PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 508 • 2018

Taksonomia 31
Klasyfikacja i analiza danych – teoria i zastosowania

ISSN 1899-3192
e-ISSN 2392-0041

Dorota Rozmus
Uniwersytet Ekonomiczny w Katowicach
e-mail: dorota.rozmus@ue.katowice.pl

WPŁYW REDUKCJI LICZBY ZMIENNYCH NA STABILNOŚĆ GRUPOWANIA

THE EFFECT OF REDUCTION OF VARIABLES TO GROUPS STABILITY

DOI: 10.15611/pn.2018.508.18
JEL Classification: C38

Streszczenie: W artykule zbadany został wpływ formalnych metod doboru zmiennych na stabilność grupowania. Kryterium stabilności bada, czy grupy, które zostały utworzone w wyniku grupowania zbioru obiektów, występują rzeczywiście (zatem struktura jest stabilna), czy też pojawiły się przypadkowo i uzyskana struktura nie odzwierciedla tej istniejącej w danych. Jako formalną metodę doboru zmiennych zastosowano analizę czynnikową, natomiast badanie stabilności grupowania przeprowadzono za pomocą metody w przybliżeniu nieobciążonego prawdopodobieństwa bootstrapowego na zbiorach danych społeczno-ekonomicznych utworzonych na podstawie danych zaczerpniętych z Głównego Urzędu Statystycznego. Uzyskane wyniki pokazują, że zastosowanie analizy czynnikowej do redukcji liczby zmiennych może wpływać zarówno na stabilność grupowania, jak i na uzyskiwaną strukturę grup.

Słowa kluczowe: grupowanie obiektów, liczba grup, stabilność grupowania, redukcja liczby zmiennych, analiza czynnikowa.

Summary: The paper examines the effect of formal methods of variables selection on groups stability. The stability criterion examines whether the groups that were formed as a result of using taxonomy methods actually exist in the data (the structure is stable) or they have come up by chance. As a formal method of selecting variables, factor analysis was used, while groups stability testing was performed using a method of approximately unbiased bootstrap probability on socio-economic data sets generated on the basis of data taken from Central Statistical Office. The results show that using factor analysis for reducing the number of variables can affect both the stability of grouping and the obtained groups structure.

Keywords: cluster analysis, number of groups, grouping stability, variables selection, factor analysis.

1. Wstęp

Jednym z najważniejszych zagadnień w taksonomii jest wybór zmiennych, na podstawie których dokonane zostanie grupowanie obiektów. Od jakości zestawu zmiennych zależą wyniki klasyfikacji, a w dalszej konsekwencji – trafność podejmowanych na ich podstawie decyzji. Konstruując zbiór danych, powinno się w nim uwzględniać tylko takie zmienne, które posiadają zdolność dyskryminacji obiektów. Nie należy natomiast stosować podejścia, które polega na uwzględnianiu wszystkich możliwych zmiennych, ponieważ, jak pokazał Milligan [1994], uwzględnianie zmiennych nieistotnych może uniemożliwić odkrycie w zbiorze obiektów właściwej struktury klas.

Do rozwiązania zagadnienia doboru zmiennych służą zasadniczo dwa ujęcia: dobór ściśle merytoryczny oraz dobór merytoryczno-formalny. W artykule uwaga zostanie skupiona na tym drugim podejściu, gdzie w pierwszej fazie na podstawie własnej hipotezy roboczej badacza bądź też współpracy z ekspertami, konstruowana jest wstępna lista zmiennych. Faza druga w tym podejściu polega na zastosowaniu formalnych algorytmów doboru zmiennych. Zastosowana może być np. analiza czynnikowa, która zastępuje oryginalne zmienne mniejszą liczbą „sztucznych" czynników o pożądanych właściwościach. Ważną zaletą analizy czynnikowej jest to, że pozwala ona na dobór takiego zestawu zmiennych, które są wzajemnie niezależne i jednocześnie zależne od zmiennych, które nie weszły do wybranego zestawu (postulat niepowielania informacji).

Jak pokazano w pracy [Milligan 1996], zastosowanie analizy czynnikowej do redukcji pierwotnej wielowymiarowej przestrzeni klasyfikacji może jednak spowodować utratę struktury klas z pierwotnej przestrzeni.

