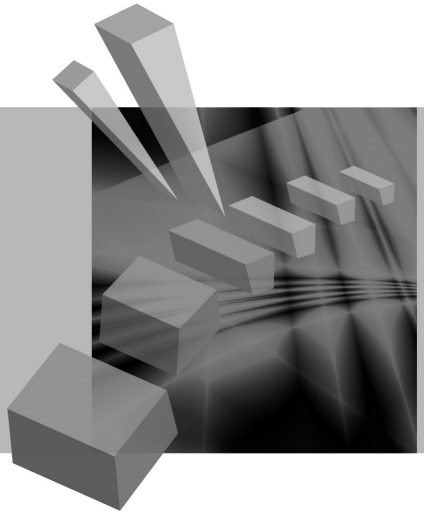
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

278

Taksonomia 20

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy danych PTS

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 278 • 2013

Taksonomia 20. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Artur Mikulec
Uniwersytet Łódzki

KRYTERIUM MOJENY I WISHARTA W ANALIZIE SKUPIEŃ – PRZYPADEK SKUPIEŃ O RÓŻNYCH MACIERZACH KOWARIANCJI

Streszczenie: Kryteria Mojeny i Wisharta są metodami wyboru optymalnego wyniku grupowania stosowanymi w przypadku metod aglomeracyjnych analizy skupień. Celem artykułu jest prezentacja wyników empirycznej analizy efektywności kryteriów Mojeny i Wisharta wyboru liczby skupień – na tle analizowanych dotychczas kryteriów Bakera i Huberta, Calińskiego i Harabasza, Daviesa i Bouldina, Huberta i Levine'a – w przypadku skupień o różnych macierzach kowariancji. Analiza empiryczna została przeprowadzona z wykorzystaniem programu *ClustanGraphics 8* oraz pakietu `clusterSim` środowiska `R`.

Słowa kluczowe: reguła górnego obszaru odrzucenia, reguła średniej ruchomej, kryteria Mojeny, kryterium Wisharta (*tree validation*), ClustanGraphics.

1. Wstęp

Etap oceny wyniku grupowania, tj. wyboru liczby skupień w analizie wykorzystującej hierarchiczne algorytmy grupowania (ze względu na ich własności), jest jednym z końcowych, lecz niezwykle ważnych etapów w klasyfikacji. Mając bowiem cały ciąg klasyfikacji $P_0, P_1, ..., P_{n-1}$, należy na podstawie pewnych formalnych kryteriów podjąć decyzję o wyborze ostatecznego wyniku grupowania.

Celem artykułu jest prezentacja wyników empirycznej analizy efektywności dwóch kryteriów Mojeny [1977] bazujących na analizie odległości łączenia kolejnych obiektów na wykresie drzewa – *best cut significance test* (*upper tail rule, moving average quality control rule*), oraz kryterium Wisharta [2006] oceny losowości podziału obiektów na wykresie drzewa – *tree validation*[1]. Wymienione wyżej kryteria zostały porównane z punktu widzenia wyboru liczby klas (oraz ich struktury) z innymi, powszechnie wykorzystywanymi w tym celu, procedurami wyboru liczby skupień: Bakera i Huberta (BH), Calińskiego i Harabasza (CH), Daviesa i Bouldina (DB) czy Huberta i Levine'a (HL). W artykule rozważane są przypadki skupień ge-

[1] Ich omówienie na tle innych metod wyboru liczby skupień odnaleźć można w pracy Mikulca [2012].

nerowanych w oparciu o różne macierze kowariancji zmiennych (skupienia zróżnicowane dla klas)[2].

2. Metody wyboru liczby skupień

Problem oceny efektywności procedur wyboru liczby klas był już w literaturze przedmiotu poruszany wielokrotnie, poczynając od historycznych i najbardziej znanych prac empirycznych prezentujących wyniki tego rodzaju analiz w kontekście metod hierarchicznych [Milligan, Cooper 1985[3]; Milligan 1996], a skończywszy na pracy, w której przedstawiono wyniki analizy empirycznej wybranych procedur analizy skupień, w tym indeksów służących ustalaniu liczby klas dla metod klasyfikacji hierarchicznej, dla danych porządkowych [Walesiak 2011]. Jeśli natomiast spojrzeć szerzej na metody ustalania liczby skupień, w kontekście oceny jakości wyniku grupowania, to kompleksowy i usystematyzowany przegląd literatury z tego zakresu z lat 1908-2011 odnaleźć można w pracy pt. *Ocena jakości wyników grupowania – przegląd bibliografii* [Migdał-Najman 2011].

