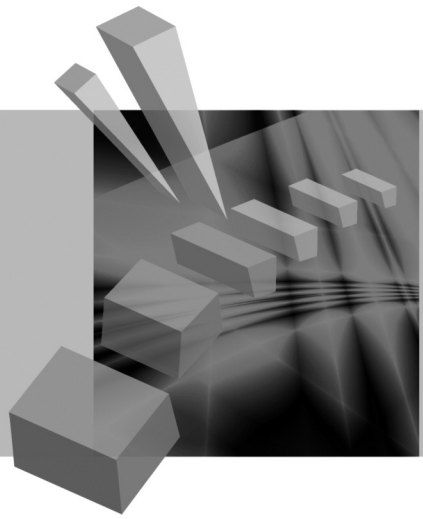
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w 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

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Artur Mikulec
Uniwersytet Łódzki

METODY OCENY WYNIKU GRUPOWANIA W ANALIZIE SKUPIEŃ

Streszczenie: W artykule dokonano przeglądu metod oceny wyniku grupowania. Omówiono trzy metody wyboru właściwej liczby skupień dla metod aglomeracyjnych zaproponowane przez Mojenę i Wisharta oraz zaimplementowane w programie *ClustanGraphics 8*. Wspomniane kryteria bazują na relatywnych wartościach różnych poziomów połączeń obiektów na wykresie drzewa – *best cut significance test* (*upper tail, moving average*) oraz sprawdzaniu losowości podziału obiektów na wykresie drzewa – *tree validation*. Treść artykułu zilustrowana została przykładem empirycznym.

Słowa kluczowe: analiza skupień, ocena wyniku grupowania, kryteria Mojeny, kryterium Wisharta.

1. Wstęp

Metody oceny wyniku grupowania – w szerokim rozumieniu – związane są z trzema zagadnieniami analizy skupień: wyborem liczby skupień, porównywaniem wyników dwóch (i więcej) klasyfikacji oraz oceną jakości wyniku grupowania. W artykule dokonano zestawienia wybranych metod oceny wyniku grupowania, a przede wszystkim omówiono trzy metody wyboru właściwej liczby skupień dla metod hierarchicznych (aglomeracyjnych) zaproponowane przez Mojenę [1977] i Wisharta dostępne w programie *ClustanGraphics 8* [Wishart 2006]. Wspomniane kryteria bazują na: analizie odległości łączenia kolejnych obiektów na wykresie drzewa – *best cut significance test* (*upper tail rule, moving average quality control rule*) oraz ocenie losowości podziału obiektów na wykresie drzewa – *tree validation*. Przypomnienie i prezentacja tych metod wydają się zasadne z dwóch powodów: po pierwsze, metody aglomeracyjne są jednymi z najbardziej rozpowszechnionych i najczęściej wykorzystywanych w praktyce metod analizy skupień; po drugie, wyniki wielu analiz wskazują na subiektywne podejście autorów do wyboru liczby skupień na wykresie drzewa, tj. brak zastosowania formalnego kryterium dla tej oceny. Treść artykułu zilustrowana została przykładem empirycznym.

2. Metody oceny wyniku grupowania

Ogólny podział metod oceny wyniku grupowania stosowany w literaturze jest następstwem informacji wykorzystywanych w tego rodzaju analizie. Wyróżnia się metody sprawdzania wyniku grupowania oparte na kryterium wewnętrznym, zewnętrznym oraz względnym[1] [Gan i in. 2007]. Oceny na podstawie **kryterium wewnętrznego** dokonuje się tylko na bazie kryteriów ilościowych i informacji pochodzących z analizowanego zbioru danych. Dla pojedynczego wyniku klasyfikacji (indywidualnych skupień) uzyskanego metodami podziałowymi analizuje się stopień podobieństwa pomiędzy strukturą grupowania *C*, tj. przynależnością obiektów do skupień, a macierzą sąsiedztwa (odległości) obiektów *P*. W przypadku ciągu klasyfikacji uzyskanego metodami hierarchicznymi analizuje się ich strukturę, tzn. odległości łączenia kolejnych obiektów na wykresie drzewa. Porównuje się stopień podobieństwa macierzy odległości łączenia obiektów na wykresie drzewa *H* z macierzą sąsiedztwa (odległości) obiektów *P* – kryteria Mojeny. Inna metoda oceny wyniku grupowania hierarchicznego opiera się na testowaniu braku struktury klas [Gordon 1996] i może odbywać się na podstawie oceny losowości podziału obiektów na wykresie drzewa – kryterium Wisharta.

