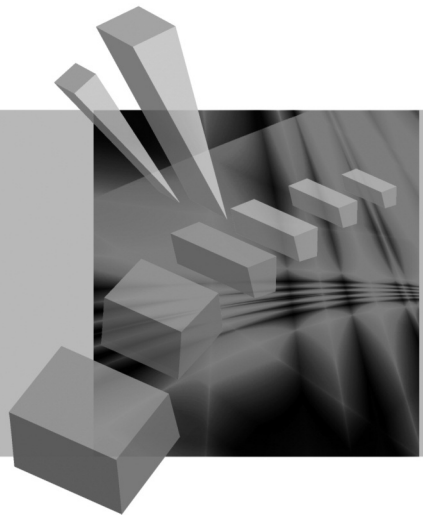
**PRACE NAUKOWE**
Uniwersytetu Ekonomicznego we Wrocławiu
**RESEARCH PAPERS**
of Wrocław University of Economics

**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

**Jerzy Korzeniewski**
Uniwersytet Łódzki

# OCENA EFEKTYWNOŚCI METODY UŚREDNIANIA ZMIENNYCH I METODY ICHINO SELEKCJI ZMIENNYCH W ANALIZIE SKUPIEŃ

**Streszczenie:** Metodę Ichino można stosować do zmiennych mierzonych na dowolnych skalach. Pełka i Wilk zbadali tę metodę, ale badanie ograniczało się do kilku modeli struktur skupień, zaś kryterium oceny była stabilność grupowania mierzona indeksem sylwetkowym. Metoda uśredniania zmiennych nieistotnych dla struktury skupień [Fraiman i in. 2008] jest nową metodą, niezbadaną dotychczas w żadnym obszerniejszym eksperymencie. W artykule przedstawione są wyniki badania symulacyjnego, w którym kryteriami są pamięć, precyzja i indeks Randa. Metody konkurencyjne, z którymi porównywana jest metoda Ichino, to HINoV oraz dwie modyfikacje HINoV (VSKM i VAF z indeksem skupialności).

**Słowa kluczowe:** analiza skupień, wybór zmiennych, graf wzajemnego sąsiedztwa, uśrednianie zmiennych.

## 1. Wstęp

W kontekście analizy skupień ważnym zagadnieniem jest wybór zmiennych istotnych dla struktury skupień zbioru danych. Jeśli wybór zostanie dokonany błędnie lub zbiór zmiennych nie zostanie w ogóle oczyszczony ze zmiennych nieistotnych, to fakt ten na ogół wpływa bardzo negatywnie na wnioski uzyskane z analizy skupień. Do chwili obecnej zaproponowano kilkanaście podejść do zagadnienia selekcji zmiennych. Metody 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, takich jak: 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 skupialości (VAF, *relative clusterability weighting method* [Steinley, Brusco 2007]). Ocena efektywności oparta była na trzech głównych wskaźnikach: pamięci (*recall*), precyzji (*precision*) oraz asymptotycznej odzyskiwalności poprawnego przypisania obserwacji do skupień (ARI*, asymptotic recovery*). Wskaźniki te definiuje się na potrzeby 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.

Od czasu cytowanego badania Steinleya i Brusco pojawiły się nowe metody. Istnieją również inne podejścia, wcześniejsze, niezbadane w cytowanym eksperymencie.

W artykule tym przedstawione są wyniki badań efektywności metody Ichino [1994] oraz metody uśredniania zmiennych [Fraiman i in. 2008]. Metoda Ichino ma bardzo ogólny charakter względem rodzaju skal pomiarowych, na których mierzone są zmienne opisujące obiekty – można ją stosować do dowolnych skal. Pełka i Wilk zbadali tę metodę, ale badanie ograniczało się do kilku modeli struktur skupień tworzonych przez zmienne ciągłe. Kryteria oceny były inne od tych zastosowanych przez Steinleya i Brusco [2008]. Efektywność oceniana była w sensie stabilności grupowania ocenianej za pomocą indeksu sylwetkowego (por. [Pełka, Wilk 2010]). Metoda uśredniania zmiennych nieistotnych dla struktury skupień jest nową metodą niezbadaną dotychczas w żadnym obszerniejszym eksperymencie. Podstawą wyników badań przedstawionych w dalszym ciągu artykułu jest eksperyment symulacyjny podobny do eksperymentu Steinleya i Brusco [2008]. Zastosowano te same miary efektywności. Obie metody oceniane były na tle trzech metod, które w badaniu Steinleya i Brusco [2008] wypadły najlepiej, tj. HINoV oraz dwóch modyfikacji HINoV (metody VSKM i VAF z indeksem skupialności).