Dwa kryteria Mojeny oraz kryterium Wisharta – analizowane w artykule – to jedne z niewielu procedur wyboru liczby skupień (obok indeksu Beale'a, Dudy i Harta, indeksu *RMSSTD* oraz *RS* [4]) dedykowane metodom klasyfikacji hierarchicznej, np. aglomeracyjnej. Niemniej także inne wymienione we wstępie procedury mogą być zastosowane jako kryteria wyboru liczby skupień dla metod aglomeracyjnych – różnią się one konstrukcją kryterium wewnętrznego oceny wyniku grupowania. W tabeli 1 zamieszczono tylko wybrane metody oceny liczby skupień, będące przedmiotem porównań w artykule.

Tabela 1. Metody oceny liczby skupień w zbiorze danych*

KRYTERIUM	Formuła, przedział zmienności	Kryterium wyboru liczby skupień
1	2	3
Bakera i Huberta	$BH(u)=\frac{S_+ - S_-}{S_+ + S_-}$, $BH(u)\in\langle -1;1\rangle$	$\hat{u}=\arg\max_u\left[BH(u)\right]$
Calińskiego i Harabasza	$CH(u)=\frac{tr(B_u)/(u-1)}{tr(W_u)/(n-u)}$, $CH(u)\in R_+$	$\hat{u}=\arg\max_u\left[CH(u)\right]$

[2] *Empiryczna analiza efektywności kryterium Mojeny i Wisharta w analizie skupień* – przypadek skupień generowanych na postawie tej samej (jednakowej) macierzy kowariancji zmiennych była tematem artykułu wygłoszonego podczas Kongresu Statystyki Polskiej w Poznaniu, 18-20 kwietnia 2012 r. [Mikulec, Fijałkowska-Kupis 2012].

[3] Analiza wykazała, że pięcioma najlepszymi regułami wyboru liczby skupień były kryteria: Calińskiego i Harabasza, Dudy i Harta, Huberta i Levine'a, Backera i Huberta oraz Beale'a (*F-ratio*). W pierwszej dziesiątce omawianych procedur znalazło się również pierwsze kryterium Mojeny (górnego obszaru odrzucenia).

[4] Indeks *RMSSTD* to miara jednorodności skupień oparta na sumie kwadratów odległości wewnątrz skupień, indeks *RS* to miara niepodobieństwa między skupieniami oparta na sumie kwadratów odległości pomiędzy skupieniami odniesionej do sumy kwadratów odległości między obiektami w całym zbiorze danych [Gan i in. 2007].

Tabela 1, cd.

1	2	3
Daviesa i Bouldina	$BD(u)=\frac{1}{u}\sum_{q=1}^{u}\max_{r,q\neq r}\left(\frac{S_q+S_r}{d(q,r)}\right)$	$\hat{u}=\arg\min_{u}\left[BD(u)\right]$
Huberta i Lewine'a	$HL(u)=\frac{D(u)-l_wD_{\min}}{l_wD_{\max}-l_wD_{\min}}$, $HL(u)\in(0;1)$	$\hat{u}=\arg\min_{u}\left[HL(u)\right]$
Górnego obszaru odrzucenia (Mojena I)	$\alpha_{x+1}>\bar{\alpha}+k\cdot s_\alpha$	klasyfikacja P_x, aby odpowiadający jej krok $x: x=1,...,n-2$ pierwszy spełniał nierówność
Średniej ruchomej (Mojena II)	$\alpha_{x+1}>\bar{\alpha}_x+L_x+b_x+k\cdot s_x$. gdzie: $L_x=\frac{(y-1)b_x}{2}$, $b_x=\frac{6\left[2\sum_{f=x-y+1}^{x}w_f\alpha_f-(y+1)\sum_{f=x-y+1}^{x}\alpha_f\right]}{y(y^2-1)}$, $w_f=w_{f-1}+1$, $f=(x-y+2),...,x$, $w_{x-y+1}=1$	klasyfikacja P_x, aby odpowiadający jej krok $x: x=y,y+1,...,n-2$ pierwszy spełniał nierówność
Losowości podziału obiektów na wykresie drzewa (Wishart)	Porównywanie wyników ciągu klasyfikacji uzyskanych metodami aglomeracyjnymi z rodziną drzew generowanych na podstawie losowej permutacji zbioru danych	H_0 mówiąca o tym, że struktura grupowania obiektów w postaci danego drzewa jest losowa (brak struktury), $H_1 :\sim H_0$

