w przedziale [–1; 1]: $z_{ij} = \left(x_{ij} - \bar{x}_j\right)/\max_i\left|x_{ij} - \bar{x}_j\right|$.

E. Przekształcenia ilorazowe: $z_{ij} = x_{0j}^{-1}x_{ij}$,

gdzie: $\bar{x}_j$, s_j, r_j – odpowiednio: średnia arytmetyczna, odchylenie standardowe i rozstęp wyznaczony na podstawie wartości j-tej zmiennej,

x_{0j} – podstawa normalizacji j-tej zmiennej (np. za x_{0j} przyjmuje się: s_j, r_j, $\max_i\{x_{ij}\}$, $\bar{x}_j \sum_{i=1}^{n} x_{ij}$, $\sqrt{\sum_{i=1}^{n} x_{ij}^2}$),

[1] Zob. [47, s. 91].

[2] Zob. [58, s. 147].

Me_j – mediana dla j-tej zmiennej,

MAD_j – medianowe odchylenie bezwzględne.

Ujednolicenie rzędów wielkości jest możliwe tylko w razie jednolitego określenia wartości zerowej dla wszystkich zmiennych (zob. [66]). Przekształcenia ilorazowe można stosować tylko wtedy, gdy zmienne są mierzone na skali ilorazowej (istnieje dla niej absolutny punkt zerowy). Gdy zbiór zawiera zmienne mierzone na skali przedziałowej lub przedziałowej i ilorazowej, wówczas do normalizacji można stosować pozostałe formuły normalizacyjne, wprowadzające jednolicie określoną wartość zerową (umowną) dla wszystkich zmiennych. Standaryzacja klasyczna (standaryzacja Webera), unitaryzacja, normalizacja w przedziale [–1; 1] określają umowną wartość zerową na poziomie średniej wartości zmiennej (mediany), a unitaryzacja zerowana – na poziomie wartości minimalnej. Zastosowanie tych formuł normalizacyjnych do zmiennych mierzonych na skali ilorazowej, aczkolwiek formalnie poprawne, spowoduje stratę informacji wskutek „przejścia" wszystkich zmiennych na skalę przedziałową. Strata informacji przejawia się m.in. ograniczeniem zastosowania różnych technik statystycznych i ekonometrycznych.

Przy wyborze formuły normalizacyjnej należy brać pod uwagę nie tylko skale pomiaru zmiennych, ale również takie charakterystyki rozkładu zmiennych, jak (por. tab. 1): średnia arytmetyczna, odchylenie standardowe i rozstęp wyznaczony dla znormalizowanych wartości zmiennych.

Analiza tab. 1 pozwala sformułować następujące wnioski (zob. [41, s. 110-111; 69, s. 20]):

a) formuły normalizacyjne (unitaryzacja, unitaryzacja zerowana, przekształcenie ilorazowe z podstawą normalizacji równą rozstępowi) są cenne, ponieważ zapewniają znormalizowanym wartościom zmiennych zróżnicowaną zmienność (mierzoną odchyleniem standardowym) i jednocześnie stały rozstęp dla wszystkich zmiennych;

b) standaryzacja klasyczna (Webera) oraz przekształcenie ilorazowe z podstawą normalizacji równą odchyleniu standardowemu powodują ujednolicenie wartości wszystkich zmiennych pod względem zmienności mierzonej odchyleniem standardowym (medianowym odchyleniem bezwzględnym). Oznacza to wyeliminowanie zmienności jako podstawy różnicowania obiektów. Standaryzację Webera należy stosować, gdy rozkład empiryczny badanych zmiennych jest silnie asymetryczny (zob. [47, s. 91]);

c) przekształcenia ilorazowe z podstawą normalizacji równą maksimum oraz pierwiastkowi z sumy kwadratów obserwacji zapewniają znormalizowanym wartościom zmiennych zróżnicowaną zmienność, średnią arytmetyczną i rozstęp;

d) przekształcenia ilorazowe z podstawą normalizacji równą sumie i średniej arytmetycznej oraz normalizacja w przedziale [–1; 1] zapewniają znormalizowanym wartościom zmiennych zróżnicowaną zmienność i rozstęp oraz stałą dla

Tabela 1. Charakterystyki rozkładu wartości zmiennych po normalizacji

Formuła	Średnia arytmetyczna*	Odchylenie standardowe*	Rozstęp
$(x_{ij} - \bar{x}_j)/s_j$	0	1	r_j/s_j
$z_{ij} = (x_{ij} - Me_j)/1{,}4826 \cdot MAD_j$	0	1	$r_j/1{,}4826 \cdot MAD_j$
$(x_{ij} - \bar{x}_j)/r_j$	0	s_j/r_j	1
$\left[x_{ij} - \min_i\{x_{ij}\}\right]/r_j$	$\left[\bar{x}_j - \min_i\{x_{ij}\}\right]/r_j$	s_j/r_j	1
$z_{ij} = (x_{ij} - \bar{x}_j)/\max_i\lvert x_{ij} - \bar{x}_j\rvert$	0	$s_j/\max_i\lvert x_{ij} - \bar{x}_j\rvert$	$r_j/\max_i\lvert x_{ij} - \bar{x}_j\rvert$
x_{ij}/s_j	$\bar{x}_j/s_j$	1	r_j/s_j
x_{ij}/r_j	$\bar{x}_j/r_j$	s_j/r_j	1
$x_{ij}/\max_i\{x_{ij}\}$	$\bar{x}_j/\max_i\{x_{ij}\}$	$s_j/\max_i\{x_{ij}\}$	$r_j/\max_i\{x_{ij}\}$
$x_{ij}/\bar{x}_j$	1	$s_j/\bar{x}_j$	$r_j/\bar{x}_j$
$x_{ij}/\sum_{i=1}^{n} x_{ij}$	$1/n$	$s_j/\sum_{i=1}^{n} x_{ij}$	$r_j/\sum_{i=1}^{n} x_{ij}$
$x_{ij}/\sqrt{\sum_{i=1}^{n} x_{ij}^2}$	$\bar{x}_j/\sqrt{\sum_{i=1}^{n} x_{ij}^2}$	$s_j/\sqrt{\sum_{i=1}^{n} x_{ij}^2}$	$r_j/\sqrt{\sum_{i=1}^{n} x_{ij}^2}$

$\bar{x}_j$, s_j, r_j – średnia arytmetyczna, odchylenie standardowe, rozstęp dla j-tej zmiennej,

* dla standaryzacji Webera: mediana i medianowe odchylenie bezwzględne.