## 2. Charakterystyka metody Ichino

Ichino opracował bardzo ogólny model populacji – model przestrzeni kartezjańskiej (*cartesian space model*). Model ten może być stosowany do zbiorów danych z szerokim zakresem kategorii zmiennych, od ciągłych do nominalnych. Model oparty jest na pojęciu operatora sumy obiektów $\oplus$ oraz operatora iloczynu obiektów $\otimes$. Modelem przestrzeni kartezjańskiej nazywamy trójkę $(U^d, \oplus, \otimes)$, gdzie $U^d$ jest $d$-wymiarową przestrzenią wartości zmiennych losowych opisujących obiekty zbioru danych

$$U^d = U_1 \times U_2 \times \ldots \times U_d, \tag{1}$$

natomiast operator sumy i iloczynu definiujemy w następujący sposób. Sumę kartezjańską dwóch obiektów $A$ oraz $B$ definiuje się w postaci iloczynu kartezjańskiego operatora sumy poszczególnych współrzędnych:

$$A \oplus B = \left(A_1 \oplus B_1\right) \times \left(A_2 \oplus B_2\right) \times \ldots \times \left(A_d \oplus B_d\right). \tag{2}$$

Operator iloczynu dwóch obiektów $A$ oraz $B$ definiujemy w postaci iloczynu kartezjańskiego operatora iloczynu poszczególnych współrzędnych:

$$A \otimes B = \left(A_1 \otimes B_1\right) \times \left(A_2 \otimes B_2\right) \times \ldots \times \left(A_d \otimes B_d\right). \tag{3}$$

W dalszym ciągu definiujemy obiekty wzajemnie sąsiadujące (*mutual neighbours*) $A_1$ oraz $A_2$ względem zbioru obiektów symbolicznych $B = \{B_1, B_2, ..., B_m\}$ jako obiekty spełniające warunek:

$$\forall B_i \in B, \quad B_i \otimes \left(A_1 \oplus A_2\right) = \varnothing. \tag{4}$$

Zbiór obiektów symbolicznych $A = \{A_1, A_2, ..., A_k\}$ nazywamy grafem wzajemnego sąsiedztwa (*mutual neighbourhood graph*) względem zbioru obiektów symbolicznych $B = \{B_1, B_2, ..., B_m\}$, gdy każda para obiektów ze zbioru $A$ jest wzajemnie sąsiadująca względem zbioru $B$. Grafy wzajemnego sąsiedztwa pozwalają opisywać separowalne skupienia obiektów. Ichino, opierając się na takim sposobie badania podobieństwa grup obiektów, zaproponował kilka metod selekcji oraz grupowania zmiennych.

W przeprowadzonym dalej eksperymencie zastosowano następującą metodę.

1. Dla każdego podzbioru $s$ zmiennych grupujemy wszystkie obiekty zbioru danych w 2 skupienia za pomocą metody $k$-średnich z losowym wyborem obiektów startowych (losujemy 50 razy i zapamiętujemy grupowanie z najmniejszą wariancją wewnątrzgrupową).

2. Dla każdego $s$ znajdujemy największą liczbę par $l(s)$ wzajemnie sąsiadujących względem drugiego skupienia.

3. Za najlepiej odzwierciedlający strukturę skupień uznajemy ten podzbiór $s$ zmiennych, dla którego różnica $l(s)–l(s\text{-}1)$ ma wartość najwyższą.

Pełka i Wilk sformułowali nieco inaczej metodę Ichino. Dla każdego podzbioru zmiennych konstruowany był zbiór grafów wzajemnego sąsiedztwa, po czym wystarczyło zliczyć liczbę par w każdym grafie (nie musimy sprawdzać, czy pary są wzajemnie sąsiadujące, gdyż zapewnia to algorytm szukania grafów).