* n – liczba obiektów $(i=1,...,n)$; m – liczba cech $(j=1,...,m)$; u – liczba grup $(q,r,s=1,...,u)$; K_q – skupienie q; S_+, S_- – liczba par odległości, odpowiednio zgodnych i niezgodnych; $tr(B_u)$, $tr(W_u)$ – ślad macierzy kowariancji, odpowiednio międzygrupowej (B_u) i wewnątrzgrupowej (W_u); $S_q=\sqrt[t]{(1/n_q)\sum_{i\in K_q}\sum_{j=1}^{m}\left|x_{ij}^{q}-z_{qj}\right|^t}$ – miara rozproszenia obiektów w grupie q (K_q), przy czym dla $t=1$ jest ona średnią odległością obiektów w skupieniu q (K_q) od środka ciężkości, tj. medoidy w grupie, a dla t = 2 jest ona odchyleniem standardowym odległości obiektów w skupieniu q (K_q) od środka ciężkości, tj. medoidy w grupie (dla grupy r miarę S_r wprowadza się analogicznie); $d(q,r)=\sqrt[p]{\sum_{j=1}^{m}\left|z_{qj}-z_{rj}\right|^p}$ – miara odległości między środkami ciężkości, tj. medoidami (z_{qj}, z_{rj}) grup q i r, odpowiednio miejskiej dla p = 1 lub euklidesowej dla p = 2; $D(u)$ – suma wszystkich odległości wewnątrzgrupowych; l_w – liczba odległości wewnątrzgrupowych; $D_{\min}, D_{\max}$ – odległość wewnątrzgrupowa, odpowiednio najmniejsza i największa; $\alpha_x=\min_{i<o}[d_{io}]$, $(i,o=1,...,n-x)$ – miara niepodobieństwa (odległości) między skupieniami; α_{x+1} – poziom (odległość) połączenia grup w kroku x + 1, $\bar{\alpha}$ – średni poziom (odległość) połączenia grup, s_α – odchylenie standardowe poziomu (odległości) połączenia grup; k-stała $k\in(2{,}75;3{,}5)$; y – liczba wartości poziomu (odległości) połączenia klas α w danym kroku (do wyznaczenia średniej ruchomej); $\bar{\alpha}_x$ – średnia ruchoma wartości parametru α obliczona w kroku x; L_x – korekta dla opóźnionego „trendu" poziomu (odległości) połączenia klas obliczona w kroku x; b_x – „ruchome" średniokwadratowe nachylenie linii trendu poziomu połączenia klas w kroku x; s_x – „ruchome" odchylenie standardowe wartości parametru α (odległości).

Źródło: opracowanie własne na podstawie [Mojena 1977; Wishart 2006; Gatnar, Walesiak (red.) 2009].

3. Założenia oraz schemat analizy empirycznej

Empiryczna analiza efektywności dwóch kryteriów Mojeny i kryterium Wisharta na tle pozostałych czterech kryteriów – Bakera i Huberta (BH), Calińskiego i Harabasza (CH), Daviesa i Bouldina (DB), czy Huberta i Levine'a (HL) – przeprowadzona została dla:

