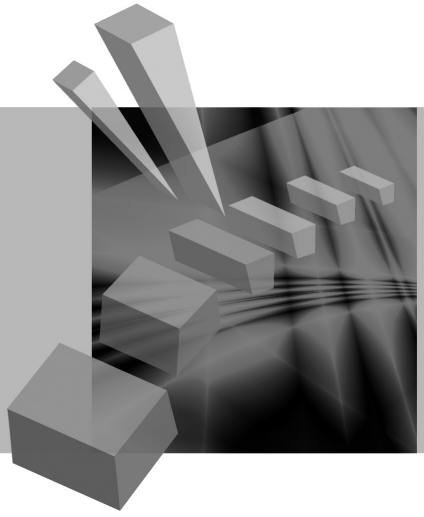
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

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

Jerzy Korzeniewski
Uniwersytet Łódzki

MODYFIKACJA METODY HINOV SELEKCJI ZMIENNYCH W ANALIZIE SKUPIEŃ

Streszczenie: Metoda HINoV selekcji zmiennych w analizie skupień jest znana od roku 1999. Oryginalna metoda HINoV jest zupełnie nieodporna na występowanie wśród zmiennych zanieczyszczających strukturę skupień zmiennych skorelowanych jednomodalnych lub równomiernych. Wadę tę próbowano wyeliminować w modyfikacji VS-KM. Częściowo się to udało. W artykule zaproponowana jest prosta modyfikacja polegająca na tym, by dla każdej liczby skupień rozważanej w procedurze replikacji zbadać stabilność podziału zbioru dla obu porównywanych podzbiorów zmiennych (rozłącznych), z których jeden trzeba wybrać. Modyfikacja oceniona jest w obszernym eksperymencie symulacyjnym na 8100 zbiorach danych ze strukturami skupień wygenerowanymi w postaci mieszanin rozkładów normalnych.

Słowa kluczowe: analiza skupień, wybór zmiennych, metoda HINoV, metoda VS-KM.

1. Wstęp

W analizie skupień bardzo ważny jest wybór zmiennych istotnych dla ewentualnej struktury skupień, która może istnieć w zbiorze danych. Poprawny wybór zmiennych warunkuje wnioski z całej analizy skupień. Do chwili obecnej zaproponowano kilkanaście podejść do zagadnienia selekcji zmiennych. Metody te charakteryzują się różnymi cechami. Niektóre uzależniają wybór od jakiejś metody grupowania obserwacji, inne nie, niektóre są skonstruowane z myślą o jednej skali pomiarowej, inne są bardziej ogólne, niektóre mają charakter modelowy, tj. taki, w którym zakładamy, że zbiór danych jest mieszaniną obserwacji z rozkładów normalnych, inne są podejściami czysto heurystycznymi.

W roku 2008 Steinley i Brusco przeprowadzili obszerny eksperyment symulacyjny, w którym dokonali porównania ośmiu metod służących do wybierania zmiennych w analizie skupień. Wybrali oni kilka podejść modelowych: metoda wyrazistości cech (*feature saliency method* [Law i in. 2003]); metoda oparta na wielkości rozproszenia (*scatter separability method* [Dy, Brodley 2000]); metoda oparta na wyborze właściwego modelu (*model selection method* [Raftery, Dean 2006]) oraz kilka niemodelowych: metoda HINoV [Carmone i in. 1999]; metoda oparta na grupowaniu obiektów na wybranych podzbiorach zmiennych (COSA, [Friedman,