W podejściu według **kryterium zewnętrznego** ocenia się wyniki algorytmu grupowania uzyskane na podstawie wcześniej ustalonej struktury klas, która wynika z tego zbioru danych i odzwierciedla intuicyjnie jego strukturę. Ocena wyników grupowania odbywa się na podstawie struktury grupowania *C* i niezależnego podziału C_0 (wynik innego grupowania C_0) lub też na podstawie macierzy sąsiedztwa obiektów *P* i innego podziału C_0 (wynik innego grupowania C_0).

Uwzględniając trzy właściwe dla analizy skupień zagadnienia oceny wyniku grupowania, tj. poszukiwanie właściwej liczby skupień, porównywanie wyników dwóch klasyfikacji i ocenę jakości klasyfikacji, metody oceny wyniku grupowania można podzielić zgodnie ze schematem przedstawionym w tab. 1. Przegląd ten z pewnością nie wyczerpuje zbioru istniejących metod, zawiera jednak metody najczęściej stosowane, sprawdzone empirycznie oraz metody względnie nowe i stosunkowo słabo rozpowszechnione w polskiej literaturze i w praktyce badawczej. Tematem dalszej części artykułu będzie prezentacja trzech metod wyboru właściwej liczby skupień dla metod hierarchicznych (aglomeracyjnych) zaproponowanych przez Mojenę i Wisharta.

[1] Bez straty dla dalszych rozważań pominięto omawianie kryterium względnego, gdyż metody na nim oparte definiowane są w ten sam sposób jak metody oparte na kryterium wewnętrznym. Ocena wyniku grupowania w przypadku kryterium względnego opiera się na iteracyjnym wyborze najlepszego rezultatu grupowania ze zbioru pewnych dopuszczalnych rozwiązań.

Tabela 1. Metody oceny wyników analizy skupień[a]

Wyszczególnienie	Kryterium	
	wewnętrznej informacji, oparte na[b]: $\{C, P\}$, $\{H, P\}$, braku struktury klas	zewnętrznej informacji, oparte na[b]: $\{C, C_0\}$ lub $\{P, C_0\}$
Wybór właściwej liczby skupień	indeks: Bakera i Huberta; Beale'a (*F-ratio*); Calińskiego i Harabasza; Daviesa-Bouldina; Dudy i Harta; Dunna; *gap* (odstępu); Hartigana; Huberta i Levine'a; Krzanowskiego i Lai; *RMSSTD*[c]; *RS*; *SD*; *S_Dbw*; metoda: *jump*; *ps*; *clest*; miara: Γ-Huberta; znormalizowana Γ-Huberta; kryterium: Mojeny – górnego obszaru odrzucenia (*upper tail rule*), średniej ruchomej (*moving average quality control rule*); Wisharta – losowości podziału obiektów na wykresie drzewa (*tree validation*); współczynnik: korelacji kofenetycznej	–
Porównywanie wyników dwóch klasyfikacji	metryka Minkowskiego; suma kwadratów odchyleń; współczynnik: korelacji kofenetycznej	indeks: Fowlkesa-Mallowsa; Jaccarda; Lermana; Wallace'a; miara: Γ-Huberta; znormalizowana Γ-Huberta; Randa; skorygowana Randa; wskaźnik: podobieństwa Nowaka
Ocena jakości klasyfikacji	indeks: sylwetkowy (*silhouette index*); miara: średniej zwartości skupień (*average of compactness*)	–

[a] Przez podkreślenie wyróżniono metody służące do oceny wyniku grupowania hierarchicznego; [b] C – struktura grupowania, P – macierz sąsiedztwa (odległości) obiektów, H – macierz odległości łączenia obiektów na wykresie drzewa, C_0 – wynik innego grupowania obiektów; [c] Dla metod hierarchicznych należy stosować łącznie z indeksem *RS*.