Z problemem struktury klas silnie łączy się pojęcie stabilności grupowania. Stabilność struktur zakłada, że przy poprawnie dobranych parametrach algorytmu (np. kryterium stopu, parametry sterujące algorytmu, liczba grup) wielokrotne grupowanie obiektów powinno dawać wyniki podziału niewiele różniące się od siebie. Shamir i Tishby [2008] stwierdzają, że jeżeli algorytm taksonomiczny jest wielokrotnie stosowany do niezależnych prób (przy niezmienionych parametrach algorytmu), dając w efekcie podobne wyniki grupowania, to można je uznać za stabilne i odzwierciedlające rzeczywistą strukturę grup. Volkovich i in. [2010] stwierdzają, że liczba grup, która maksymalizuje stabilność struktur, może służyć jako odpowiedź na pytanie, na ile grup należy dokonać podziału.

Ponieważ kryterium stabilności grupowania znajduje zastosowanie zwłaszcza przy ustalaniu jednego z najważniejszych parametrów metod taksonomicznych, tj. liczby grup (k), w artykule zbadany zostanie wpływ formalnych metod doboru zmiennych na stabilność grupowania. Do redukcji liczby zmiennych zastosowana zostanie analiza czynnikowa, natomiast stabilność grupowania badana będzie za pomocą metody w przybliżeniu nieobciążonego prawdopodobieństwa testowego (*approximately unbiased p-value* – AU) zaproponowaną przez Suzuki i Shimodaira

[2004]. Obliczenia przeprowadzone zostaną w programie **R** na zbiorach danych społeczno-ekonomicznych pochodzących z GUS.

2. Metody badawcze

2.1. Analiza czynnikowa

Analiza czynnikowa stanowi zespół metod i procedur statystycznych pozwalających na badanie wzajemnych relacji między dużą liczbą zmiennych oraz wykrywanie ukrytych uwarunkowań, które wyjaśniają ich występowanie. Umożliwia ona sprowadzenie dużej liczby badanych zmiennych do mniej licznego zbioru wzajemnie niezależnych (nieskorelowanych) czynników[1].

Przed przystąpieniem do analizy czynnikowej należy sprawdzić, czy zmienne zawarte w zbiorze danych są wystarczająco ze sobą skorelowane. Jeżeli są słabo skorelowane, to jest mało prawdopodobne, że utworzą silne i łatwe w interpretacji czynniki. Do badania stopnia skorelowania zmiennych można zastosować wskaźnik Kaisera-Mayera-Olkina (*KMO*) [Zakrzewska 1994]:

$$KMO = \frac{\sum_j \sum_{h \neq j} r_{jh}^2}{\sum_j \sum_{h \neq j} r_{jh}^2 + \sum_j \sum_{h \neq j} \hat{r}_{jh}^2},$$

gdzie: r_{jh} to współczynnik korelacji między zmiennymi o numerach j i h, $\hat{r}_{jh}$ – współczynnik korelacji cząstkowej między nimi.

Wartości *KMO* niższe od 0,7 sugerują potrzebę usunięcia części zmiennych.

Można także wyliczyć miarę adekwatności doboru każdej indywidualnej zmiennej (MSA_h) [Zakrzewska 1994]:

$$MSA_h = \frac{\sum_{j \neq h} r_{jh}^2}{\sum_{j \neq h} r_{jh}^2 + \sum_{j \neq h} \hat{r}_{jh}^2}.$$

Usuwamy te zmienne, dla których wartość MSA_h jest niska.

2.2. Stabilność grupowania

Do badania stabilności grupowania zostanie zastosowana metoda w przybliżeniu nieobciążonego prawdopodobieństwa testowego, którą można znaleźć w pakiecie `pvclust` [Suzuki, Shimodaira 2006] w programie **R**.

---

[1] Dokładny opis metody można znaleźć m.in. w pracach: [Kim, Müller 1978a; 1978b; Walesiak, Gatnar (red.) 2009].