Źródło: opracowanie własne na podstawie [41, s. 109; 47, s. 91].

wszystkich zmiennych średnią arytmetyczną. Pierwsza formuła stanowi podstawę normalizacji w badaniach strukturalnych;

e) wszystkie formuły normalizacyjne, będące przekształceniami liniowymi obserwacji na każdej zmiennej, zachowują skośność i kurtozę rozkładu zmiennych. Ponadto dla każdej pary zmiennych wszystkie formuły normalizacyjne nie zmieniają wartości współczynnika korelacji liniowej Pearsona.

5. Wybór miary odległości

Stosowanie konkretnych konstrukcji miar odległości jest uzależnione od skal pomiaru zmiennych. W literaturze wypracowano wiele propozycji miar odległości znajdujących zastosowanie do zmiennych mierzonych na skali: ilorazowej, przedziałowej i (lub) ilorazowej, porządkowej, nominalnej (w tym dla zmiennych binarnych).

Bardzo dobry przegląd różnych typów miar odległości przedstawiono m.in. w pracach: [4, s. 98-130; 13; 14, s. 10; 25, s. 20-21; 43, s. 4-37; 69, s. 23-31].

Przy wyborze miar odległości obiektów opisanych zmiennymi mierzonymi na skali przedziałowej i (lub) ilorazowej należy wziąć pod uwagę również zastosowaną formułę normalizacji wartości zmiennych (zob. [69, s. 29]).

Na wybór konkretnej formuły odległości wpływa też spełnianie przez daną formułę dodatkowych własności. Każda funkcja będzie nazywana miarą odległości wtedy i tylko wtedy, gdy spełnia warunki: nieujemności, zwrotności i symetryczności. Dodatkową cenną zaletą miar odległości jest spełnianie warunku nierówności trójkąta (miara odległości zwana jest wtedy metryką). Spośród miar odległości obiektów opisanych zmiennymi mierzonymi na skali przedziałowej i (lub) ilorazowej najczęściej wykorzystuje się z tego powodu odległość euklidesową i jej kwadrat.

Sytuacja komplikuje się wtedy, gdy w zbiorze znajdują się zmienne mierzone na skalach różnych rodzajów. Na podstawie literatury przedmiotu (por. [22, s. 25-27; 39; 43, s. 32-37; 44; 67]) do rozwiązania tego problemu można wykorzystać następujące sposoby:

1. Przeprowadzić klasyfikację zbioru obiektów osobno dla każdej grupy zmiennych. Gdy tak otrzymane rezultaty są w miarę zgodne, problem można uznać za rozwiązany. Sytuacja komplikuje się wtedy, gdy wyniki te znacznie odbiegają od siebie.

2. Wykorzystać w analizie tylko zmienne jednego ustalonego typu (dominującego w zbiorze zmiennych) z odrzuceniem zmiennych innego typu. Wyniki uzyskane na podstawie zbioru zmiennych uzyskanego w taki sposób są na ogół bardzo zniekształcone (wskutek tego, że musimy zrezygnować z części informacji, które niosą odrzucone zmienne).

3. W praktyce zaniedbać to, że zmienne są mierzone na skalach różnych typów i stosować metody właściwe dla zmiennych jednego typu. Zmienne nominalne i porządkowe traktuje się zazwyczaj tak jak przedziałowe i ilorazowe: stosuje się więc do nich techniki właściwe tym skalom. Sposób ten, choć atrakcyjny z aplikacyjnego punktu widzenia, jest nie do przyjęcia ze względów metodologicznych (następuje tu bowiem sztuczne wzmocnienie skali pomiaru).

4. Dokonać transformacji zmiennych tak, by sprowadzić je do skali jednego typu. Podstawowa reguła teorii pomiaru mówi, że jedynie rezultaty pomiaru w skali mocniejszej mogą być transformowane na liczby należące do skali słabszej. Wynika z tego, że wszystkie obserwacje na zmiennych należy przekodować na pomiary na skali najsłabszej. Tej operacji towarzyszy jednak utrata informacji. Proponowane są również w tym względzie procedury wzmacniania skal pomiaru (por. [4, s. 53-69; 54]). Są to aproksymacyjne metody przekształcania skal słabszych w silniejsze, opierające się na pewnych dodatkowych informacjach. Z punktu widzenia teorii pomiaru wzmacnianie skal jest jednak niemożliwe, ponieważ z mniejszej ilości informacji nie można uzyskać większej ilości informacji.

5. Posłużyć się miarami podobieństwa dopuszczającymi wykorzystanie zmiennych mierzonych na różnych skalach. W literaturze miary takie zaproponowali: [8, s. 152; 15; 26; 70].

6. Wybór metody klasyfikacji

Do rozwiązania problemu wyboru właściwej dla danego typu danych empirycznych metody klasyfikacji proponuje się w literaturze przedmiotu cztery zasadnicze podejścia (por. [23]).

W pierwszym z nich poprawność poszczególnych metod ocenia się na podstawie zadanych typów struktur danych. Dana metoda klasyfikacji jest poprawna, jeśli wyniki klasyfikacji uzyskane za jej pomocą odpowiadają znanej strukturze danych. Przykłady zastosowania tego typu podejścia można znaleźć m.in. w pracach [27; 28; 29; 48; 49].

Podstawową wadą tego podejścia jest to, że opiera się na wygenerowanych strukturach danych, w których konfiguracje obiektów są na ogół przedstawiane w przestrzeniach dwuwymiarowych i trójwymiarowych. Trudno jest więc uogólnić wyniki na przypadek wielowymiarowy. Nawet wtedy, gdy podejście to opiera się na danych symulacyjnych (uzyskanych za pomocą odpowiednio skonstruowanych wielowymiarowych generatorów zmiennych losowych o zadanej postaci rozkładu), trudno jest uogólnić wyniki, ponieważ każda empirycznie uzyskana struktura danych jest inna i tak uzyskane wnioski mają ograniczony zasięg zastosowania.