## 3. Charakterystyka metody uśredniania zmiennych

Podstawą metody jest spostrzeżenie, że zmienne nieistotne dla struktury skupień powinny mieć we wszystkich skupieniach takie same rozkłady i, co za tym idzie, wartości średnie. Wobec tego, gdy zbiór danych został podzielony na pewną liczbę skupień, wystarczy sprawdzić, jak często podział oparty na wszystkich zmiennych pokrywał się z podziałem opartym na wybranym podzbiorze zmiennych, który uznamy za zbiór zmiennych istotnych, przy wartościach pozostałych zmiennych równych wartościom średnim.

Szczegółowe sformułowanie algorytmu metody jest następujące.

1. Dla zbioru wszystkich *di* zmiennych oryginalnych grupujemy wszystkie obiekty zbioru danych w tę samą liczbę skupień za pomocą metody *k*-średnich z losowym wyborem obiektów startowych (losujemy 50 razy i zapamiętujemy grupowanie z najmniejszą wariancją wewnątrzgrupową).

2. Dla każdego podzbioru *s* zmiennych ($s < di$) zmodyfikowanych (*s* zmiennych ma wartości oryginalne, zaś wartości pozostałych *di*-*s* zmiennych zastępujemy wartościami średnimi) przypisujemy każdy obiekt do najbliższego centrum skupienia wyznaczonego w kroku pierwszym.

3. Spośród wszystkich podzbiorów *s* zmiennych za najlepiej odzwierciedlający strukturę skupień uznajemy minimalny podzbiór, który ma największą zgodność grupowania z grupowaniem uzyskanym dla zbioru wszystkich zmiennych.

Zgodność grupowania mierzymy za pomocą zmodyfikowanego indeksu Randa.

## 4. Eksperyment badawczy

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. Se-

parowalność 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, zaś poza przekątną jest liczba wylosowana z odcinka [0.3; 0.8].
- 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.

## 5. Wyniki i wnioski

Wyniki przedstawione w tab. 1 pozwalają stwierdzić, że obie metody okazały się dużo gorsze od metod HINoV, VSKM oraz VAF. Te trzy metody, na tych samych zbiorach danych, uzyskały pamięć i precyzję rzędu 0,8-0,9 oraz wskaźnik *ARI* rzędu 0,7. Dla metody Ichino oraz metody uśredniania zmiennych wartości wskaźników są dużo gorsze. W przypadku metody Ichino można uznać, że wyniku takiego można było się spodziewać, gdyż autor tej metody raczej konstruował ją z myślą o słabszych skalach pomiarowych. We wszystkich publikacjach Ichino dominują przykłady ze zbiorami, w których występują tylko zmienne mierzone na skali nominalnej. Metoda jest bardzo czasochłonna, konieczne jest zbadanie wszystkich możliwych podzbiorów zbioru wszystkich zmiennych.

Słaba efektywność metody uśredniania zmiennych jest zaskakująca, gdyż zdaniem autorów metoda jest dobra. Porównując badanie autorów metody z naszym eksperymentem, należy zauważyć, że rozważali oni jedynie prosty przypadek dwóch zmiennych tworzących trzy raczej bardzo wyraźnie separowane skupienia. Odpo-

**Tabela 1.** Średnia pamięć, precyzja i asymptotyczna odzyskiwalność dla obu metod w zależności od liczby zmiennych nieistotnych dla struktury skupień oraz występowania w skupieniach korelacji wewnątrzgrupowej

| Liczba zmiennych nieistotnych | Występowanie korelacji wewnątrz skupień | Metoda Ichino | | | Metoda uśredniania zmiennych | | |
|---|---|---|---|---|---|---|---|
| | | pamięć | precyzja | ARI | pamięć | precyzja | ARI |
| 2 | Nie | 0,947 | 0,696 | 0,508 | 0,787 | 0,688 | 0,481 |
| 4 | Nie | 0,951 | 0,543 | 0,429 | 0,824 | 0,531 | 0,408 |
| 6 | Nie | 0,950 | 0,448 | 0,407 | 0,838 | 0,433 | 0,366 |
| 2 | Tak | 0,946 | 0,698 | 0,499 | 0,799 | 0,701 | 0,439 |
| 4 | Tak | 0,949 | 0,543 | 0,426 | 0,846 | 0,547 | 0,384 |
| 6 | Tak | 0,957 | 0,437 | 0,400 | 0,847 | 0,451 | 0,350 |
| **Średnio** | | **0,933** | **0,561** | **0,499** | **0,821** | **0,559** | **0,405** |