W pakiecie tym stabilność struktur mierzy się przez prawdopodobieństwo testowe (*p-value*) liczone dla każdej grupy, wykorzystując do tego losowanie bootstrapowe. Dostępne są dwa rodzaje prawdopodobieństwa:

- prawdopodobieństwo bootstrapowe (*bootstrap probability value* – BP) [Efron 1979; Felsenstein 1985];
- w przybliżeniu nieobciążone prawdopodobieństwo testowe (*approximately unbiased p-value* – AU) [Shimodaira 2002; 2004].

Do wyliczenia w przybliżeniu nieobciążonego prawdopodobieństwa testowego wykorzystuje się wieloskalowe losowanie bootstrapowe (*multiscale bootstrap*).

2.2.1. Prawdopodobieństwo bootstrapowe

Schemat wyliczania prawdopodobieństwa bootstrapowego można przedstawić następująco:

1. Utwórz próby bootstrapowe.

2. Do każdej z nich zastosuj hierarchiczną metodę grupowania, uzyskując tzw. bootstrapowe replikacje dendrogramów.

3. Wśród wszystkich bootstrapowych replikacji dendrogramów oblicz odsetek tych dendrogramów, które zawierają grupę hipotetyczną.

Powyższe postępowanie prowadzi do uzyskania tzw. prawdopodobieństwa bootstrapowego, które stosowane jest do określenia prawdopodobieństwa wystąpienia danej grupy.

Jednakże tak liczone prawdopodobieństwo testowe jest obciążone [Hillis, Bull 1993; Zharkikh, Li 1992; Sanderson, Wojciechowski 2000]. W związku z tym zaproponowano sposoby korekty tego obciążenia. Jedną z nich jest metoda wieloskalowego losowania bootstrapowego, które daje w efekcie w przybliżeniu nieobciążone prawdopodobieństwo bootstrapowe.

2.2.2. W przybliżeniu nieobciążone prawdopodobieństwo bootstrapowe

Przy wyliczaniu prawdopodobieństwa bootstrapowego liczebność podprób bootstrapowych jest taka sama jak liczebność pierwotnego zbioru danych, w metodzie Shimodaira natomiast rozmiar podprób ulega zmianie. Ma to na celu wprowadzenie korekty obciążenia bootstrapowego prawdopodobieństwa testowego na podstawie wariacji wyników dla różnych rozmiarów próby.

Schemat wyliczania w przybliżeniu nieobciążonego prawdopodobieństwa bootstrapowego można przedstawić następująco:

1. Utwórz próby bootstrapowe dla założonych wartości liczebności próby (wielkoskalowe losowanie bootstrapowe).

2. Do każdej z nich zastosuj hierarchiczną metodę grupowania, uzyskując tzw. bootstrapowe replikacje dendrogramów.

3. Określ prawdopodobieństwa bootstrapowe dla każdego rozmiaru próby.

4. Na ich podstawie określ wartość w przybliżeniu nieobciążonego prawdopodobieństwa testowego, korzystając z równania:

$$AU = 1 - \Phi(d - c),$$

gdzie: c i d wyznaczane są przez dopasowanie teoretycznych wartości $BP(\tau) = 1 - \Phi(d/\tau + c\tau)$ do zaobserwowanych wartości $BP(\tau)$ uzyskanych na podstawie wieloskalowego prawdopodobieństwa bootstrapowego ($\tau = \sqrt{n/n'}$, gdzie n' to liczba obserwacji w podpróbach bootstrapowych w wieloskalowym losowaniu bootstrapowym).

2.2.3. Pakiet `pvclust`

Poniżej zamieszczona została funkcja programu **R** z najważniejszymi parametrami i ich wartościami, które zostały zastosowane w badaniu:

```
stabilnosc <- pvclust(data = dane, method.hclust = "ward",
method.dist = "correlation", nboot = 1000, r = seq(.5,
1.5, by = .1))
```

Jako metodę grupowania wybrano metodę Warda, parametr `method.dist = „correlation"` informuje, że odległość między obiektami policzona została za pomocą współczynnika korelacji. Liczba prób bootstrapowych równa była 1000 (`nboot = 1000`), natomiast parametr `r = seq(.5, 1.5, by = .1)` informuje, że w wieloskalowym losowaniu bootstrapowym liczebność zbiorów równa była od 50 do 150% liczebności pierwotnego zbioru danych i za każdym razem zwiększana była o 10 p.p.