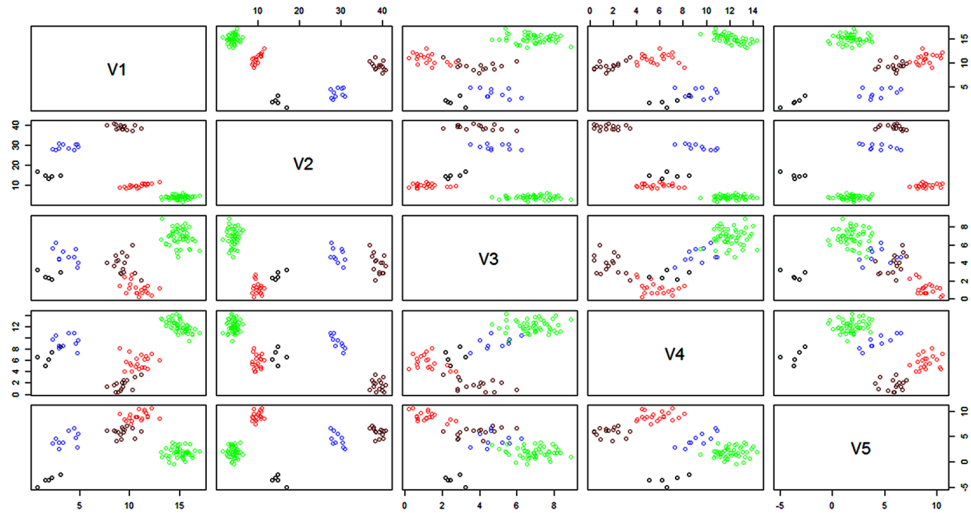
- 2-5 skupień;
- 2-5 zmiennych;
- skupień o następującej strukturze dla 100 obiektów:
  - 2 skupienia zawierające odpowiednio 40 i 60 obiektów,
  - 3 skupienia zawierające odpowiednio 20, 30 i 50 obiektów,
  - 4 skupienia zawierające odpowiednio 5, 15, 25 i 55 obiektów,
  - 5 skupień zawierających odpowiednio 5, 10, 15, 20 i 50 obiektów;
- skupień bez zmiennych zakłócających;
- skupień generowanych na podstawie różnych macierzy kowariancji zmiennych, powodujących różne rozproszenie obiektów w skupieniach, a więc różny kształt skupień (skupienia zróżnicowane dla klas), zob. rys. 1,
- miary odległości euklidesowej,
- trzech najczęściej stosowanych metod aglomeracyjnych – pełnego wiązania, średniego wiązania i Warda.

W rezultacie analizie poddano 16 zbiorów danych[5], biorąc pod uwagę 4 warianty liczby skupień, 4 warianty liczby zmiennych, wykorzystując w tym celu 3 metody aglomeracyjne. Na rysunku 1 zaprezentowano jeden z analizowanych zbiorów danych wygenerowany dla 5 skupień i 5 zmiennych na podstawie różnych macierzy kowariancji zmiennych, powodujących zróżnicowanie kształtu skupień (wydłużone, sferyczne) oraz różny stopień ich separowalności.

Obliczenia przeprowadzone zostały na zbiorach danych wygenerowanych poleceniem `cluster.Gen` pakietu `clusterSim` [Walesiak, Dudek 2012] środowiska `R` oraz z wykorzystaniem programu *ClustanGraphics 8* [Wishart 2006]. Ich schemat był następujący:

- krok 1 – wygenerowano zbiory danych według przyjętych założeń (16 zbiorów), dla których znano właściwą strukturę skupień (2-5 skupień),
- krok 2 – w programie *ClustanGraphics 8* dokonano analizy skupień z wykorzystaniem 3 algorytmów grupowania aglomeracyjnego (48 wyników), a wynik wykresu drzewa zawierającego wszystkie podziały zbioru obiektów zapisano do pliku,
- krok 3 – w programie *ClustanGraphics 8* dokonano wyboru liczby skupień (wyniku grupowania) według dwóch kryteriów Mojeny i kryterium Wisharta,

[5] Ze względu na ograniczoną objętość artykułu nie jest możliwe przedstawienie pełnej charakterystyki analizowanych zbiorów danych.