Źródło: opracowanie własne na podstawie [Cormack 1971; Denoeud i in. 2005; Fowlkes, Mallows 1983; Gan i in. 2007; Gatnar, Walesiak 2009; Gordon 1987; Kaufman, Rousseeuw 2005; Migdał-Najman, Najman 2008; Milligan, Cooper 1985; Mojena 1977; Nowak 1985; Sugar, James 2003; Tibshirani i in. 2001; Wishart 2006].

3. Kryterium *upper tail rule* oraz *moving average quality control rule*

Jedną z najbardziej znanych w literaturze przedmiotu prac poświęconych metodom wyboru liczby skupień jest artykuł [Milligan, Cooper 1985], dotyczący empirycznej analizy i oceny 30 procedur wyboru liczby skupień. Autorzy zidentyfikowali pięć najlepszych wówczas „reguł zatrzymania" (według kolejności): Calińskiego i Hara-

basza, Dudy i Harta, Huberta i Levine'a, Backera i Huberta oraz Beale'a (*F-ratio*)[2]. W pierwszej dziesiątce omawianych procedur znalazło się też pierwsze kryterium Mojeny – *upper tail rule* bazujące na względnych wartościach połączeń obiektów na wykresie drzewa.

Grupowanie *N* obiektów metodami hierarchicznymi (aglomeracyjnymi), które na początku stanowią *N* skupień, polega na krokowym łączeniu pojedynczych obiektów, tj. redukowaniu o 1 liczby skupień do czasu, aż wszystkie obiekty zostaną włączone do jednego skupienia. Wyniki grupowania hierarchicznego (aglomeracyjnego) N obiektów można opisać za pomocą zbioru kolejnych klasyfikacji $P_0, P_1, ..., P_{N-1}$ oraz korespondujących z nimi wartości kryterium klasyfikacji $\alpha_0, \alpha_1, ..., \alpha_{N-1}$, gdzie krok (subskrypt) $j = 0,1, ..., N-1$ odpowiada uzyskanej kolejno $N, N-1, ..., 1$ liczbie skupień. Struktura grupowania podziału P_j jest reprezentowana przez skupienia $C_1, C_2, ..., C_{N-j}$.

Mojena zauważył, że różnice pomiędzy wynikami poszczególnych metod aglomeracyjnych są jedynie następstwem odmiennego sposobu definiowania miary połączenia klas (wszystkie metody opierają się na procedurze Lance'a i Williamsa i dla każdej z nich uzyskuje się inny ciąg klasyfikacji $P_0, P_1, ..., P_{N-1}$ i korespondujące z nimi wartości kryterium $\alpha_0, \alpha_1, ..., \alpha_{N-1}$ – poziomu połączenia klas) oraz znaczenia, w jakim zdefiniowany jest parametr α. Jeżeli w analizie w charakterze miary połączenia przyjmie się miarę niepodobieństwa (odległości) między skupieniami, a więc jej większe wartości implikują większe niepodobieństwo grup, to [Mojena 1977]:

$$\alpha_j = \min_{i<m} [d_{im}], \; i, m = 1, ..., N - j, \tag{1}$$

w przeciwnym razie (przyjmując miarę podobieństwa między skupieniami):