3. Zastosowane zbiory danych

W badaniu zastosowane zostały dwa zbiory danych społeczno-ekonomicznych pochodzące z GUS. Pierwszy z nich powstał na postawie aplikacji STRATEG i znajdują się w nim zmienne, które mają na celu monitorowanie realizacji polityki spójności, m.in. w aspekcie inteligentnego rozwoju, w województwach Polski. Dane dotyczące inteligentnego rozwoju podzielone są przez GUS na pięć grup tematycznych:

- jakość edukacji, umiejętności i uczenia się przez całe życie,
- jakość i dostępność technologii informacyjno-komunikacyjnych (ICT),
- podnoszenie konkurencyjności małych i średnich przedsiębiorstw, sektora rolnego oraz sektora rybołówstwa i akwakultury,
- wspieranie badań naukowych, rozwoju technologicznego i innowacji,
- zrównoważony transport i wysoka przepustowość kluczowych sieci infrastrukturalnych.

Zastosowane w badaniu dane pochodzą z 2015 roku i obejmują 214 zmiennych z kompletnymi danymi.

Drugi zbiór danych utworzony został na podstawie aplikacji Wskaźniki Zrównoważonego Rozwoju, która monitoruje realizację polityki zrównoważonego rozwoju w państwach UE. Dane te podzielone są przez GUS na cztery grupy, monitorujące realizację polityki zrównoważonego rozwoju w ramach następujących ładów:

- społecznego,
- gospodarczego,
- środowiskowego,
- instytucjonalno-politycznego.

W badaniu wykorzystano dane z 2015 roku, które obejmują 51 zmiennych z kompletnymi danymi.

4. Wyniki badań empirycznych

4.1. Wyniki dla zbioru dotyczącego inteligentnego rozwoju

Z uwagi na to, że w zbiorze tym znajdują się zmienne pogrupowane przez GUS na pięć obszarów, w pierwszym kroku zbadano wartości wskaźnika Kaisera-Mayera-Olkina dla każdego obszaru osobno[2]. Wartości te zawarte są w tab. 1.

Tabela 1. Wartości wskaźnika KMO dla poszczególnych obszarów inteligentnego rozwoju

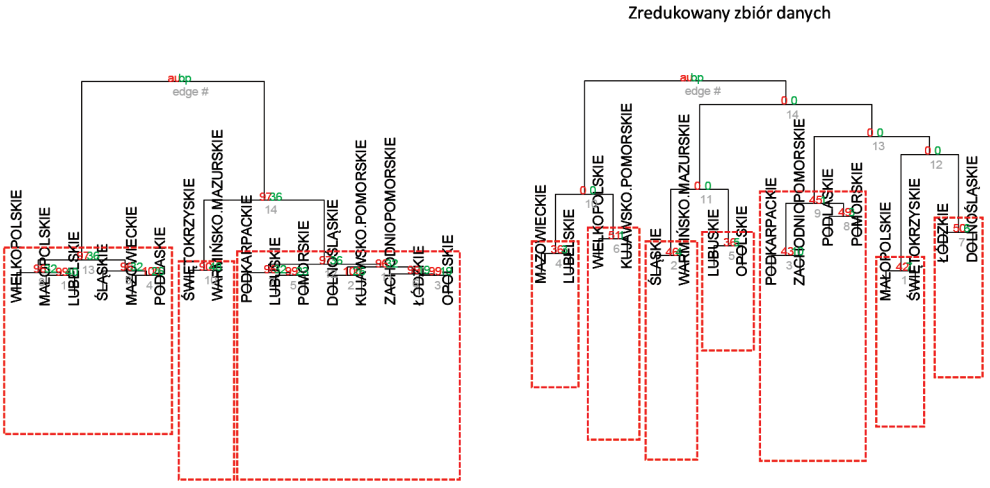
Obszar inteligentnego rozwoju	Wartość miary KMO
Jakość edukacji	0,6587
Dostępność ICT	0,4596
Konkurencyjność przedsiębiorstw	0,6405
Badania naukowe i innowacje	0,5832
Transport	0,5729

Źródło: obliczenia własne.