Meulman 2004]); metoda kolejnych rzutowań (*projection pursuit* [Montanari, Lizzani 2001]); metoda oparta na grupowaniu *k*-średnich (*VS-KM method* [Brusco, Cradit 2001]); metoda oparta na grupowaniu *k*-średnich z indeksem odnoszącym wariancję zmiennej do jej rozstępu (*VAF*, *relative clusterability weighting method* [Steinley, Brusco 2007]). Ocena efektywności oparta była na trzech głównych miarach: **pamięci** (*recall)*, **precyzji** (*precision*) oraz **asymptotycznej odzyskiwalności** poprawnego przypisania obserwacji do skupień (*ARI, asymptotic recovery*) (por. [Steinley, Brusco 2007]). Wskaźniki te definiuje się dla potrzeb eksperymentu symulacyjnego, w którym dla każdego zbioru znany jest zbiór zmiennych istotnych dla struktury skupień oraz zbiór zmiennych nieistotnych. Wówczas przez pamięć rozumiemy stosunek liczby wybranych zmiennych istotnych do liczby wszystkich zmiennych istotnych, precyzja to stosunek liczby wybranych zmiennych istotnych do liczby wszystkich wybranych zmiennych, zaś asymptotyczna odzyskiwalność rozumiana jest w sensie średniej arytmetycznej (ze wszystkich zbiorów) wartości skorygowanego indeksu Randa mierzącego zgodność podziału, opartego na wybranym podzbiorze zmiennych z podziałem wynikającym ze sposobu generowania zbioru danych. Wartość indeksu Randa obliczana jest dla podziału zbioru obiektów otrzymanego w następujący sposób: za pomocą metody *k*-średnich z losowym wyborem obserwacji startowych powtarzamy 100 razy grupowanie i zapamiętujemy grupowanie o najmniejszej wariancji wewnątrzgrupowej.

Z badania tego wynika, że jedną z najlepszych metod okazał się HINoV. Ta metoda jest jednak nieodporna na ewentualne skorelowanie zmiennych, które nie tworzą żadnej struktury skupień (por. tab. 1). Nieco lepsze wyniki uzyskała modyfikacja VS-KM opracowana z myślą poprawienia efektywności w takich przypadkach. Badanie Steinleya i Brusco ma jednak wiele mankamentów. Najważniejsze z nich to założenie o znajomości poprawnej liczby skupień oraz arbitralnie ustalany, jednolity zbiór zmiennych maskujących struktury skupień. W przypadku analizowania empirycznych zbiorów danych typy zbiorów zmiennych maskujących mało realistycznie oddają problemy, przed którymi staje statystyk. Imitując świat realny, należałoby raczej położyć nacisk na istnienie różnych rodzajów zmiennych, których zadaniem jest maskowanie struktur skupień.

Celem niniejszego artykułu jest zaproponowanie modyfikacji metody HINoV, która byłaby bardziej efektywna od HINoV, zwłaszcza w przypadku, gdy zmienne zanieczyszczające strukturę skupień są skorelowane. W dalszym ciągu artykułu podane są miary efektywności metod HINoV oraz VS-KM, jakie uzyskały one w eksperymencie podobnym do eksperymentu Steinleya i Brusco, charakterystyka proponowanej modyfikacji metody HINoV oraz wyniki, jakie uzyskała ona i metody konkurencyjne w drugim eksperymencie z inaczej dobranymi zborami zmiennych maskujących struktury skupień.

2. Efektywność metod HINoV i VS-KM

Metodę HINoV dla potrzeb eksperymentu określamy w następujący sposób, analogicznie do sformułowania z pracy Steinleya i Brusco [2008]. Dla każdej zmiennej v przeprowadzamy 50 razy grupowanie metodą k-średnich i zapamiętujemy to grupowanie, które miało najmniejszą sumę kwadratów odchyleń obserwacji od środków skupień, tzn.

$$SSE(v) = \sum_{k=1}^{K} \sum_{i \in C_k} \left(x_{iv} - \overline{x}_{iv} \right)^2, \qquad \text{gdzie} \qquad \overline{x}_{iv} = \frac{1}{N_k} \sum_{i \in C_k} x_{iv} . \tag{1}$$