$$\alpha_j = \max_{i<m} [d_{im}], \; i, m = 1, ..., N - j. \tag{2}$$

Statystyczne reguły służące selekcji najlepszego „podziału" badanej zbiorowości mogą być oparte na rozkładzie wartości kryterium α lub na odpowiedniej transformacji tego kryterium. Jeżeli większe wartości miary połączenia implikują niepodobieństwo jednostek, to rozkład wartości parametru α jest monotonicznie rosnący dla metod aglomeracyjnych: pojedynczego i kompletnego połączenia, średniej klasowej, ważonej średniej klasowej i Warda, przy zastosowaniu wzoru (1) – przypadek A oraz monotonicznie malejący dla tych metod aglomeracyjnych, przy zastosowaniu wzoru (2) – przypadek B. Stąd też Mojena zaproponował dwie metody, które bazują na ocenie „istotności" zmian wartości parametru α stanowiącego kryterium klasyfikacji.

Pierwsza reguła wykorzystuje $N-1$ wartości parametru α do wyznaczenia jego wartości średniej i odchylenia standardowego z próby celem zdefiniowania

[2] Kryteria Dudy i Harta oraz Beale'a są odpowiednie dla zagnieżdżonej struktury klas.

„istotnej" wartości α, która jako pierwsza leży w „górnym" – wzór (1) lub w „dolnym" – wzór (2) obszarze odrzucenia rozkładu jego wartości. Dla metod aglomeracyjnych: pojedynczego i kompletnego połączenia, średniej klasowej, ważonej średniej klasowej i Warda – wzór (1), przypadek A – ze zbioru kolejnych klasyfikacji $P_1, ..., P_{N-1}$ należy wybrać taką klasyfikację P_j, aby odpowiadający jej krok j $(j = 1 ..., N-2)$ pierwszy spełniał nierówność sformułowaną dla poziomu połączenia klas α [Mojena 1977]:

$$\alpha_{j+1} > \overline{\alpha} + k \cdot s_\alpha, \tag{3}$$

natomiast dla przypadku B uwzględniającego wzór (2) sens nierówności (3) jest odwrotny – należy wybrać taką klasyfikację P_j, aby odpowiadający jej krok j, $(j = 1 ..., N-2)$ pierwszy spełniał nierówność [Balicki 2009]:

$$\alpha_{j+1} < \overline{\alpha} - k \cdot s_\alpha, \tag{4}$$

gdzie α_{j+1} to poziom połączenia klas w kroku $j+1$, $\overline{\alpha}$ i s_α to odpowiednio średnia i nieobciążone odchylenie standardowe poziomu połączenia klas, a k to pewna stała[3].

Jeżeli żadna z wartości α_j $(j = 1, ..., N-2)$ poziomu połączenia klas nie spełnia nierówności (3) lub (4), to należy: (a) przyjąć, że wszystkie analizowane obiekty tworzą jedno skupienie – brak rozwiązania, (b) wybrać taki krok *j*, dla którego w kroku *j* + 1 wartość $(\alpha_{j+1} - \overline{\alpha}) / s_\alpha$ jest maksymalna, (c) wybrać inną regułę podziału obiektów[4].

Druga reguła podziału zaproponowana przez Mojenę jest bardziej skomplikowana i bazuje na metodzie średniej ruchomej, mającej szerokie zastosowanie w kontroli jakości. Mojena zauważył, że zachowanie się wartości α (poziomu połączenia klas) nie jest odmienne od zachowania się wartości szeregu czasowego z trendem, co zasugerowało, że lepszym podejściem jest sformułowanie 1-krokowego modelu przewidującego, który zdefiniuje istotną wartość parametru α

[3] Autor sugerował przyjąć w obliczeniach wartość $k \in (2{,}75; 3{,}50)$, z kolei Milligan i Cooper [1985] na podstawie przeprowadzonych badań empirycznych sugerowali wartość *k* = 1,25. Alternatywnie zamiast k można przyjąć wartość z rozkładu *t*-Studenta, co oznacza dodatkowe założenie, że α, tj. poziom połączenia klas na wykresie drzewa, podlega rozkładowi normalnemu.