Wartości wskaźnika Kaisera-Mayera-Olkina dla każdego obszaru są niższe niż 0,7, dlatego w następnym kroku, sugerując się wartościami miary adekwatności doboru każdej indywidualnej zmiennej (MSA_h) w każdym obszarze z osobna, usunięto te zmienne, dla których $MSA_h < 0{,}5$. Po dokonanej w ten sposób redukcji ostatecznie w zbiorze danych (obejmującym wszystkie obszary) pozostały 164 zmienne. Analiza czynnikowa przeprowadzona na tak zredukowanym wstępnie zbiorze danych pozwoliła wyodrębnić 15 czynników, które wyjaśniały 98,7% wariancji.

[2] Przeprowadzenie wstępnej redukcji zmiennych (za pomocą miary *KMO* i MSA_h) w obrębie każdego obszaru z osobna ma na celu zapewnienie, że zredukowany zbiór zmiennych również będzie reprezentował wszystkie te obszary.

W kolejnym kroku zbadano stabilność grupowania województw na podstawie pierwotnego i zredukowanego do 15 czynników zbioru danych. Uzyskane rezultaty pokazane zostały na rys. 1.



Rys. 1. Wyniki grupowania i badania stabilności dla zbioru danych dotyczącego inteligentnego rozwoju

Źródło: opracowanie własne.

W przypadku grupowania województw na podstawie pierwotnego zbioru danych uzyskany dendrogram i wartości w przybliżeniu nieobciążonego prawdopodobieństwa testowego (AU) sugerują utworzenie trzech grup. Pierwsza, obejmująca województwa: wielkopolskie, małopolskie, lubelskie, śląskie, mazowieckie i podlaskie, charakteryzuje się wysoką wartością miernika stabilności równą 0,97. Dla drugiej grupy województw, obejmującej świętokrzyskie oraz warmińsko-mazurskie, w przybliżeniu nieobciążone prawdopodobieństwo testowe przyjmuje wartość 0,90. Ostatnia grupa, w którym znalazły się pozostałe województwa, również charakteryzuje się wysoką wartością miernika stabilności wynoszącą 0,97.

Grupowanie województw na podstawie zredukowanego zbioru danych (czyli na podstawie uzyskanych czynników) ma zupełnie inną strukturę: powstało siedem grup obiektów, w tym aż sześć grup obejmuje po zaledwie dwa województwa; również wartość w przybliżeniu nieobciążonego prawdopodobieństwa testowego uległa znacznemu obniżeniu dla większości grup, np. dla grupy obejmującej województwa łódzkie i dolnośląskie wynosi zaledwie 0,50.

Wyraźnie zatem widać na tym przykładzie, że redukcja liczby zmiennych doprowadziła do zmiany struktury grup, a także do obniżenia się stabilności grupowania.

4.2. Wyniki dla zbioru dotyczącego zrównoważonego rozwoju

Do badania zrównoważonego rozwoju GUS zaproponował podział zmiennych na cztery obszary, zwane ładami. Dlatego, podobnie jak w powyżej opisywanym przypadku, najpierw policzone zostaną wartości wskaźnika Kaisera-Mayera-Olkina dla każdego ładu z osobna; ich wartości zestawione są w tab. 2. Następnie, w razie potrzeby, z każdego podzbioru zostaną usunięte zmienne o niskich wartościach miar adekwatności doboru każdej indywidualnej zmiennej (MSA_h)[3].

Tabela 2. Wartości wskaźnika KMO dla poszczególnych obszarów zrównoważonego rozwoju

Obszar zrównoważonego rozwoju	Wartość miary KMO
Ład gospodarczy	0,5610
Ład społeczny	0,5388
Ład środowiskowy	0,5418
Ład instytucjonalno-polityczny	0,8075

Źródło: obliczenia własne.