$$m = \begin{bmatrix} 2 & 15 & 2 & 7 & -3 \\ 4 & 29 & 5 & 10 & 5 \\ 9 & 39 & 4 & 1 & 6 \\ 11 & 10 & 1 & 6 & 9 \\ 15 & 4 & 7 & 12 & 2 \end{bmatrix} \quad \text{cov1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0.1 & 0 \\ 0 & 0 & 1 & 0 & -0.1 \\ 0 & 0.1 & 0 & 1 & 0 \\ 0 & 0 & -0.1 & 0 & 1 \end{bmatrix} \quad \text{cov2} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0.5 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.7 \\ 0.5 & 0 & 0 & 0.7 & 1 \end{bmatrix}$$

$$\text{cov3} = \begin{bmatrix} 1 & 0 & 0.3 & 0.6 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0.3 & 0 & 1 & 0 & 0 \\ 0.6 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{cov4} = \begin{bmatrix} 1 & 0.8 & 0 & 0 & 0 \\ 0.8 & 1 & 0.2 & 0 & 0 \\ 0 & 0.2 & 1 & -0,6 & 0 \\ 0 & 0 & -0,6 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{cov5} = \begin{bmatrix} 1 & 0 & 0 & -0.7 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ -0.7 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Rys. 1. Zbiór wygenerowany dla 5 skupień i 5 zmiennych

Źródło: opracowanie własne.

- krok 4 – w środowisku `R` (`clusterSim`) obliczono pozostałe indeksy wyboru liczby skupień – Bakera i Huberta (BH), Calińskiego i Harabasza (CH), Daviesa i Bouldina (DB), Huberta i Levine'a (HL) dla podziałów w zakresie 2-10 skupień, a następnie wybrano optymalne rozwiązanie według danego kryterium wyboru liczby klas: BH (max), CH (max), DB (min), HL (min),
- krok 5 – mając rozwiązanie, tj. wynik analizy skupień – strukturę skupień wskazaną przez każde kryterium wyboru liczby skupień dla każdego analizowanego zbioru danych, obliczono skorygowany indeks Randa[6] zgodności

[6] Skorygowany indeks Randa, niemający tendencji do wzrostu wartości w przypadku zwiększania liczby klas, porównuje przynależność wszystkich par obiektów według dwóch porównywanych klasyfikacji. Pozwala określić odsetek par obiektów zgodnych w obydwu porównywanych klasyfikacjach. Szczegóły dotyczące skorygowanego indeksu Randa odnaleźć można w pracy Walesiaka [2011].

przynależności obiektów do skupień powstałych na bazie danego kryterium na tle właściwego podziału analizowanego zbioru obiektów (znaną strukturą klas),

- krok 6 – biorąc pod uwagę wszystkie wyniki grupowania, tj. poprawne pod względem struktury i przynależności obiektów do skupień – o wartościach skorygowanych indeksów Randa zbliżonych do jedności, oraz niepoprawne pod względem struktury i przynależności obiektów do skupień – o wartościach skorygowanych indeksów Randa zbliżonych do zera, oceniono efektywność analizowanych kryteriów wyboru liczby skupień, uśredniając wartość tego indeksu dla każdej z metod aglomeracyjnych i każdego kryterium.

4. Wyniki analizy empirycznej

W tabeli 2 dla każdego analizowanego kryterium wyboru liczby skupień przedstawiono liczbę poprawnych oraz błędnych wskazań liczby skupień względem metody grupowania aglomeracyjnego dla 16 analizowanych zbiorów danych, w których skupienia wygenerowane zostały na podstawie różnych macierzy kowariancji zmiennych – służą one ocenie trafności wskazań liczby skupień przez poszczególne kryteria.

Kryterium Bakera i Huberta – niezależnie od metody grupowania aglomeracyjnego w ok. 2/3 przypadków wskazało właściwą liczbę poszukiwanych klas. Zbliżone wyniki poprawności dla tego kryterium uzyskano we wcześniejszej analizie, tj. dla zbiorów o skupieniach generowanych za pomocą jednakowej macierzy kowariancji zmiennych.

Lepsze wyniki poprawności wyboru liczby skupień uzyskano na podstawie kryterium Calińskiego i Harabasza, dla którego poziom poprawności wskazań – bez względu na metodę grupowania aglomeracyjnego – wynosił co najmniej 81,25%. W przypadku wcześniejszej analizy – dla skupień z tą samą macierzą kowariancji zmiennych – uzyskane wyniki były zbliżone, choć nieco gorsze.