Następnie obliczamy poprawiony indeks Randa $ARI(u,v)$ (por. np. [Gatnar, Walesiak (red.) 2004]). Wartość indeksu $ARI(u,v)$ interpretuje się jako miarę podobieństwa dwóch podziałów/grupowań tego samego zbioru obserwacji. Następnie dla każdej zmiennej obliczamy sumę

$$TOPRI(u) = \sum_{u \neq v} ARI(u,v), \tag{2}$$

którą możemy interpretować jako miarę siły związku podziałów zbioru, w których brane były pod uwagę wartości tej zmiennej ze wszystkimi podziałami, w których nie były one brane pod uwagę. Zmienne o największych wartościach $TOPRI(u)$ mają, według autorów metody, najsilniejszy związek ze strukturą skupień. Ostatnim etapem jest podzielenie wszystkich zmiennych na dwa zbiory: zmiennych istotnych i maskujących. W tym celu można wykorzystać kryterium największego skoku wskaźnika $TOPRI(u)$, tzn. po uporządkowaniu wartości tego wskaźnika malejąco obliczamy ilorazy

$$RR(s) = \frac{\left(TOPRI\left(k(s)\right) - TOPRI\left(k(s+1)\right) \right)}{\left(TOPRI\left(k(s-1)\right) - TOPRI\left(k(s)\right) \right)} \tag{3}$$

i wybieramy s początkowych zmiennych do największej wartości $RR(s)$.

Metodę VS-KM określamy również tak jak w pracy Steinleya i Brusco [2008]. Idea tej metody polega na tym, by rozpocząć poszukiwanie zbioru zmiennych istotnych dla struktury skupień od pary zmiennych mającej najwyższą wartość $ARI(u, v)$. Następnie do tej pary dołączane są sekwencyjnie pojedyncze zmienne. W każdym kroku dołączana jest zmienna s, która ma najwyższy wskaźnik podobieństwa podziału opartego na zmiennej s oraz podziału opartego na wszystkich zmiennych połączonych do tego kroku. Tak rozumiana metoda wymaga określenia kryterium stopu, czyli minimalnego progu dla wartości wskaźnika podobieństwa podziałów, poniżej którego kończymy dołączanie zmiennych. Przyjęto takie same wartości wszystkich progów koniecznych dla działania metody jak w pracy Steinleya i Brusco [2008].

W celu zachowania porównywalności badania przeprowadzono eksperyment na wzór eksperymentu Steinleya i Brusco [2008]. Poszerzono nieco zakres eksperymentu, dopuszczając zbiory z 2 i 3 skupieniami i maskujące rozkłady równomierne. Wszystkie zbiory składały się z 200 obiektów, różniły się między sobą następującymi cechami.

Pierwsza cecha: liczba skupień, może być równa – 2, 3, 4, 6 lub 8.

Druga cecha: liczebności skupień, możliwe są trzy warianty: (a) równe liczebności wszystkich skupień; (b) 10% obserwacji i (c) 60% obserwacji w jednym skupieniu, a pozostałe skupienia równoliczne.

Trzecia cecha: liczba zmiennych istotnych, może być równa 2, 4 lub 6.

Czwarta cecha: prawdopodobieństwo „zachodzenia na siebie" (*overlap*) skupień na każdej ze zmiennych istotnych, może być równe – 0, 0.1, 0.2, 0.3, 0.4. Separowalność skupień jest typu „łańcuchowego", tj. na każdym wymiarze jest $k - 1$ par skupień (k – liczba skupień), przy czym każde dwa kolejne zachodzą na siebie w stopniu (jednakowym dla wszystkich par), na który wskazuje prawdopodobieństwo.

Piąta cecha: siła korelacji wewnątrz skupień, możliwe są dwa warianty: (a) macierz kowariancji w każdym skupieniu jest macierzą jednostkową; (b) w każdym skupieniu jest taka sama macierz kowariancji z jedynkami na przekątnej i liczbami wybranymi losowo z odcinka [0.3; 0.8] poza przekątną.

Szósta cecha: liczba zmiennych maskujących, może być równa – 2, 4 lub 6.