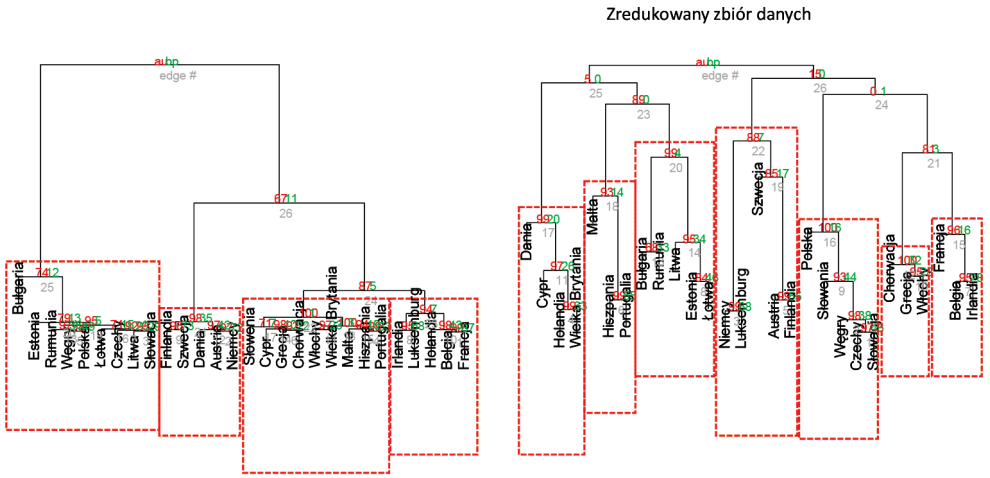
Na podstawie danych zawartych w tab. 2 widać, że jedynie ład instytucjonalno-polityczny ma wartość wskaźnika Kaisera-Mayera-Olkina powyżej 0,7. Z pozostałych obszarów konieczne będzie usunięcie zmiennych. Kierując się w tym celu wartościami miar adekwatności doboru każdej indywidualnej zmiennej (MSA_h), usunięto z każdego obszaru te zmienne, dla których $MSA_h < 0,5$. Ostatecznie całościowy zbiór danych, który będzie poddany redukcji za pomocą analizy czynnikowej (zawierający zmienne ze wszystkich ładów), zawiera 32 zmienne. Przeprowadzenie analizy czynnikowej na tym zbiorze doprowadziło do utworzenia 7 czynników, które wyjaśniły 85,7% wariancji.

Wyniki grupowania państw Unii Europejskiej pod względem realizacji polityki zrównoważonego rozwoju oraz wartości miar stabilności pokazuje rys. 2.

Dla oryginalnego zbioru danych dendrogram i wartości w przybliżeniu nieobciążonego prawdopodobieństwa testowego sugerują istnienie czterech grup; ponadto uzyskane wartości miernika stabilności (AU) w większości pokazują wysoką stabilność uzyskanego grupowania. Pierwsza grupa państw, obejmująca Bułgarię, Estonię, Rumunię, Węgry, Polskę, Łotwę, Czechy, Litwę i Słowację, charakteryzuje się najniższą stabilnością równą 0,74. Druga grupa, dla której AU = 0,98, obejmuje: Finlandię, Szwecję, Danię, Austrię oraz Niemcy. Trzecia, z miernikiem stabilności na poziomie 1, obejmuje: Słowenię, Cypr, Grecję, Chorwację, Włochy, Wielką Brytanię, Maltę, Hiszpanię i Portugalię. I ostatnia grupa państw, obejmująca Irlandię,

[3] Podobnie jak w przypadku zbioru dotyczącego inteligentnego rozwoju, przeprowadzenie wstępnej redukcji zmiennych (za pomocą miary *KMO* i MSA_h) w obrębie każdego ładu z osobna ma na celu zapewnienie, że zredukowany zbiór danych również będzie reprezentował wszystkie te łady.

Luksemburg, Holandię, Belgię oraz Francję, to grupa, dla której miara stabilności równa jest 0,94.



Rys. 2. Wyniki grupowania i badania stabilności dla zbioru danych dotyczącego zrównoważonego rozwoju

Źródło: opracowanie własne.

Natomiast grupowanie państw Unii Europejskiej na podstawie zredukowanego zbioru danych wskazuje na istnienie siedmiu grup (zatem inna jest struktura w porównaniu z grupowaniem obiektów z pierwotnego zbioru danych) przy jednoczesnym podtrzymaniu dosyć wysokich wartości miernika stabilności. Najniższą wartość odnotowujemy dla zgrupowania obejmującego Niemcy, Luksemburg, Szwecję, Austrię i Finlandię, gdzie AU = 0,88. Jednocześnie widać także, że w przypadku dwóch skupień miara stabilności przyjmuje wartość 1 (Polska, Słowenia, Węgry, Czechy Słowacja oraz Chorwacja, Grecja, Włochy), a dla dwóch – 0,99 (Dania, Cypr, Holandia, Wielka Brytania oraz Bułgaria, Rumunia, Litwa, Estonia, Łotwa).