Na podstawie wielu analiz poprawności odkrywania zadanych typów struktur danych Milligan [49, s. 358] wskazał, że najlepsze wśród hierarchicznych metod aglomeracyjnych są metody Warda i giętka (*β-flexible*).

Podejście drugie polega na tym, że do klasyfikacji zbioru obiektów wykorzystuje się różne metody klasyfikacji, a następnie ocenia się zgodność wyników klasyfikacji i wybiera się te metody, które dają zbliżone wyniki. Wyniki klasyfikacji z użyciem tych metod podlegają w dalszej fazie syntetyzacji w celu wyłonienia zgodnej klasyfikacji.

Godne odnotowania propozycje mierników służących do porównywania wyników dwóch różnych podziałów podali: Rand [55]; Hubert i Arabie [35]; Fowlkes i Mallows [20]; Lerman [46]; Goodman i Kruskal [21]; Wallace [71]. W literaturze polskiej propozycje takie przedstawili: Nowak [51], Szmigiel [64] i Sokołowski [62].

Dany jest niepusty zbiór obiektów badania A o elementach A_i $(i = 1, 2, \ldots, n)$ oraz dwie klasyfikacje (dwa podziały) tego zbioru na u i v klas otrzymane na podstawie jednolitej procedury klasyfikacyjnej. W celu oceny podobieństwa wyników dwóch podziałów zbioru obiektów konstruuje się tablicę kontyngencji i na tej podstawie otrzymuje się ich klasyfikację krzyżową (zob. tab. 2).

Tabela 2. Tablica kontyngencji (klasyfikacja krzyżowa wyników dwóch podziałów)

Podziały	Klasy	Podział $P^{(t)}$				Sumy
		$P_1^{(t)}$	$P_2^{(t)}$	...	$P_v^{(t)}$	
Podział $P^{(q)}$	$P_1^{(q)}$	n_{11}	n_{12}	...	n_{1v}	$n_{1\cdot}$
	$P_2^{(q)}$	n_{21}	n_{22}	...	n_{2v}	$n_{2\cdot}$
	⋮	⋮	⋮		⋮	⋮
	$P_u^{(q)}$	n_{u1}	n_{u2}	...	n_{uv}	$n_{u\cdot}$
Sumy		$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot v}$	$n_{\cdot\cdot} = n$

gdzie: $P^{(t)}$, $P^{(q)}$ – klasyfikacja (podział zbioru obiektów A) t, q;
n_{sr} – liczba obiektów, które jednocześnie należą do klas $P_s^{(q)}$ i $P_r^{(t)}$;
$r = 1, 2, \ldots, v$; $s = 1, 2, \ldots, u$; v (u) – liczba klas w podziale $P^{(t)}$ ($P^{(q)}$);
$n_{\cdot r}$ – liczba obiektów w klasie $P_r^{(t)}$ (kolumna t);
$n_{s\cdot}$ – liczba obiektów w klasie $P_s^{(q)}$ (wiersz s).

Źródło: opracowanie własne.

W literaturze do oceny podobieństwa wyników klasyfikacji zbioru obiektów najczęściej wykorzystywana jest miara Randa [55]. W jej koncepcji porównuje się zaklasyfikowanie wszystkich par obiektów w podziałach $P^{(t)}$, $P^{(q)}$ i wyróżnia się cztery typy par obiektów (zob. tab. 3):

typ (I): obiekty tworzące parę znajdują się w tych samych klasach w podziałach $P^{(t)}$ i $P^{(q)}$;

typ (II): obiekty tworzące parę znajdują się w różnych klasach w podziałach $P^{(t)}$ i $P^{(q)}$;

typ (III): obiekty tworzące parę znajdują się w różnych klasach w $P^{(q)}$ i w tej samej klasie w $P^{(t)}$;

typ (IV): obiekty tworzące parę znajdują się w tej samej klasie w $P^{(q)}$ i w różnych klasach w $P^{(t)}$.

Typy (I) i (II) są interpretowane jako pary zgodne w obu klasyfikacjach $P^{(t)}$ i $P^{(q)}$, natomiast typy (III) i (IV) – jako pary niezgodne. W tab. 3 przedstawiono formuły pozwalające ustalić liczby par obiektów do każdego typu, będące funkcjami n, $n_{s\cdot}$, $n_{\cdot r}$, i n_{sr}.

Można zauważyć, że podobieństwo dwóch podziałów $P^{(t)}$ i $P^{(q)}$ wzrasta w miarę wzrostu wartości Z. Na tej podstawie Rand [55] skonstruował miarę pozwalającą oceniać podobieństwo wyników dwóch podziałów zbioru obiektów:

$$R = Z \Big/ \binom{n}{2} = 1 - N \Big/ \binom{n}{2}, \tag{2}$$

Tabela 3. Formuły służące do określania liczby par obiektów zakwalifikowanych do jednego z czterech typów

Typ	Formuła
(I)	$\sum_{s=1}^{u}\sum_{r=1}^{v} n_{sr}(n_{sr}-1)$
(II)	$\frac{1}{2}\left[n^2+\sum_{s=1}^{u}\sum_{r=1}^{v} n_{sr}^2-\left(\sum_{s=1}^{u} n_{s\cdot}^2+\sum_{r=1}^{v} n_{\cdot r}^2\right)\right]$
(III)	$\frac{1}{2}\left(\sum_{r=1}^{v} n_{\cdot r}^2-\sum_{s=1}^{u}\sum_{r=1}^{v} n_{sr}^2\right)$
(IV)	$\frac{1}{2}\left(\sum_{s=1}^{u} n_{s\cdot}^2-\sum_{s=1}^{u}\sum_{r=1}^{v} n_{sr}^2\right)$
$(I)+(II)=Z=\binom{n}{2}+\sum_{s=1}^{u}\sum_{r=1}^{v} n_{sr}^2-\frac{1}{2}\left(\sum_{s=1}^{u} n_{s\cdot}^2+\sum_{r=1}^{v} n_{\cdot r}^2\right)$	
$(III)+(IV)=N=\frac{1}{2}\left(\sum_{s=1}^{u} n_{s\cdot}^2+\sum_{r=1}^{v} n_{\cdot r}^2\right)-\sum_{s=1}^{u}\sum_{r=1}^{v} n_{sr}^2$	
$Z+N=\binom{n}{2}$	

Źródło: opracowano na podstawie pracy [35, s. 196].

gdzie: Z i N są określone wzorami w tab. 3.