[4] Dla tej reguły Mojeny wybranie wariantu (a) lub (b) daje ten sam wynik dla monotonicznych wartości α_j, jeżeli $\overline{\alpha}$ i s_α są deterministyczne (ustalone), oraz różny wynik, jeżeli $\overline{\alpha}$ i s_α są stochastyczne.

w kroku $j+1$ [5]. Dla metod aglomeracyjnych: pojedynczego i kompletnego połączenia, średniej klasowej, ważonej średniej klasowej i Warda – wzór (1), przypadek A – celem reguły jest wybór takiej klasyfikacji P_j, aby odpowiadający jej krok j $\left(j = r, r+1, ..., N-2\right)$ pierwszy spełniał nierówność [Mojena 1977]:

$$\alpha_{j+1} > \overline{\alpha}_j + L_j + b_j + k \cdot s_j, \tag{5}$$

gdzie: r oznacza liczbę wartości poziomu połączenia klas α w danym kroku – liczbę wartości do wyznaczania średniej ruchomej[6], $\overline{\alpha}_j$ jest średnią ruchomą wartości parametru α obliczoną w kroku j, L_j jest korektą dla opóźnionego trendu poziomu połączenia klas obliczoną w kroku j, b_j jest „ruchomym" średniokwadratowym nachyleniem linii trendu poziomu połączenia klas w kroku j, k jest pewną stałą, a s_j jest „ruchomym" nieobciążonym oszacowaniem odchylenia standardowego wartości parametru α, przy czym wartości L_j i b_j wyraża się następująco:

$$L_j = \frac{(r-1)b_j}{2}, \tag{6}$$

$$b_j = \frac{6\left[2\sum_{f=j-r+1}^{j} w_f \alpha_f - (r+1)\sum_{f=j-r+1}^{j} \alpha_f\right]}{r\left(r^2-1\right)}, \tag{7}$$

gdzie: $w_f = w_{f-1} + 1$, $f = \left(j-r+2\right), ..., j$ oraz $w_{j-r+1} = 1$.

Dla przypadku B opisanego wzorem (2) sens nierówności (5) jest odwrotny, należy wybrać taką klasyfikację P_j, aby odpowiadający jej krok j $\left(j = r, r+1, ..., N-2\right)$ pierwszy spełniał nierówność:

$$\alpha_{j+1} < \overline{\alpha}_j - L_j - b_j - k \cdot s_j. \tag{8}$$

---

[5] Idea drugiej reguły polega na wyznaczaniu oczekiwanej wartości poziomu połączenia klas α na każdym *j*-tym etapie, na podstawie liniowej funkcji trendu dopasowywanej (korygowanej) względem pierwszych *j* – 1 wartości poziomu połączenia klas na wykresie drzewa.

[6] W regule drugiej również oblicza się średnią arytmetyczną i odchylenie standardowe wartości parametru α (poziomu połączenia klas) oznaczone odpowiednio $\overline{\alpha}_j$ i s_j, jednakże wyznacza się je sekwencyjnie po każdym kolejnym *j*-tym kroku procedury $\left(j = r, r+1, ..., N-2\right)$. Do wyznaczenia optymalnej liczby klas w regule pierwszej należy znać wszystkie $N-1$ wartości poziomu połączenia klas, zaś w regule drugiej wystarczy znać tylko *r* wartości poziomu połączenia klas $\left(r \le N-2\right)$ [Gatnar, Walesiak 2004].

Jeżeli żadna z wartości α_j $\left(j = r, r+1, ..., N-2\right)$ poziomu połączenia klas nie spełnia nierówności (5) lub (8), to procedura dalszego postępowania jest analogiczna jak w przypadku pierwszej reguły Mojeny – warianty (a), (b) i (c). Warto jednak zauważyć, że $\overline{\alpha}_j + L_j$ jest oczekiwaną wartością α_j, a $\overline{\alpha}_j + L_j + b_j$ jest oczekiwaną wartością α_{j+1}, stąd wartość znormalizowana obliczana w wariancie (b) równa się $\left(\alpha_{j+1} - \overline{\alpha}_j - L_j - b_j\right) / s_\alpha$.