Siódma cecha: rozkład zmiennych maskujących. Możliwych jest siedem wariantów: (a) wszystkie zmienne niezależnie wygenerowane z rozkładu skośnego jednomodalnego (rozkład gamma z jednym stopniem swobody dla licznika i jednym dla mianownika); (b) wszystkie zmienne niezależnie wygenerowane z rozkładu normalnego ze średnią zero i jednostkową wariancją; (c) wszystkie zmienne niezależnie wygenerowane z rozkładu normalnego ze średnim wektorem zerowym i jedynkami na przekątnej w macierzy kowariancji i 0,25 poza przekątną; (d) wszystkie zmienne niezależnie wygenerowane z rozkładu normalnego ze średnim wektorem zerowym i jedynkami na przekątnej w macierzy kowariancji i 0,5 poza przekątną; (e) wszystkie zmienne niezależnie wygenerowane z rozkładu normalnego ze średnim wektorem zerowym i jedynkami na przekątnej w macierzy kowariancji i 0,75 poza przekątną; (f) wszystkie zmienne wygenerowane z niezależnych rozkładów normalnych ze średnimi równymi zeru i wariancjami losowanymi z odcinka [1; 20] ; (g) wszystkie zmienne niezależnie wygenerowane z rozkładów równomiernych na odcinku [1; 20]. Po uwzględnieniu wszystkich układów parametrów otrzymujemy razem liczbę 11 550 zbiorów.

Z liczb zawartych w tab. 1 i 2 wynika, że HINoV jest bardzo nieodporny na skorelowanie zmiennych zakłócających strukturę skupień. Wadę tę udało się częściowo wyeliminować w modyfikacji VS-KM, ale nadal efektywność metody jest słabsza niż w innych przypadkach. Ponadto metoda VS-KM jest ograniczona tym, że można ją zastosować tylko do zmiennych mierzonych na skali co najmniej interwałowej.

Tabela 1. Efektywność metody HINoV względem typu zmiennych nieistotnych

Typ zmiennych nieistotnych	Beta(1,1)	N(0,1) brak korelacji	N(0,1) słaba korelacja	N(0,1) średnia korelacja	N(0,1) silna korelacja	Normalne o dużej wariancji	Równo-mierne
Pamięć	0,851	0,855	0,805	0,669	0,466	0,853	0,852
Precyzja	0,921	0,928	0,882	0,724	0,506	0,931	0,959
ARI	0,643	0,648	0,601	0,598	0,531	0,652	0,655

Źródło: obliczenia własne.

Tabela 2. Efektywność metody VS-KM względem typu zmiennych nieistotnych

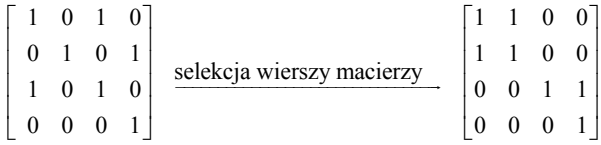
Typ zmiennych nieistotnych	Beta(1,1)	N(0,1) brak korelacji	N(0,1) słaba korelacja	N(0,1) średnia korelacja	N(0,1) silna korelacja	Normalne o dużej wariancji	Równo-mierne
Pamięć	0,885	0,885	0,871	0,852	0,827	0,887	0,830
Precyzja	0,957	0,974	0,951	0,903	0,806	0,973	0,888
ARI	0,697	0,699	0,687	0,666	0,613	0,696	0,620

Źródło: obliczenia własne.

3. Opis nowej metody

Sumowanie wskaźników podobieństwa podziałów występujące we wzorze (2) nie ma sensu, gdyż zmienne dość silnie skorelowane, nietworzące żadnej struktury skupień, zawsze będą mogły uzyskać wysoką wartość wskaźnika *TOPRI,* gdy podstawą wyboru będzie najwyższa stabilność podziału opartego na wybranych zmiennych spośród wszystkich możliwych liczb skupień z pewnego zakresu. Zmienne skorelowane zawsze uzyskują wysokie wartości wskaźnika *TOPRI* dla 2 lub 3 skupień. W celu uniezależnienia się od liczby zmiennych zaproponujmy grupowanie zbioru wszystkich zmiennych w dwa podzbiory oddzielnie dla każdej ze zmiennych. Grupowanie można uzyskać metodą *k*-średnich przy $k = 2$. Dla zmiennej v punktami startowymi są dwie skrajne wartości $ARI(u,v)$ spośród wszystkich zmiennych $u \neq v$. Taka metoda podziału jest stabilniejsza od metody największego skoku (por. wzór (3)). Po wielokrotnym (dla każdej zmiennej) podzieleniu zbioru wszystkich d zmiennych na dwa podzbiory otrzymamy macierz zero-jedynkową (po lewej stronie rys. 1), w której jedynką oznaczymy zmienne przyłączone do zmiennej v (reprezentowanej przez wiersz o numerze v), natomiast zerem oznaczymy zmienne przyłączone do innej zmiennej, najmniej „podobnej" do v. Z macierzy tej należy wybrać niektóre wiersze.