Przedział zmienności tej miary zaczyna się od 0, kiedy to dwa podziały $P^{(t)}$ i $P^{(q)}$ są zupełnie niepodobne (jeden podział zawiera tyle klas, ile jest obiektów, a drugi jedną klasę zawierającą wszystkie obiekty), a kończy na 1, kiedy podziały są identyczne. Miarę Randa (2) interpretuje się jako odsetek par obiektów zgodnych w obu klasyfikacjach $P^{(t)}$ i $P^{(q)}$ w ogólnej liczbie par obiektów określonych na zbiorze A.

Wadą miary Randa jest to, że wykazuje tendencję do wzrostu wartości w przypadku zwiększania liczby klas (por. [17, s. 182]). Hubert i Arabie [35, s. 198] zaproponowali skorygowany indeks Randa:

$$R_{HA}=\frac{R-E(R)}{R_{\max}-E(R)}, \tag{3}$$

gdzie: $R_{\max}$ – maksymalna wartość miary Randa ($R_{\max}=1$),

$E(R)$ – wartość oczekiwana miary Randa określona wzorem:

$$E(R)=1+2\sum_{r}\binom{n_{\cdot r}}{2}\sum_{s}\binom{n_{s\cdot}}{2}\Big/\binom{n}{2}^2-\left[\sum_{r}\binom{n_{\cdot r}}{2}+\sum_{s}\binom{n_{s\cdot}}{2}\right]\Big/\binom{n}{2}.$$

Skorygowana miara Randa przyjmuje postać [35, s. 198]:

$$R_{HA} = \frac{\sum_{r,s}\binom{n_{rs}}{2} - \sum_{r}\binom{n_{\cdot r}}{2}\sum_{s}\binom{n_{s\cdot}}{2}\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_{r}\binom{n_{\cdot r}}{2} + \sum_{s}\binom{n_{s\cdot}}{2}\right] - \sum_{r}\binom{n_{\cdot r}}{2}\sum_{s}\binom{n_{s\cdot}}{2}\Big/\binom{n}{2}}. \quad (4)$$

$R_{HA} = 0$ oznacza, że podziały zbioru obiektów na r i s klas zostały wyodrębnione losowo.

W podejściu trzecim (stosowanym dla metod klasyfikacji hierarchicznej) za właściwą dla danego typu danych metodę klasyfikacji należy uznać taką, która daje minimalne zniekształcenia przy transformacji wyjściowej macierzy odległości $[d_{ik}]$ w macierz wartości kofenetycznych $[h_{ik}]$ (inaczej wartości poziomu połączenia klas w dendrogramie). Wartości h_{ik} (dla każdego i, k) w macierzy $[h_{ik}]$ odczytuje się z dendrogramu, który wskazuje wartości poziomu połączenia klas P_i oraz P_k.

W tab. 4 przedstawiono trzy mierniki pomiaru zniekształcenia przy transformacji $[d_{ik}] \rightarrow [h_{ik}]$. Małe wartości D.2 i D.3 oraz duże wartości D.1 oznaczają małe zniekształcenia przy transformacji $[d_{ik}] \rightarrow [h_{ik}]$ przez daną metodę klasyfikacji.

Tabela 4. Miary zniekształcenia przy transformacji $[d_{ik}] \rightarrow [h_{ik}]$

Lp.	Nazwa	Miara	Źródło
D.1	Współczynnik korelacji kofenetycznej	$\frac{\sum_{i,k}(d_{ik}-\bar{d})(h_{ik}-\bar{h})}{\left[\sum_{i,k}(d_{ik}-\bar{d})^2\sum_{i,k}(h_{ik}-\bar{h})^2\right]^{0,5}}$	[61]
D.2	Suma kwadratów odchyleń	$\sum_{i,k} w_{ik}(d_{ik}-h_{ik})^2$	[32]
D.3	Metryka Minkowskiego	$\begin{cases}\left[\sum_{i,k}\lvert d_{ik}-h_{ik}\rvert^{\lambda}\right] & (0<\lambda\le 1)\\ \max_{i,k}\{\lvert d_{ik}-h_{ik}\rvert\} & (\lambda=0)\end{cases}$	[42]

w_{ik} – wagi (na ogół wszystkie odległości są jednakowo ważne, więc $w_{ik} = 1$).

Źródło: opracowano na podstawie prac [13; 23].

Pewną słabością oparcia się w wyborze właściwej (w odniesieniu do danego typu danych) metody klasyfikacji na miarach tego typu jest to, że na ogół metodą wybieraną przez D.1 jest metoda średniej klasowej [60; 61], a przez D.3 – metoda pojedynczego połączenia [23].

W **czwartym podejściu** analizuje się formalne własności metod klasyfikacji, które mogą stanowić pomocne kryterium wyboru właściwej metody. Pierwsze własności formalne wypracowali Jardine i Sibson [42]. Zostały one następnie wzbogacone nowymi w pracach: [18; 65], a zwięzły ich przegląd w literaturze zawierają następujące monografie: [22; 25, s. 98-100; 53; 2]. Wybrane własności formalne metod klasyfikacji można ująć w postaci następujących punktów:

a) własność wypukłości – metody klasyfikacji mają tę własność, jeżeli w wyniku zastosowania otrzymuje się podział zbioru obiektów A na klasy P_1, ..., P_H, w którym wypukłe otoczenia klas się nie przecinają,

b) własność poprawnej struktury według klas – wszystkie odległości wewnątrzklasowe są mniejsze od wszystkich odległości międzyklasowych,

c) własność poprawnej struktury według drzewka połączeń – metody klasyfikacji mają tę własność, jeżeli rezultaty klasyfikacji hierarchicznej zbioru obiektów dadzą się przedstawić w postaci drzewka połączeń (dendrogramu) zgodnego z kolejnością podobieństwa między elementami tego zbioru (wartości poziomu połączenia klas h_{ik} są rozłożone monotonicznie rosnąco, gdy stosujemy miary odległości między obiektami),

d) własność monotoniczności – metody klasyfikacji mają tę własność, jeżeli monotoniczna transformacja każdego elementu macierzy odległości nie zmienia wyników klasyfikacji,

e) własność powtarzania punktów – metody klasyfikacji mają tę własność, jeśli po dodaniu jednego lub wielu obiektów, identycznych z obiektami należącymi do klas P_1, ..., P_H, i ponownym zastosowaniu danej metody granice klas się nie zmienią (zmieni się tylko ich liczebność).