5. Podsumowanie

Celem artykułu było zbadanie wpływu redukcji liczby zmiennych za pomocą analizy czynnikowej na stabilność grupowania. Bazując na uzyskanych wynikach, można stwierdzić, że zastosowanie analizy czynnikowej do redukcji liczby zmiennych może spowodować obniżenie się stabilności grupowania.

Warto także zwrócić uwagę na wyniki badań prowadzonych przez Yeung i Ruzzo [2001], którzy pokazali wpływ redukcji danych za pomocą analizy głównych składowych na jakość grupowania, gdzie jakość grupowania mierzona była za pomocą skorygowanego indeksu Randa. Z rezultatów ich badań wynika, że stosowanie

kilku pierwszych składowych o największych wartościach własnych rzeczywiście może powodować obniżenie się jakości grupowania, i sugerują, że możliwe jest uzyskanie poprawy jakości, gdy uwzględniona będzie odpowiednia liczba składowych. Pytanie, jak dotąd niestety bez odpowiedzi, brzmi: ile składowych należy uwzględnić? Próba odpowiedzi na to pytanie będzie elementem dalszych badań.

Literatura

Efron B., 1979, *Bootstrap methods: Another look at the jackknife*, Annals of Statistics, vol. 7, s. 1-26.

Felsenstein J., 1985, *Confidence limits on phylogenies: An approach using the bootstrap*, Evolution, 39, s. 783-791.

Hennig C., 2007, *Cluster-wise Assessment of Cluster Stability*, Computational Statistics and Data Analysis, vol. 52, s. 258-271.

Hillis D., Bull J., 1993, *An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis*, Systematic Biology, vol. 42, s. 182-192.

Kim J.O., Müller C.W., 1978a, *Factor Analysis, Statistical Methods and Practical Issues*, Sage, Beverly Hills.

Kim J.O., Müller C.W., 1978b, *Introduction to Factor Analysis. What it is and How to do it*, Sage, Beverly Hills.

Milligan G.W., 1994, *Issues in applied classification: Selection of variables to cluster*, Classification Society of North America Newsletters, November, Issue 37.

Milligan G.W., 1996, *Clustering validation: results and implications for applied analyses*, [w:] Arabie P., Hubert L.J., de Soete G. (red.), *Clustering and Classification*, World Scientific, Singapore.

Sanderson M.J., Wojciechowski M.F., 2000, *Improved bootstrap confidence limits in large-scale phylogenies, with an example from Neo-Astragalus (Leguminosae)*, Systematic Biology, vol. 49, s. 671-685.

Shamir O., Tishby N., 2008, *Cluster stability for finite samples*, Advances in Neural Information Processing Systems, vol. 20, s. 1297-1304.

Shimodaira H., 2002, *An approximately unbiased test of phylogenetic tree selection*, Systematic Biology, vol. 51, s. 492-508.

Shimodaira H., 2004, *Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling*, Annals of Statistics, vol. 32, s. 2616-2641.

Suzuki R., Shimodaira H., 2004, *An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: how accurate are these clusters?*, Proceedings by the Fifteenth International Conference on Genome Informatics (GIW 2004).

Suzuki R., Shimodaira H., 2006, *Pvclust: An R package for assessing the uncertainty in hierarchical clustering*, Bioinformatics, vol. 22, no.12, s. 1540-1542.

Volkovich Z., Barzily Z., Toledano-Kitai D., Avros R., 2010, *The Hotteling's metric as a cluster stability measure*, Computer Modelling and New Technologies, vol. 14, no. 4, s. 65-72.

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

Yeung K.Y., Ruzzo W.L., 2001, *An empirical study on principal component analysis for clustering gene expression data*, Bioinformatics, vol. 17(9), s. 763-774.

Zakrzewska M., 1994, *Analiza czynnikowa w budowaniu i sprawdzaniu modeli psychologicznych*, Wydawnictwo UAM, Poznań.

Zharkikh A., Li W.H., 1992, *Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock*, Molecular Biology and Evolution, vol. 9, s. 1119-1147.