Źródło: obliczenia własne.

wiadałoby to niemalże zbiorom z „overlap" równym 0 w naszym badaniu. Dla tych zbiorów wskaźnik ARI jest znacznie wyższy, równy ok. 0,7. Bardzo duży jest wtedy również odsetek zbiorów z bezbłędnym grupowaniem. Wadą metody jest jednak to, że niekiedy mamy do wyboru kilka podzbiorów zmiennych mających dokładnie taką samą wartość kryterium selekcji. Wówczas w algorytmie przyjęto, że wybrać należy podzbiór najmniej liczny. Ten sposób wybierania jednego z podzbiorów bardzo negatywnie wpływa na średnią precyzję, jaką uzyskała metoda. Jeżeli w przypadku równej wartości kryterium selekcji wybieralibyśmy wszystkie zmienne wchodzące w skład podzbiorów dających taką samą wartość kryterium, to precyzja metody byłaby nieco wyższa. Ciekawe jest to, że w przypadku występowania korelacji wewnątrz skupień metoda spisuje się wyraźnie gorzej. Zapewne gdy skupienia są „wydłużone", to przypisywanie obiektu do wyznaczonego wcześniej (przy uwzględnieniu wszystkich zmiennych) centrum skupienia jest częściej błędne. Metoda w rozważanej postaci jest czasochłonna, ale dla dużych zbiorów zmiennych autorzy proponują algorytm filtrujący zbiór wszystkich zmiennych, typu *forward-backward search*, pracujący na zasadzie dołączania pojedynczych zmiennych do zbioru istniejącego. Ten algorytm został jednak zbadany tylko na dwóch empirycznych zbiorach danych. Efektywność obu metod bardzo szybko spada wraz ze wzrostem liczby zmiennych nieistotnych dla struktury skupień. Należy nadmienić, że dla metody uśredniania zmiennych badanie przeprowadzone było przy założeniu znajomości liczby skupień. To założenie w praktyce zbiorów empirycznych na ogół nie jest spełnione.

## Literatura

Brusco M., Cradit D., *A variable-selection heuristic for k-means clustering*, "Psychometrika" 2001, no 66.

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

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

Fraiman R., Justel A., Svarc M., *Selection of variables for cluster analysis and classification rules*, "JASA" 2008, no 103.

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

Ichino M., *Feature Selection for Symbolic Data Classification*, [w:] *New Approaches in Classification and Data Analysis,* Springer-Verlag,1994.

Law M., Jain A., Figueiredo M., *Feature Selection in Mixture-Based Clustering*, IEEE Transactions on Pattren Analysis and Machine Inteligence, 2003.

Montanari A., Lizzani L., *A projection pursuit approach to variable selection*, „Computational Statistics and Data Analysis" 2001, vol. 35(4).

Pełka M., Wilk J., *Metody selekcji zmiennych symbolicznych w zagadnieniach klasyfikacji,* [w:] Taksonomia 17, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 107, UE, Wrocław 2010.

Raftery A.E., Dean N., *Variable selection for model based clustering,* „JASA" 2006, no 101.

Steinley D., Brusco M., *A new variable weighting and selection procedure for k-means cluster analysis*, "Psychometrika" 2007.

Steinley D., Brusco M., *Selection of variables in cluster analysis: an empirical comparison of eight procedures*, „Psychometrika" 2008, no 73.

## EFFICIENCY ASSESSMENT OF ICHINO METHOD AND MEAN VALUE METHOD OF SELECTING VARIABLES IN CLUSTER ANALYSIS

**Summary:** Ichino method can be applied to variables measured on any type of scale. Pełka and Wilk examined the method but their investigation was limited to a couple of cluster structures and the criterion used was the stability of grouping based on the silhouette index. The mean value method [Fraiman et al. 2008] is a relatively new method not investigated until now in a broad simulation experiment. In the paper the results of simulation experiment are presented. The criteria used are: recall, precision and Rand index. The competing methods are HINoV and its two modifications: VSKM and VAF with clusterability index.

**Keywords:** cluster analysis, variable selection, mutual neighbourhood graph, taking mean values of variables.