f) własność opuszczania klas – niech dany będzie podział zbioru obiektów A na klasy P_1, ..., P_k, ..., P_H. Jeśli po odrzuceniu obiektów należących do klasy P_k i ponownym zastosowaniu algorytmu klasyfikacji otrzymamy podział zbioru $A - P_k$ na klasy P_1, ..., P_{k-1}, P_k, P_{k+1}, ..., P_H, to dana metoda klasyfikacji ma własność opuszczania klas.

W tab. 5 w sposób syntetyczny przedstawiono formalne własności hierarchicznych metod aglomeracyjnych. Znajomość określonych własności poszczególnych metod klasyfikacji pozwala na właściwe ich wykorzystanie w badaniach empirycznych.

Tabela 5. Formalne własności wybranych hierarchicznych metod aglomeracyjnych

Metoda	Własności					
	A	B	C	D	E	F
Pojedynczego połączenia (*single-link*)	–	+	+	+	+	+
Kompletnego połączenia (*complete-link*)	–	+	+	+	+	+
Średniej klasowej (*group average-link*)	–	+	+	–	–	+
Powiększona suma kwadratów odległości (*incremental sum of squares*)	+	–	+	–	–	+
Środka ciężkości (*centroid*)	–	–	–	–	–	+

A – wypukłości, B – poprawnej struktury według klas, C – poprawnej struktury według drzewka połączeń, D – monotoniczności, E – powtarzania punktów, F – opuszczania klas.
„+" – spełnia; „–" – nie spełnia.

Źródło: opracowano na podstawie prac [22, s. 131; 36; 53; 67, s. 59; 65, s. 424].

7. Ustalanie liczby klas

Milligan i Cooper przetestowali na podstawie zbiorów danych o znanej strukturze klas 30 procedur[3] pozwalających wyznaczyć liczbę klas. Większość metod przedstawionych w pracy Milligana i Cooper [50] opierała się, przy wyznaczaniu liczby klas, na wyjściowej macierzy danych. Niektóre z nich wykorzystywały odległości wewnątrzklasowe i międzyklasowe. Trzy najlepsze kryteria globalne są następujące:

1. Indeks Calińskiego i Harabasza [11]:

$$G1(u) = \frac{\text{trace}(\mathbf{B})/(u-1)}{\text{trace}(\mathbf{W})/(n-u)}, \tag{5}$$

gdzie: **B** (**W**) – macierz kowariancji międzyklasowej (wewnątrzklasowej),
trace – ślad macierzy,
u – liczba klas,
n – liczba obiektów.

Indeks Calińskiego i Harabasza nazywany jest pseudostatystyką F (zob. [45, s. 291]).

2. Indeks Huberta i Levine [34]:

$$G2(u) = \frac{D(u) - r \cdot D_{\min}}{r \cdot D_{\max} - r \cdot D_{\min}}, \tag{6}$$

gdzie: $D(u)$ – suma wszystkich odległości wewnątrzklasowych,
r – liczba odległości wewnątrzklasowych,
$D_{\min}$ ($D_{\max}$) – najmniejsza (największa) odległość wewnątrzklasowa.

[3] Przedstawiony przegląd nie wyczerpuje zbioru istniejących sposobów wyznaczania liczby klas. Inne sposoby zawarte są m.in. w pracach [63; 66].

3. Indeks Gamma Bakera i Huberta [6]:

$$G3(u) = \frac{s(+) - s(-)}{s(+) + s(-)}, \tag{7}$$

gdzie: $s(+)$ – liczba odległości wewnątrzklasowych mniejszych od odległości międzyklasowych,

$s(-)$ – liczba odległości wewnątrzklasowych większych od odległości międzyklasowych.

Maksymalna wartość $G1(u)$ i $G3(u)$ oraz minimalna $G2(u)$ wskazuje najlepszy podział zbioru obiektów, a zarazem wyznacza liczbę klas.

8. Walidacja wyników klasyfikacji

Analiza replikacji (powtórzenie klasyfikacji)

Replikacja w przypadku zagadnienia klasyfikacji dotyczy przeprowadzenia procesu klasyfikacji zbioru obiektów na podstawie dwóch prób wylosowanych z danego zbioru danych, a następnie oceny zgodności otrzymanych rezultatów. Procedura replikacji składa się z następujących etapów [25, s. 184; 49, s. 368-369]:

1. Podzielić losowo zbiór danych (zbiór n obiektów opisanych m zmiennymi) na dwa podzbiory A i B.
2. Zastosować wybraną metodę klasyfikacji do podziału zbioru A na ustaloną liczbę klas u. Wyznaczyć środki ciężkości (*centroids*) dla poszczególnych klas.
3. Obliczyć odległości obiektów ze zbioru B od środków ciężkości klas wyznaczonych na podstawie podzbioru A.
4. Przydzielić obiekty z podzbioru B do klas zawierających najbliższy środek ciężkości. Prowadzi to do podziału podzbioru B na nie więcej niż u klas.
5. Zastosować tę samą metodę klasyfikacji do podziału podzbioru B na u klas.
6. Policzyć, np. za pomocą skorygowanej miary Randa, zgodność wyników dwóch podziałów podzbioru B. Poziom zgodności wyników dwóch podziałów podzbioru B odzwierciedla stabilność przeprowadzonej klasyfikacji zbioru obiektów.