Wyboru dokonujemy, rozpatrując wszystkie kombinacje wierszy i kolumn o dowolnej liczbie elementów (od 1 do liczby d wszystkich zmiennych). Dla każdej kombinacji m-elementowej wierszy i każdej kombinacji n-elementowej kolumn



Rys. 1. Idea selekcji zmiennych reprezentowanych przez wiersze macierzy

Źródło: opracowanie własne.

wartość kryterium definiujemy jako sumę dwóch ilorazów: liczby jedynek w bloku składającym się z wybranych wierszy i wybranych kolumn podzielonej przez „pole" bloku, czyli przez *mn,* oraz liczby jedynek w bloku składającym się z niewybranych wierszy i niewybranych kolumn podzielonej przez „pole" bloku, czyli przez (*d*–*m*) (*d*–*n*). W przykładzie z rys. 1 największą wartość kryterium selekcji wierszy uzyskamy, wybierając pierwszy i trzeci wiersz, gdyż dla kombinacji drugiej i trzeciej kolumny otrzymamy 4/4+3/4 = 1,75. Ten sposób obliczania wartości kryterium został przedstawiony w prawej macierzy rys. 1, w której dla lepszego pokazania bloków drugi wiersz został zamieniony z trzecim (z oryginalnej macierzy lewej) oraz druga kolumna została zamieniona z trzecią. Dla żadnego innego podziału macierzy na cztery bloki nie można uzyskać większej wartości sumy obu ilorazów.

Podział zbioru wszystkich zmiennych na dwa rozłączne podzbiory przeprowadzamy dla każdej liczby skupień ze zbioru {2, 3, …, 10}. Z otrzymanych 18 podzbiorów zmiennych wybieramy ten, który ma najwyższą stabilność podziału w replikacji z 20 powtórzeniami z wykorzystaniem metody *k*-średnich z losowym 50-krotnym wyborem punktów startowych. Wszystkie podzbiory zmiennych w takiej procedurze traktowane są jednakowo, nie rozważamy tylko zbiorów początkowych (odrzucając końcowe, tj. te, które są „poniżej łokcia") tak jak w metodzie HINoV.

4. Drugi eksperyment badawczy

Wszystkie struktury skupień z pierwszego eksperymentu zostały zachowane, zmieniono jedynie układ zmiennych maskujących te struktury. Do każdej struktury skupień dołączone zostały 4 zmienne zakłócające. W pierwszym wariancie do każdej struktury skupień dołączone zostały dwie zmienne nieskorelowane o rozkładach równomiernych na odcinku [0; 20] i dwie zmienne o rozkładzie normalnym ze średnim wektorem zerowym i jedynkami na przekątnej w macierzy kowariancji i 0,50 poza przekątną. W drugim wariancie – dwie zmienne o rozkładzie normalnym ze średnim wektorem zerowym i jedynkami na przekątnej w macierzy kowariancji i 0,75 poza przekątną.