Indeks Daviesa i Bouldina zdecydowanie częściej wskazywał poprawną liczbę skupień w analizowanych zbiorach danych dla metody pełnego wiązania oraz Warda, natomiast dla metody średniego wiązania poziom trafności jego wskazań nie przekraczał 50%. Niemalże identyczna sytuacja pod względem trafności wyboru liczby klas dla tego kryterium według metody aglomeracyjnej występowała w analizie skupień generowanych na podstawie jednakowej macierzy kowariancji zmiennych.

Dwa kolejne kryteria, tj. Huberta i Levine'a oraz górnego obszaru odrzucenia (Mojeny), w ogóle nie sprawdziły się z punktu widzenia wyboru liczby skupień w zbiorach danych, w których skupienia są wygenerowane na podstawie różnych macierzy kowariancji zmiennych. Zdecydowanie częściej wskazywały niepoprawną liczbę skupień, podobnie jak w przypadku wcześniejszej analizy – zbiorów danych o skupieniach z jednakową macierzą kowariancji zmiennych.

Na podstawie wyników wskazań poprawnej liczby skupień dla drugiego kryterium Mojeny (średniej ruchomej) można stwierdzić, że w przypadku metod średnie-

Tabela 2. Wskazania liczby skupień według kryteriów wyboru liczby skupień

METODA	Wskazanie			
	poprawne		błędne	
Bakera i Huberta (BH)				
Średniego wiązania	10	62,50%	6	37,50%
Pełnego wiązania	11	68,75%	5	31,25%
Warda	11	68,75%	5	31,25%
Calińskiego i Harabasza (CH)				
Średniego wiązania	14	87,50%	2	12,50%
Pełnego wiązania	13	81,25%	3	18,75%
Warda	14	87,50%	2	12,50%
Daviesa i Bouldina (DB)				
Średniego wiązania	7	43,75%	9	**56,25%**
Pełnego wiązania	13	81,25%	3	18,75%
Warda	13	81,25%	3	18,75%
Huberta i Levine'a (HL)				
Średniego wiązania	3	18,75%	13	**81,25%**
Pełnego wiązania	6	37,50%	10	**62,50%**
Warda	1	6,25%	15	**93,75%**
Górnego obszaru odrzucenia (Mojena I)				
Średniego wiązania	2	12,50%	14	**87,50%**
Pełnego wiązania	1	6,25%	10	**93,75%**
Warda	6	37,50%	15	**62,50%**
Średniej ruchomej (Mojena II)				
Średniego wiązania	11	68,75%	5	31,25%
Pełnego wiązania	9	56,25%	7	43,75%
Warda	0	0,00%	16	**100,00%**
Losowości podziału obiektów na wykresie drzewa (Wishart)				
Średniego wiązania	10	62,50%	6	37,50%
Pełnego wiązania	8	50,00%	8	50,00%
Warda	9	56,25%	7	43,75%

Źródło: opracowanie własne.

go i pełnego wiązania w większości przypadków pozwalało ono wybrać właściwą liczbę skupień w analizowanych zbiorach danych, niemniej poziom trafności tych wskazań jest stosunkowo niski (56,25 i 68,7%). Charakterystyczne jest, iż kryterium to całkowicie nie sprawdziło się w analizie skupień metodą Warda. Warto dodać, że podobne wyniki (zbliżoną poprawność tego kryterium dla pierwszych dwóch metod aglomeracyjnych oraz jego nieprzydatność przy metodzie Warda) uzyskano we wcześniejszych analizach dla zbiorów danych o skupieniach tworzonych na podsta-

wie tych samych macierzy kowariancji zmiennych. Trudno jednak stwierdzić, czy w przypadkach obydwu tych analiz błąd kryterium średniej ruchomej nie był wynikiem zastosowania metody Warda z miarą odległości euklidesowej zamiast kwadratu tej odległości.

Ostatnie z analizowanych kryteriów – losowości podziału obiektów na wykresie drzewa – w ponad połowie przypadków, bez względu na metodę grupowania aglomeracyjnego, poprawnie wskazało liczbę poszukiwanych klas. Warto dodać, że w analizie zbiorów danych o skupieniach z jednakową macierzą kowariancji zmiennych, wykonanej wcześniej, omawiane kryterium częściej wskazywało właściwą, poszukiwaną liczbę skupień.