Ocena jakości klasyfikacji

Syntetyczny miernik pozwalający mierzyć prawidłowość zaklasyfikowania poszczególnych obiektów do klas, prawidłowość wyodrębnienia poszczególnych klas oraz ogólną jakość klasyfikacji (relatywną zwartość i separowalność klas) zaproponował Rousseeuw [57] (zob. [43, s. 83-88]). Wskaźnik Rousseeuwa (*silhouette index*), pozwalający oceniać prawidłowość zaklasyfikowania poszczególnych obiektów do klas, przyjmuje postać:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{8}$$

gdzie: $a(i)$ – średnia odległość obiektu i od pozostałych obiektów należących do klasy P_r; $a(i) = \sum_{k \in \{P_r \setminus i\}} d_{ik}/(n_r - 1)$;

$b(i) = \min_{s \neq r}\{d_{iP_s}\}$; d_{iP_s} – średnia odległość obiektu i od obiektów należących do klasy P_s ($d_{iP_s} = \sum_{k \in P_s} d_{ik}/n_s$);

$r, s = 1, 2, \ldots, u$ – numer klasy,

u – liczba klas.

Indeks $S(i)$ przyjmuje wartości z przedziału [–1; 1]. Im bliżej wartości jeden, tym dany obiekt silniej należy do wyodrębnionej klasy. W przypadku klas jednoelementowych $S(i) = 0$.

Indeksy pozwalające mierzyć prawidłowość wyodrębnienia poszczególnych klas oraz ogólną jakość klasyfikacji (relatywną zwartość i separowalność klas) są następujące:

$$S(P_r) = \sum_{i \in P_r} S(i)/n_r, \tag{9}$$

$$S(P) = \sum_{r} S(P_r)/u. \tag{10}$$

Subiektywną ocenę przedziałów wartości miernika $S(P)$ zawiera tab. 6.

Tabela 6. Interpretacja wartości miernika $S(P)$

$S(P)$	Interpretacja
(0,70; 1,00]	silna struktura klas
(0,50; 0,70]	poważna struktura klas
(0,25; 0,50]	słaba struktura klas (należy zastosować inne metody klasyfikacji)
0,25 i mniej	nie odkryto struktury klas

Źródło: [43, s. 88].

9. Opis (interpretacja) i profilowanie klas

W wyniku zastosowania do klasyfikacji zbioru obiektów wybranej metody klasyfikacji otrzymuje się podział tego zbioru na klasy P_1, ..., P_u. W badaniach marketingowych podstawowym zagadnieniem stają się w związku z tym:

- **Opis (interpretacja) otrzymanych wyników**, tj. wskazanie cech charakterystycznych poszczególnych klas oraz wyjaśnienie, jakimi czynnikami różnią się wyodrębnione klasy. Podstawą opisu (interpretacji) wyodrębnionych klas są zmienne, które brały udział w procesie klasyfikacji zbioru obiektów.

Dla ułatwienia interpretacji otrzymanych rezultatów klasyfikacji wyznacza się środki ciężkości poszczególnych klas (średnie arytmetyczne obliczone z wartości pierwotnych każdej zmiennej na podstawie obiektów tworzących daną klasę) oraz odchylenia standardowe zmiennych w poszczególnych klasach. Na ten sposób rozwiązania problemu interpretacji rezultatów klasyfikacji wskazują m.in.: Hair, Anderson, Tatham i Black [31, s. 443]; Sokołowski [63, s. 47]; Jajuga [40, s. 134]; Robles i Sarathy [56]. Oczywiście taki sposób opisu klas możliwy jest do zastosowania tylko wtedy, gdy zmienne użyte w zagadnieniu klasyfikacji zbioru obiektów są mierzone na skali przedziałowej i (lub) ilorazowej (dopiero bowiem w tych skalach dopuszcza się użycie średniej arytmetycznej i odchylenia standardowego).

Jeśli klasyfikacja jest przeprowadzana na podstawie zmiennych mierzonych na skali porządkowej lub nominalnej, to możliwe jest wyznaczenie opisowej (werbalnej) charakterystyki poszczególnych klas dla każdej zmiennej. Można wyznaczyć frakcje i odsetki występowania w danej klasie poszczególnych kategorii zmiennych.

- **Profilowanie klas**. Celem profilowania klas jest wskazanie cech charakterystycznych poszczególnych klas pozwalających na ukazanie różnic pomiędzy nimi. Profilowanie klas przeprowadza się na podstawie zmiennych, które nie brały udziału w procesie klasyfikacji zbioru obiektów.

  Typowymi zmiennymi stosowanymi w profilowaniu klas w badaniach marketingowych są zmienne demograficzne, geograficzne, socjoekonomiczne, psychograficzne i in., które charakteryzują konsumentów (nabywców) poszczególnych klas. Profilowanie przeprowadza się zwykle z wykorzystaniem takich metod, jak (por. np. [31, s. 501, 513-515; 59, s. 180]): analiza dyskryminacyjna, drzewa klasyfikacyjne, tabulacja krzyżowa (tablice kontyngencji).

10. Podsumowanie

W artykule w syntetycznej formie zaprezentowano problemy decyzyjne wymagające rozstrzygnięcia w procesie klasyfikacji zbioru obiektów obejmującym osiem etapów. Wykorzystując światową literaturę przedmiotu, scharakteryzowano podejścia służące rozstrzygnięciu pojawiających się problemów decyzyjnych w procesie klasyfikacji zbioru obiektów. Opracowanie ma charakter porządkujący wiedzę z omawianego zakresu. Przy prezentacji niektórych etapów wprowadzono własne komentarze i rozwiązania.

Literatura

[1] Abrahamowicz M., *Konstrukcja syntetycznych mierników rozwoju w świetle twierdzenia Arrowa*, Prace Naukowe AE we Wrocławiu nr 311, 1985, 5-25.

[2] Ajvazjan S.A., Bežaeva Z.I., Staroverov O.V., *Klassifikacija mnogomernych nabludenij*, Statistika, Moskva 1974.

[3] Aldenderfer M.S., Blashfield R.K., *Cluster Analysis*, Sage, Beverly Hills 1984.

[4] Anderberg M.R., *Cluster Analysis for Applications*, Academic Press, New York, San Francisco, London 1973.

[5] Arabie P., Hubert L.J., de Soete G. (Eds.), *Clustering and Classification*, World Scientific, Singapore 1996.