5. Wyniki i wnioski

Wyniki drugiego eksperymentu przedstawione są w tab. 1. Modyfikacja osiągnęła te wyniki przy słabszych założeniach, gdyż dopuszczalna była sytuacja, w której selekcja zmiennych mogła zakończyć się wyborem jednej zmiennej. Stosując metodę HINoV z największym skokiem (por. wzór (3)), przy mocniejszych założeniach o możliwości wyboru co najmniej dwóch zmiennych, narzucamy sztucznie selekcję co najmniej dwóch zmiennych, co przy parametrach eksperymentu jest ułatwieniem. Zaproponowana modyfikacja okazała się bardzo stabilna, tzn. niezależnie od tego, czy zmienne zanieczyszczające strukturę skupień są silnie czy tylko średnio silnie skorelowane, osiągnęła bardzo podobne wskaźniki efektywności. Metoda HINoV osiągnęła bardzo niestabilne wyniki, na przykład w przypadku silnie skorelowanych zmiennych zanieczyszczających precyzja była o około 0,14 gorsza niż w przypadku średnio silnej korelacji. Największą zaletą modyfikacji jest to, że wskaźniki efektywności, jakie ona osiągnęła, były w każdym przypadku lepsze o kilka setnych od wyników HINoV. Dotyczy to zarówno precyzji, pamięci, jak i asymptotycznej odzyskiwalności. Modyfikacja osiągnęła dużo wyższą asymptotyczną odzyskiwalność zwłaszcza w przypadku silnego skorelowania zmiennych zanieczyszczających.

Tabela 3. Efektywność porównywanych metod w drugim eksperymencie

	Pierwszy wariant 4 zmiennych zakłócających		Drugi wariant 4 zmiennych zakłócajacych	
	HINoV	Modyfikacja	HINoV	Modyfikacja
Pamięć	0,832	**0,897**	0,801	**0,896**
Precyzja	0,895	**0,940**	0,758	**0,916**
ARI	0,661	**0,695**	0,587	**0,695**

Źródło: obliczenia własne.

Podsumowując wyniki zawarte w tab. 1-3, należy zauważyć, że w niektórych przypadkach rozkładów zanieczyszczających metoda VS-KM ma efektywność bardzo zbliżoną do proponowanej modyfikacji, ale podobnie jak dla HINoV jest ona niestabilna, tzn. dużo słabsza, gdy zmienne zanieczyszczające są skorelowane.

Literatura

Carmone F.J. Jr., Kara A., Maxwell S. (1999), *HINoV: a new model to improve market segment definition by identifying noisy variables* , "Journal of Marketing Research", vol. 36.

Dy J., Brodley C. (2000), *Feature subset selection and order identification for unsupervised learning*, Proc. 17th International Conf. on Machine Learning.

Friedman J., Meulman J. (2004), *Clustering objects on subsets of attributes*, "Journal of the Royal Statistical Society", Series B 66.

Gatnar E., Walesiak M. (red.) (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych,* Wydawnictwo AE we Wrocławiu.
Law M., Jain A., Figueiredo M. (2003), *Feature Selection in Mixture-Based Clustering* , IEEE Transactions on Pattren Analysis and Machine Inteligence.
Montanari A., Lizzani L. (2001), *A projection pursuit approach to variable selection*, "Computational Statistics and Data Analysis", vol. 35(4).
Raftery A.E., Dean N. (2006), *Variable selection for model based clustering,* JASA 101.
Steinley D., Brusco M. (2007), *A new variable weighting and selection procedure for k-means cluster analysis*, "Psychometrika".
Steinley D., Brusco M. (2008), *Selection of variables in cluster analysis: an empirical comparison of eight procedures*, "Psychometrika" 73.

MODIFICATION OF THE HINOV METHOD OF SELECTING VARIABLES IN CLUSTER ANALYSIS

Summary: The HINoV metod of variable selection has been known since 1999. The original method is not resistant to the existence of correlated variables among the noisy variables. This drawback was partially eliminated in the VS-KM modification of HINoV. In the article a modification of HINoV is proposed, consisting in the assessment of stability of the data division for each number of clusters and for both of the compared sets of variables. In the simplest variant one has to choose the subset whose stability criterion is highest. A new way of dividing the set of variables into two subsets is also proposed. The stability criterion is based on repeated drawing of roughly half of the data and comparing the divisions received with the help of the *k*-means. The modification is assessed in a broad simulation experiment comprising 8100 data sets with cluster structures generated in the form of the mixtures of normal distributions.

Keywords: cluster analysis, variable selection, HINOV method, VS-KM method.