Należy zdawać sobie sprawę, iż sama poprawność (częstość) wskazywania przez poszczególne kryteria właściwej liczby skupień jest pierwszą, ale nie dostateczną przesłanką do oceny efektywności, tzn. przydatności, danego kryterium w zakresie wyboru liczby skupień. Istotna jest również zgodność danego wyniku grupowania pod względem przynależności obiektów do ich właściwych skupień, a więc zgodność wyniku grupowania ze znaną strukturą klas dla wygenerowanych zbiorów danych, którą oceniono za pomocą skorygowanego indeksu Randa.

Tabela 3. Zgodność wyniku grupowania według kryteriów wyboru liczby skupień

METODA	Kryterium	Średnia wartość skorygowanego indeksu Randa
Średniego wiązania	Bakera i Huberta (BH)	0,902
	Calińskiego i Harabasza (CH)	**0,948**
	Daviesa i Bouldina (DB)	0,859
	Huberta i Levine'a (HL)	0,849
	Górnego obszaru odrzucenia (Mojena I)	**0,390**
	Średniej ruchomej (Mojena II)	0,947
	Losowości podziału obiektów na wykresie drzewa (Wishart)	**0,948**
Pełnego wiązania	Bakera i Huberta (BH)	0,904
	Calińskiego i Harabasza (CH)	**0,936**
	Daviesa i Bouldina (DB)	0,935
	Huberta i Levine'a (HL)	0,879
	Górnego obszaru odrzucenia (Mojena I)	**0,284**
	Średniej ruchomej (Mojena II)	0,867
	Losowości podziału obiektów na wykresie drzewa (Wishart)	0,828
Warda	Bakera i Huberta (BH)	0,930
	Calińskiego i Harabasza (CH)	**0,937**
	Daviesa i Bouldina (DB)	0,935
	Huberta i Levine'a (HL)	0,696
	Górnego obszaru odrzucenia (Mojena I)	0,708
	Średniej ruchomej (Mojena II)	**0,491**
	Losowości podziału obiektów na wykresie drzewa (Wishart)	0,870

Źródło: opracowanie własne.

Stąd też w tab. 3 zaprezentowano wyniki zgodności wyniku grupowania ze znaną strukturą klas, uśredniając wartości skorygowanego indeksu Randa względem każdej z metod aglomeracyjnych i każdego kryterium, a uśrednienia tego dokonano, biorąc pod uwagę strukturę skupień wszystkich wyników analizy skupień (odnoszących się do wszystkich 16 zbiorów danych), zarówno tych o „poprawnym", jak i o „błędnym" wskazaniu liczby skupień przez poszczególne kryterium (zob. tab. 2).

Tym samym w ocenie efektywności uwzględniono dwa aspekty – liczbę „dobrych" i „złych" rozwiązań (wyników grupowania) wskazanych przez poszczególne kryteria wyboru liczby skupień oraz zgodność każdego wyniku grupowania ze znaną strukturą klas.

5. Podsumowanie i wnioski

Biorąc pod uwagę wyniki przeprowadzonych analiz (por. tab. 2, 3), można stwierdzić, że spośród rozpatrywanych procedur wyboru liczby skupień dla metod aglomeracyjnych najbardziej efektywne okazały się kryteria: Calińskiego i Harabasza (CH) oraz losowości podziału obiektów na wykresie drzewa (Wisharta) – które może nie zawsze okazywało się tym najlepszym na tle pozostałych (jak w przypadku metody pełnego wiązania i Warda), ale z reguły wskazywało właściwą liczbę wyodrębnionych klas i gwarantowało stosunkowo wysoką zgodność wyniku grupowania ze znaną strukturą klas. Natomiast w przypadku metody aglomeracyjnej średniego wiązania kryterium losowości podziału obiektów na wykresie drzewa (Wisharta) okazało się równie wysoce skuteczne jak kryterium Calińskiego i Harabasza (CH).

Zestawiając ze sobą trafność wskazań liczby skupień oraz zgodność wyniku grupowania ze znaną strukturą klas dla wszystkich wyników analizy wyraźnie należy zauważyć, że najsłabsze okazały się kryteria: Huberta i Levine'a (HL) oraz górnego obszaru odrzucenia (pierwsze kryterium Mojeny, Mojena I).