[6] Baker F.B., Hubert L.J., *Measuring the power of hierarchical cluster analysis*, „Journal of the American Statistical Asscociation" 1975, 70, 31-38.

[7] Bartosiewicz S., *Ekonometria. Technologia ekonometrycznego przetwarzania informacji*, PWE, Warszawa 1989.

[8] Bock H.H., Diday E. (Eds.), *Analysis of Symbolic Data*, Springer-Verlag, Berlin, Heidelberg 2000.

[9] Borys T., *Kategoria jakości w statystycznej analizie porównawczej*, Prace Naukowe Akademii Ekonomicznej nr 284, Seria: Monografie i opracowania nr 23, AE, Wrocław 1984.

[10] Brundage T.W., *Use and Stability of Classifications of Firms in Oregon's Health Care Industry*. [w:] *Proceedings of the Business and Economics Statistics Section*, New Orleans, Luisiana 1988, August, 22-25, 218-222.

[11] Caliński R.B., Harabasz J., *A Dendrite Method for Cluster Analysis*, „Communications in Statistics", 1974, vol. 3, 1-27.

[12] Cieślak M., *Dobór syndromu zmiennych do porządkowania liniowego obiektów wielowymiarowych*, Prace Naukowe Akademii Ekonomicznej nr 328, AE, Wrocław 1986, 19-27.

[13] Cormack R.M., *A Review of Classification (with Discussion)*, „Journal of the Royal Statistical Society" 1971, Ser. A, part 3, 321-367.

[14] Cox T.F., Cox M.A.A., *Multidimensional Scaling*, Chapman & Hall, London 1994.

[15] Cox T.F., Cox M.A.A., *A General Weighted Two-way Dissimilarity Coefficient*, „Journal of Classification" 2000, Vol. 17, 101-121.

[16] Czekanowski J., *Zarys metod statystycznych w zastosowaniu do antropologii*, Towarzystwo Naukowe Warszawskie, Warszawa 1913.

[17] Everitt B.S., Landau S., Leese M., *Cluster Analysis*, Edward Arnold, London 2001.

[18] Fisher L., van Ness J.W., *Admissible Clustering Procedures*, „Biometrika" 1971, no. 1, 91-104.

[19] Florek K., Łukasiewicz J., Perkal J., Steinhaus H., Zubrzycki S., *Taksonomia wrocławska*, „Przegląd Antropologiczny" 1951, (17), 193-211.

[20] Fowlkes E.B., Mallows C.L., *A Method for Comparing Two Hierarchical Clusterings*, „Journal of the American Statistical Association" 1983, no. 383, 553-569.

[21] Goodman L.A., Kruskal W.H., *Measures of Association for Cross Classifications*, Springer-Verlag, New York, Heidelberg 1979.

[22] Gordon A.D., *Classification*, Chapman and Hall, London 1981.

[23] Gordon A.D., *A Review of Hierarchical Classification*, „Journal of the Royal Statistical Society" 1987, ser. A, 119-137.

[24] Gordon A.D., *Hierarchical Classification*, [w:] P. Arabie, L.J. Hubert, G. de Soete (Eds.), *Clustering and Classification*, World Scientific, Singapore 1996, 65-121.

[25] Gordon A.D., *Classification*, Chapman and Hall/CRC, London 1999.

[26] Gower J.C. (), *A General Coefficient of Similarity and Some of its Properties*. „Biometrics" 1971, (27), 857-874.

[27] Grabiński T., Wydymus S., Zeliaś A., *Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych*, PWN, Warszawa 1989.
[28] Grabiński T., *Problemy analizy poprawności procedur taksonomicznych*, [w:] J. Pociecha (red.), *Materiały z konferencji nt. Taksonomia – teoria i jej zastosowania*, Wydawnictwo AE, Kraków 1990, 61-71.
[29] Grabiński T., *Metody taksonometrii*, Wydawnictwo AE, Kraków 1992.
[30] Hair J.F., Anderson R.E., Tatham R.L., Black W.C., *Multivariate Data Analysis with Readings*, Prentice Hall, Englewood Cliffs 1995.
[31] Hair J.F., Anderson R.E., Tatham R.L., Black W.C., *Multivariate Data Analysis*, Prentice-Hall, Englewood Cliffs 1998.
[32] Hartigan J.A., *Representation of Similarity Matrices by Trees*, „Journal of the American Statistical Association" 1967, vol. 62, 1140-1158.
[33] Hartigan J.A., *Classification*, [w:] *Encyclopedia of Statistical Sciences*, vol. 2, Wiley, New York, 1982, 1-10.
[34] Hubert L.J., Levine J.R., *Evaluating Object Set Partitions: Free Sort Analysis and Some Generalizations*, „Journal of Verbal Learning and Verbal Behaviour" 1976, 15, 549-570.
[35] Hubert L.J., Arabie P., *Comparing Partitions*, „Journal of Classification" 1985, no. 1, 193-218.
[36] Hussain M., *Taksonomiczne metody podziału zbiorów skończonych*, AE, Krakowie 1982 (praca doktorska).
[37] Jajuga K., *Metody analizy wielowymiarowej w ilościowych badaniach przestrzennych*, AE, Wrocław 1981 (praca doktorska).
[38] Jajuga K., *O sposobach określania ilości klas w zagadnieniach klasyfikacji i klasyfikacji rozmytej*, Prace Naukowe Akademii Ekonomicznej nr 262, AE, Wrocław 1984, 17-29.
[39] Jajuga K., *Podstawowe metody analizy wielowymiarowej w przypadku występowania zmiennych mierzonych na różnych skalach*, Akademia Ekonomiczna we Wrocławiu 1989, praca wykonana w ramach CPBP 10.09.
[40] Jajuga K., *Statystyczna teoria rozpoznawania obrazów*, PWN, Warszawa 1990.
[41] Jajuga K., Walesiak M., *Standardisation of Data Set Under Different Measurement Scales*, [w:] Decker R., Gaul W. (Eds.), *Classification and Information Processing at the Turn of the Millennium*, Springer-Verlag, Berlin, Heidelberg 2000, 105-112.
[42] Jardine N., Sibson R., *Mathematical Taxonomy*, Wiley, New York 1971.
[43] Kaufman L., Rousseeuw P.J., *Finding groups in Data: an Introduction to Cluster Analysis*, Wiley, New York 1990.
[44] Kolonko J., *O wykorzystaniu w badaniach taksonomicznych danych pierwotnych mierzonych na skalach różnego typu*. Materiały konferencyjne nt. *Metody taksonomiczne i ich zastosowanie w badaniach ekonomicznych*, Szklarska Poręba, 25 października 1979 r. (materiał powielony).
[45] Lattin J.M., Carroll J.D., Green P.E., *Analyzing Multivariate Data*, Brooks/Cole, Pacific Grove 2003.
[46] Lerman J.C., *Comparing Partitions (Mathematical and Statistical Aspects)*, [w:] H.H. Bock (ed.): *Classification and Related Methods of Data Analysis*, North-Holland, Amsterdam 1988, 121-131.
[47] Lira J., Wagner W., Wysocki F., *Mediana w zagadnieniach porządkowania liniowego 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 2002, 87-99
[48] Milligan G.W., *A Review of Monte Carlo Tests of Cluster Analysis*, „Multivariate Behavioral Research" 1981, (16), 379-407.
[49] Milligan G.W., *Clustering Validation: Results and Implications for Applied Analyses*, [w:] P. Arabie, L.J. Hubert, G. de Soete (Eds.), *Clustering and Classification*, World Scientific, Singapore 1996, 341-375.