4. Kryterium *tree validation*

Trzecie kryterium, którym warto zainteresować się w kontekście wyboru właściwej liczby skupień w przypadku metod hierarchicznych (aglomeracyjnych), opiera się na testowaniu braku struktury klas – ocenie losowości podziału obiektów na wykresie drzewa.

Idea metody *tree validation* (*bootstrap validation*) [Wishart 2006] polega na porównywaniu wyników ciągu klasyfikacji uzyskanych metodami aglomeracyjnymi, mających postać drzewa, z rodziną drzew generowanych na podstawie losowej permutacji tego samego zbioru danych albo skojarzonej macierzy sąsiedztwa. W tym celu wykorzystywane są techniki bootstrapowego, wielokrotnego losowania prób (bez zwracania), za pomocą których konstruowane są pojedyncze dendrogramy wyników grupowania [Domański i in. 1998]. Na każdym poziomie od 1 do N skupień, które są możliwe w przypadku grupowania N obiektów, oceniana jest jakość drzewa uzyskanego na podstawie analizowanych (oryginalnych) danych w stosunku do uśrednionej wersji drzewa wyznaczonej na podstawie prób losowych. Metoda oceny losowości podziału obiektów na wykresie drzewa stara się odrzucić hipotezę H_0 mówiącą o tym, że struktura grupowania obiektów w postaci danego drzewa jest losowa (brak struktury klas) na rzecz alternatywnej. Inaczej mówiąc, na podstawie analizowanego zbioru danych metoda *tree validation* poszukuje na wykresie drzewa, będącym ciągiem klasyfikacji, podziału, który jest „najbardziej oddalony" od podziału losowego.

5. Przykład empiryczny

Dla zilustrowania omówionych procedur wyboru właściwej liczby skupień (*upper tail rule*, *moving average quality control rule* i *tree validation*) programu *ClustanGraphics 8* dokonano klasyfikacji przykładowego zbioru danych Mammals.xls (pochodzącego z płyty CD *ClustanGraphics 8*) zawierającego skład mleka 25 ssaków opisany za pomocą 5 cech, tj. zawartości: wody, białka, tłuszczu, laktozy oraz składników mineralnych (ash).

W pierwszym kroku wyznaczono macierz kwadratowych odległości euklidesowych. Następnie, wskazując metodę aglomeracyjną – Warda, uzyskano wyjściowy

wykres drzewa przedstawiający ciąg klasyfikacji. Po wybraniu z menu polecenia *Tree/Best Cut*, a w dalszej kolejności *Upper tail* otrzymano wszystkie „istotne" podziały dla aktualnego drzewa według pierwszego kryterium Mojeny (2 lub 3 skupienia). Na podstawie ciągu wartości poziomu połączenia klas obliczany jest średni poziom połączenia klas, jego odchylenie standardowe oraz statystyka *t*-Studenta (standaryzowane odchylenie poziomu połączeń klas od średniej). Z kolejnych poziomów połączenia klas wybierane jest pierwsze „istotne" na poziomie 5%, które jest automatycznie „przenoszone" na wykres drzewa.

Wybierając z menu polecenie *Tree/Best Cut*, a w dalszej kolejności *Moving average* otrzymano „istotne" podziały dla aktualnego drzewa według drugiego kryterium Mojeny, które w tym przypadku wskazało inny podział mleka ssaków – 9 skupień.

W ostatnim kroku, wybierając z menu polecenie *Tree/Validate* oraz decydując, czy drzewa mają być generowane na podstawie losowej permutacji zbioru danych czy macierzy sąsiedztwa, otrzymano wynik oceny losowości podziału obiektów na wykresie drzewa. Metoda *tree validation* wskazała, że „najbardziej oddalony" od podziału losowego jest podział mleka ssaków na 3 skupienia, jednorodne pod względem zawartości wody, białka, tłuszczu, laktozy oraz składników mineralnych (ash).