Z kolei kryterium średniej ruchomej (drugie kryterium Mojeny, Mojena II) charakteryzowało się stosunkowo dobrą efektywnością w przypadku metody aglomeracyjnej średniego i pełnego wiązania, lecz w ogóle nie sprawdziło się w analizie skupień z wykorzystaniem aglomeracyjnej metody Warda – być może wynikało to np. z zastosowania w obliczeniach miary odległości euklidesowej.

Uzyskane rezultaty w pewnym stopniu zależą od analizowanych przykładów – założeń analizy empirycznej, w tym sposobu generowania danych o znanej strukturze klas. Generowanie danych losowo z wielowymiarowego rozkładu normalnego na podstawie macierzy wartości średnich i macierzy kowariancji to jeden z możliwych sposobów – często wykorzystywany w analizach symulacyjnych. Być może losowe generowanie danych na podstawie innych rozkładów wielowymiarowych oraz tzw. funkcji połączenia (*copula*) lub wykorzystanie danych zawierających skupienia o zadanym stopniu separowalności (nowsze podejście do generowania danych) pozwoliłoby na sformułowanie bardziej ogólnych wniosków. Niemniej jednak wciąż

podstawowym problemem pozostaje fakt, iż dla danych o znanej strukturze klas istnieje nieskończenie wiele kształtów skupień dla dowolnej liczby wymiarów i nie jest możliwe przebadanie każdego z nich.

Literatura

Gan G., Ma C., Wu J., *Data Clustering: Theory, Algorithms, and Applications*, SIAM, Philadelphia 2007.

Gatnar E., Walesiak M. (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo PWN, Warszawa 2009.

Migdał-Najman K., *Ocena jakości wyników grupowania – przegląd bibliografii*, „Przegląd Statystyczny" 2011, vol. 3-4, s. 281-299.

Mikulec A., *Metody oceny wyniku grupowania w analizie skupień*, [w:] K. Jajuga, M. Walesiak (red.), Taksonomia 19, *Klasyfikacja i analiza danych – teoria i zastosowania*, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2012.

Mikulec A., Fijałkowska-Kupis A., *An empirical analysis of the effectiveness of Wishart and Mojena criteria in cluster analysis*, „Statistics in Transition – new series" 2012, vol. 13(3), p. 569-580.

Milligan G.W., *Clustering Validation: Results and Implication for Applied Analysis*, [w:] P. Arabie, L.J. Hubert, G. De Soete (red.), *Clustering and Classification,* World Scientific Publishing Co. Pte. Ltd., Singapore 1996.

Milligan G.W., Cooper M.C., *An examination of procedures for determining the number of clusters in a data set*, „Psychometrika" 1985, vol. 50(2), p. 159-179.

Mojena R., *Hierarchical grouping methods and stopping rules: an evaluation*, „Computer Journal" 1977, vol. 20(4), p. 359-363.

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

Walesiak M., Dudek A., *The clusterSim package* (wersja 0.41-5, 1 marca 2012), http://keii.ue.wroc.pl/clusterSim/, Wrocław 2012.

Wishart D., *Clustangraphics Primer: a Guide To Cluster Analysis*, (4th edition), Edinburgh 2006.

MOJENA AND WISHART CRITERION IN CLUSTER ANALYSIS – THE CASE OF CLUSTERS WITH DIFFERENT COVARIANCE MATRICES

Summary: Mojena and Wishart criteria are designed to facilitate the choice of the optimal clustering solution in the case of agglomeration methods in cluster analysis. The aim of the paper is to present the empirical study on efficiency of Mojena and Wishart criteria in the choice of the number of clusters. The study was conducted with the focus on clusters with different covariance matrices and the results were compared to previously analysed criteria of Baker and Hubert, Caliński and Harabasz, Davies and Bouldin, Hubert and Levine. The empirical analysis was made with the use of *ClustanGraphics 8* program and `clusterSim` package of `R` environment.

Keywords: upper tail rule, moving average quality control rule, Mojena criteria, Wishart criterion (tree validation), ClustanGraphics.