[50] Milligan G.W., Cooper M.C., *An Examination of Procedures for Determining the Number of Clusters in a Data Set*, „Psychometrika" 1985, no. 2, 159-179.
[51] Nowak E., *Wskaźnik podobieństwa wyników podziałów*, „Przegląd Statystyczny" 1985, z. 1, 41-48.
[52] Nowak E., *Metody taksonomiczne w klasyfikacji obiektów społeczno-gospodarczych*, PWE, Warszawa 1990.
[53] Pociecha J., *Kryteria oceny procedur taksonomicznych*, „Przegląd Statystyczny" 1982, z. 1/2, 183-190.
[54] Pociecha J., *Statystyczne metody segmentacji rynku*, Zeszyty Naukowe AE w Krakowie, Seria specjalna: Monografie nr 71, Kraków 1986.
[55] Rand W.M., *Objective Criteria for the Evaluation of Clustering Methods*, „Journal of the American Statistical Association" 1971, no. 336, 846-850.
[56] Robles F., Sarathy R., *Segmenting the Commuter Aircraft Market with Cluster Analysis*, „Industrial Marketing Management" 1986, vol. 15, 1-12.
[57] Rousseeuw P.J., *Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis*, „Journal of Computational and Applied Mathematics" 1987, 20, 53-65.
[58] Rybaczuk M., *Graficzna prezentacja struktury danych wielowymiarowych*, [w:] K. Jajuga, M. Walesiak (red.), *Klasyfikacja i analiza danych – teoria i zastosowania*. Taksonomia 10, Prace Naukowe Akademii Ekonomicznej nr 942, AE, Wrocław 2002, 146-153.
[59] Sagan A., *Badania marketingowe. Podstawowe kierunki*, Wydawnictwo AE, Kraków 1998.
[60] Sneath P.H.A., *Evaluation of Clustering Methods (with discussion)*, [w:] A.J. Cole (ed.), *Numerical Taxonomy*, Academic Press, London 1969, 257-271.
[61] Sokal R.R., Rohlf F.J., *The Comparison of Dendrograms by Objective Methods*, „Taxon" 1962, no. 2, 33-40.
[62] Sokołowski A., *Metoda porównywania wyników podziału zbioru skończonego*, [w:] XII Konferencja Naukowa Ekonometryków, Statystyków i Matematyków Akademii Ekonomicznych Polski Południowej, Karpacz 1976.
[63] Sokołowski A., *Empiryczne testy istotności w taksonomii*, Zeszyty Naukowe AE w Krakowie, Seria specjalna: Monografie nr 108, Kraków 1992.
[64] Szmigiel C., *Wskaźnik zgodności kryteriów podziału*, „Przegląd Statystyczny" 1976, z. 4, 491-498.
[65] van Ness J.W., *Admissible Clustering Procedures*, „Biometrika" 1973, (60), 422-424.
[66] Walesiak M., *Sposoby wyznaczania optymalnej liczby klas w zagadnieniu klasyfikacji hierarchicznej*, Prace Naukowe Akademii Ekonomicznej nr 449, AE, Wrocław 1988, 63-72.
[67] Walesiak M., *Statystyczna analiza wielowymiarowa w badaniach marketingowych*, Prace Naukowe Akademii Ekonomicznej nr 654, Seria: Monografie i Opracowania nr 101, AE, Wrocław 1993.
[68] Walesiak M., *Metody analizy danych marketingowych*, PWN, Warszawa 1996.
[69] Walesiak M., *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, Wydawnictwo AE, Wrocław 2002.
[70] Walesiak M., *Miara odległości obiektów opisanych zmiennymi mierzonymi na różnych skalach pomiaru*, XXXIX Konferencja Ekonometryków, Statystyków i Matematyków Akademii Ekonomicznych Wrocławia, Krakowa i Katowic, Lądek Zdrój, 2-5 marca 2003. Prace Naukowe Akademii Ekonomicznej we Wrocławiu (w redakcji).
[71] Wallace D.L., *Comment*, „Journal of the American Statistical Association" 1983, (78), no. 383, 569-576.

DECISION PROBLEMS IN A CLUSTER ANALYSIS PROCEDURE

Summary

In the article eight major steps in a cluster analysis are discussed (see [Milligan 1996, 342-343]):
1. Selection of objects to cluster,
2. Selection of variables to be used,
3. Decisions concerning variable normalisation,
4. Selection of a distance measure,
5. Selection of clustering method,
6. Determining the number of clusters,
7. Cluster validation,
8. Describing and profiling clusters.

The sequence represents the critical steps, or decisions, that are made in a cluster analysis.

Marek Walesiak jest pracownikiem Katedry Ekonometrii i Informatyki w Akademii Ekonomicznej we Wrocławiu.