6. Podsumowanie

W artykule omówiono kryteria Mojeny i Wisharta wyboru właściwej liczby skupień dla metod hierarchicznych (aglomeracyjnych), a ich działanie zilustrowane zostało przykładem z programu *ClustanGraphics 8*. Należy zauważyć, że w załączonym przykładzie dwa spośród trzech kryteriów wskazały ten sam wynik grupowania – 3 skupienia, który wydaje się tym właściwym. Celem kolejnego artykułu z tego zakresu będzie empiryczna analiza i porównanie trzech omówionych kryteriów na tle innych powszechnie stosowanych miar wyboru optymalnej liczby skupień.

Literatura

Balicki A., *Statystyczna analiza wielowymiarowa i jej zastosowania społeczno-ekonomiczne*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 2009.

Cormack R., *A review of classification*, „Journal of the Royal Statistical Society", Series A 1971, vol. 134(3).

Denoeud L., Garreta H., Guénoche A., *Comparison of distance indices between partition*, Conference „International Symposium on Applied Stochastic Models and Data Analysis" 2005 (ASMDA2005), http://conferences.telecom-bretagne.eu/asmda2005.

Domański Cz., Pruska K., Wagner W., *Wnioskowanie statystyczne przy nieklasycznych założeniach*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 1998.

Fowlkes E.B., Mallows C.L., *A method for comparing two hierarchical clusterings*, „Journal of the American Statistical Association" 1983, vol. 78(383).

Gan G., Ma C., Wu J., *Data clustering: theory, algorithms, and applications*, SIAM, Philadelphia 2007.

Gatnar E., Walesiak M. (red.), Metody *statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo AE, Wrocław 2004.
Gatnar E., Walesiak M. (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo PWN, Warszawa 2009.
Gordon A.D., *A Review of hierarchical classification*, „Journal of the Royal Statistical Society", Series A 1987, vol. 150(2).
Gordon A.D., *Hierarchical Classification*, [w:] Arabie P., Hubert L.J., De Soete G. (red.), *Clustering and Classification*, World Scientific, Singapore 1996.
Kaufman L., Rousseeuw P.J., *Finding Groups in Data. An Introduction to Cluster Analysis*, (reprint 2005), John Wiley & Sons, New York 2005.
Migdał-Najman K., Najman K., *Wykorzystanie wskaźnika Dunna do ustalania optymalnej liczby skupień*, „Wiadomości Statystyczne" 2008, nr 11.
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).
Mojena R., *Hierarchical grouping methods and stopping rules: an evaluation*, „Computer Journal" 1977, vol. 20(4).
Nowak E., *Wskaźnik podobieństwa wyników podziałów*, „Przegląd Statystyczny" 1985, nr 1.
Sugar C.A., James G.M., *Finding the number of clusters in a data set: an information-theoretic approach*, „Journal of the American Statistical Association" 2003, vol. 98(463).
Tibshirani R., Walther G., Hastie T., *Estimating the number of clusters in a data set via the gap statistic*, „Journal of the Royal Statistical Society" 2001, Series B, vol. 63(2).
Wishart D., *Clustangraphics Primer: a Guide to Cluster Analysis*, (4th edition), Edinburgh 2006.

EVALUATION METHODS FOR THE GROUPING RESULT IN CLUSTER ANALYSIS

Summary: The paper reviews evaluation methods for the grouping result. It discusses three methods how to choose the correct number of clusters for agglomeration methods proposed by Mojena and Wishart, which are implemented in *ClustanGraphics 8*. The criteria which are mentioned are based on the relative sizes of the different fusion levels of the objects in the dendrogram – best cut significance test (*upper tail, moving average*) and checking randomness of objects clustering in the dendrogram – *validation tree*. The text is followed by an empirical example.

Keywords: cluster analysis, evaluation of the grouping result, Mojena criteria, Wishart criterion.